TARTU ÜLIKOOL
Matemaatika-informaatikateaduskond
Matemaatilise statistika eriala

Kaido Lepik

Spordiennustused: kihlveokontoritega konkureerimine NBA-s

Magistritöö (30 EAP)

Juhendaja: Jüri Lember, PhD

Tartu 2014

Spordiennustused: kihlveokontoritega konkureerimine NBA-s

Käesolev magistritöö püüab näidata, et spordikihlvedusid võib sõlmida professionaalsetel alustel, arvestades riskiga ja baseerides panustamisotsused matemaatikale. Töös on sporditulemustele ennustamist vaadeldud mitmekülgselt, alustades teema motiveerimisega ja probleemistiku uurimisega, kogudes ja korrastades suurel hulgal olulisi andmed, tutvustades juba varasemalt tehtud töid ja ideid; pakutud on uusi lahendusi, implementeeritud mitmeid algoritme ja teostatud kogutud andmetel põhjalik analüüs.

Magistritöö jaoks on veebiroboti abil kogutud enam kui 15000 korvpallimängu andmed aastatelt 2000 kuni 2013 ja rohkem kui 5000 korvpallimängu koefitsiendid paljudelt kihlveokontoritelt. Mängude kohta kogutud informatsioon hõlmab nii meeskondade, mängijate ja viisikute kohta käivaid kokkuvõtlikke statistikuid kui ka sündmus-sündmus andmeid. Kõik andmed on korrastatud ja organiseeritud relatsioonilisse andmebaasi.

Analüüsi osas veenduti esialgu teoreetiliselt tõestatud tulemuses, et juhuslikult spordisündmustele panustamine on keskmiselt kahjumlik. Seejärel püüti kasumlikult panustada lihtsate mudelite abil, mis klassifitseerisid korvpallimängu võitja meeskondade eelnevate omavaheliste mängude põhjal. Leiti mudel, mis suurest testandmetel tehtud klassifitseerimisveast (41,4%) hoolimata andis panustamissituatsioonis suure tulususe.

Kihlveokontoreid püüti võita ka tehisõppe meetodite abil. Selleks kasutati logistilist regressiooni ja AdaBoosti, sobivate tunnuste valikuks implementeeriti mitmed heuristikud. Ükski nimetatud meetoditega treenitud klassifitseerija ei olnud panustamisel kasumlik, samas suutis parim logistilise regressiooni mudel klassifitseerida korrektselt 68,9% testmängudest.

Lihtsate mudelite ja tehisõppe meetoditega leitud mudelite põhjal veendusime, et parem klassifitseerija ei pruugi anda suuremat kasumit. Seetõttu on klassifitseerijate ehitamisel treeningriski minimiseerimise asemel proovitud maksimiseerida ka treeningkasumit. Ideed on püütud jõuga realiseerida otsustuspuude abil. Samuti on implementeeritud modifitseeritud AdaBoosti meetod, mis kaalus vaatlusi vastavalt koefitsientide suurusele ja töötas kohati paremini kui originaalne AdaBoost. Lisaks on korvpallimängude võitjaid proovitud ennustada korvpallitulemuste simuleerimise abil Poissoni protsesside põhjal.

Märksõnad: algoritmid, andmekaeve, andmevalmendus, juhuslikud protsessid, klassifitseerimine, korvpall, matemaatilised mudelid, mustriotsing, optimiseerimine, simulatsioon, spordiennustused, statistiline andmetöötlus, tehisõpe.

Sports betting: trying to beat the bookmakers in NBA

This master's thesis tries not to succumb to the preconception that sports betting is essentially gambling, rather it treats sports betting as a form of investment. The thesis covers a lot of what is important in order to have an outside chance to succeed in sports betting. It starts with an overview of the field which has been presented in a compact mathematical format, using proofs. After covering the problem space, a lot of data has been gathered and organized for the analysis, previous work in the literature has been revised, many new approaches have been proposed, several algorithms have been implemented and a thorough analysis of the data has been conducted.

For the analysis, data has been gathered using a web bot: statistics about more than 15000 NBA games from 2000-2013, including play-by-play data; odds on more than 5000 games by several bookmakers. All the data has been organized into a relational database.

The analysis starts with an empirical verification of the theoretical proof that a punter loses in the long term if he places his bets randomly. After that, however, it has been tried to bet profitably using simple models which classify the outcome of a match in accordance with teams' past meetings. Somewhat surprisingly, an average model with predictive power of only 58,6% has been found to be highly profitable.

For more complex models, machine learning techniques with feature selection algorithms have been implemented and applied. Several logistic regression and AdaBoost classifiers have been trained and tested, however, none of these models were found profitable, although the best logistic regression classifier managed to classify 68,9% of the test matches correctly.

It is deducted from the performance of the models that profitability does not go hand in hand with accuracy. Therefore, instead of focusing on minimizing training error, it has been tried to train the models in such a way as to maximize profit, using decision trees. For the purposes of studying whether the punter should concentrate on games with higher or smaller odds, a modified version of AdaBoost has been devised, placing weights on training instances either proportional or inverse proportional to odds. Also, it has been tried to predict the winners of basketball matches by simulating basketball scores by Poisson processes.

Keywords: algorithms, automatic learning, basketball, data acquisition, data mining, classification, mathematical models, optimization, pattern mining, simulation, sports betting, statistical data processing, stochastic processes.

Sisukord

Sissejuhatus

Spordiennustustega tegelemine on üldjuhul midagi, millest avalikult ei räägita. Subjektiivses või igapäevases käsitluses ei kõla see prestiižselt nagu finantsturgudel kauplemine: ei assotsieeru uhkete ülikondade, tuntud ettevõtete või suure vastutusega; seda ei asetatata samasse konteksti riskide maandamise, keerukate mudelite või matemaatikal põhinevate otsustega. Pigem heidetakse spordiennustused ühte hasartmängurlusega.

See ei pea tingimata olema nii. Ka sporditulemuste ennustamisse on võimalik suhtuda professionaalselt: riske on võimalik vähendada, ennustused võivad baseeruda kvantitatiivsetel mudelitel, otsustamisel saab lähtuda emotsioonide asemel matemaatikast.

Käesolev magistritöö orienteerub spordiennustuste valdkonnas ja üritab selles keskkonnas navigeerimisel tugineda just numbriliste meetodite abile. Eesmärk on võrdlemisi praktiline: korvpalliliiga NBA[1] mängude ennustamises püütakse kihlveokontorit pikas perspektiivis võita. Sellest tulenevalt on töö valdavalt ka rakendusliku iseloomuga. Teoreetilist käsitlust esitatakse vaid nii palju, kui tundub materjali mõistmiseks tarvis olevat. Kasutatavate võtete sügavamaks mõistmiseks suunatakse lugejat põhjalikuma käsitlusega allikatele. Sellegipoolest püüab töö olla võimalikult iseseisev ning vähemalt põgusalt argumenteerida nii rakendatavaid lahenduskäike ja -ideid kui ka tulemusi, millel need baseeruvad.

Töö kirjeldab paljusid etappe, mida võib olla tarvis spordiennustustega tegelemisel läbida. Esimeses peatükis defineeritakse olulisemaid mõisteid, avatakse valdkonna tagamaid ning loodetavasti luuakse lugejas uskumine, et ka ebaausast heitlusest on võimalik võitjana väljuda. Seejuures on terve peatüki matemaatiline käsitlus autori enda panus, mis üksnes tugineb väheses spordiennustuste alases kirjanduses esitatud jutustavale ja

[1] *National Basketball Association*

mõneti üldistatud tekstile. Teises peatükis esitatakse lühidalt matemaatiline kirjaoskus, mida on hilisema analüüsi mõistmiseks vaja: välja tuuakse notatsioon ja kasutust leidvad tulemused tõenäosusteooriast, klassifitseerimisteooriast ja juhuslike protsesside teooriast. Kolmandas peatükis kogutakse, organiseeritakse, varustatakse ja töödeldakse vajalikud andmed – nii NBA mängude statistika kui kihlveokontorite koefitsiendid –, et neid oleks mugav analüüsiks kasutada. Kõik andmed on autor ise hankinud ja analüüsiks sobivale kujule töödelnud. Neljandas peatükis esitatakse lühidalt teoreetilised tulemused, millel analüüs põhineb: kirjeldatakse kasutatavate meetodite põhimõtet ja ideed. Samuti tutvustatakse korvpallimängude klassifitseerimisel juba varasemalt tehtud töid. Viiendas peatükis asutakse viimaks sporditulemusi ennustama: ühelt poolt kasutatakse selleks erinevaid klassifitseerijaid, teiselt poolt üritatakse probleemile läheneda mängutulemusi simuleerides. Peatükis pakutakse välja ka mitmeid uusi lähenemisi, kuidas kasumlikult panustada Kõikide kasutatud ja realiseeritud meetodite headust võrreldakse nii klassifitseerimistäpsuse kui kasumlikkuse alusel tegelikus panustamissituatsioonis.

Märkused: töö autor ei ole ise kunagi sporditulemustele panustanud; töö raames valminud programmi kood on ligipääsetav lisas (A) toodud lingi alt.

Peatükk 1

Probleemi domeen

Enne probleemi kallale asumist tuleb selgeks teha mängureeglid. Järgnevalt tutvustataksegi lugejale valdkonda, milles käesolev magistritöö orienteerub. Defineeritakse seonduvad mõisted ja püütakse lugejat veenda, et tegelikult võib *spordikihlvedusid*[2] käsitleda ühena paljudest investeerimisvormidest. Esitatud materjali formaalne käsitlus on autori enda looming.

1.1 Spordiennustus

Üldises käsitluses tähendab ennustamine arvamuse esitamist mingi tulevikus toimuva nähtuse kohta. Enamasti kaasneb selle nähtusega teatav juhuslikkus, st tulemus ei ole kindlalt fikseeritud. Näiteks võime ennustada, kas homme on ilm pilves või mitte. Mõlemad variandid on võimalikud, mis teeb korrektselt ennustamise keeruliseks ülesandeks.

Spordikihlveo sõlmimine või spordiennustuse tegemine tähendab mingi *spordisündmuse* peale raha paigutamist, eesmärgiga saada tulu. Enamasti on ühes kontekstis võimalik panustada erinevatele (üksteist välistavatele) sündmustele. Näiteks korvpallimängu kahe meeskonna vahel võib võita nii üks kui teine meeskond, seega on võimalikke seotud sündmusi kaks ja ennustamise variante samuti kaks: kihlvedu saab sõlmida nii ühe kui teise meeskonna võidule. Kihlveo sõlmijat nimetatakse seejuures *mängijaks*, investeeringu suurust aga *panuseks*. Investeeringu tulusus sõltub sündmuse toimumisest: kui sündmust ei toimu, siis mängija kaotab oma panuse; kui sündmus toimub, siis lisaks oma esialgse panuse tagasisaamisele võidab

[2]Kasutatakse sünonüümina spordiennustustele

mängija mingi täiendava summa. Viimase suuruse määrab koefitsient.

1.2 Koefitsiendid

Erineva haridusliku taustaga mängijad võivad *koefitsiendi*[3] mõistet tõlgendada mitmeti. Spordiennustustes ja seega ka käesolevas töös mõeldakse selle termini all kihlveokontori poolt sündmusele seatud teatud sorti suurust, mis fikseerib sündmuse toimumise korral mängija kasu ehk kihlveokontori väljamakse. Koefitsienti saab mitmel erineval kujul kirja panna ka spordiennustuste valdkonnas. Käsitleme mõistet esialgu holistlikumalt ja alustame koefitsiendi esitamist murru abil, sest selline kirjapilt on teisi valdkondi arvesse võttes universaalsem ja võimaldab lihtsamat tõlgendust väljaspool spordivaldkonda.

Definitsioon 1.1. *Me ütleme, et mingi sündmuse A toimumise koefitsient k^i on inglise tüüpi, kui see on esitatud mittenegatiivse ratsionaalarvuna:*

$$k^i = \frac{1-p-o}{p+o} = \frac{x^*}{y^*}, \tag{1.1}$$

kus p tähistab sündmuse A toimumise tõenäosust ning $-p < o \leq 1-p$ on mingi väike positiivne suurus; $x^* \in \mathbb{N} \cup \{0\}$ ja $y^* \in \mathbb{N}$.

Paneme tähele, et kui definitsioonis võtta $o = 0$, siis vähemalt statistikat tundvale mängijale võib esitatud lähenemine ja kirjapilt tuttav olla. Samamoodi tähistatakse ka *šansside suhet*. Seega võiks koefitsiendi interpretatsioon $o = 0$ korral kõlada järgmiselt: mingi sündmuse iga y^* toimumise kohta vastab keskmiselt x^* mittetoimumist. Koefitsient on seega tihedalt seotud tõenäosusega. Selguse huvides näitame koefitsiendi ja sündmuse toimumise tõenäosuse vahelise seose otseselt.

Omadus. *Sündmuse toimumise tõenäosus p avaldub sündmuse koefitsiendi k^i kaudu kujul $p = \frac{1}{k^i+1} - o$.*

Tõestus. Järeldub vahetult definitsioonist:

$$pk^i + p = 1 - o - k^i o \iff p = (k^i+1)^{-1} - o$$

. □

[3]Magistritöös mõeldakse siin ingliskeelset terminit *odds*

Näide 1.1. *Olgu $o = 0$. Vaatleme täringu veeretamist: oletame, et meile pakub huvi, kas saame täringuviskel 2 silma. Ausa täringu puhul toimub see keskmiselt 1 korral kuuest viskest, ülejäänud 5 korral saame midagi muud. Täringuviskel 2 silma saamise koefitsiendi võime seega kirjutada kui* $5/1$.

Ülaltoodud näites toodud loogikat võib kasutada ka teist tüüpi sündmuste koefitsientide leidmiseks (endiselt $o = 0$): nädalast juhusliku päeva valimisel tööpäevale sattumise koefitsient võiks olla $2/5$, ausa mündi viskamisel tuleb iga kulli kohta keskmiselt ka 1 kiri ning sellisel juhul võiks koefitsient kulli (või ka kirja) saamisele olla $1/1$. Näeme, et kui sündmus toimub suurema tõenäosusega kui 0,5, siis on $k^i < 1$, kui sündmuse ja tema vastandsündmuse toimumise tõenäosused on võrdsed, siis $k^i = 1$ ja kui sündmus toimub väiksema tõenäosusega kui 0,5, siis on $k^i > 1$.

Näide 1.2. *Vaatleme täringu veeretamise näidet ka tõenäosuste kaudu. Olgu huvipakkuvate täringu silmade hulk $A = \{2\}$ ning olgu p_{A^c} ja p_A tõenäosused, et täringuviskel vastavalt ei tule ja tuleb 2 silma. Olgu $o_A = 0$. Kahe silma saamise koefitsient k^i_A on siis*

$$k^i_A = \frac{p_{A^c}}{p_A} = \frac{1 - p_A}{p_A} = \frac{5/6}{1/6} = \frac{5}{1}.$$

Siinkohal võib lugejal tekkida küsimus, mida tähendab koefitsiendi definitsioonis o. Põhimõtteliselt määrab o koefitsiendi aususe: koefitsient on aus, kui $o = 0$.

Näide 1.3. *Jätkame eelmist näidet. Kui ka $o_B = 0$, siis koefitsient täringuviskel 2 silma mittesaamisele on* $\frac{1}{5}$ *ehk* $k^i_{A^c} = \frac{1}{k^i_A}$.

Kui sündmuse ja selle vastandsündmuse koefitsiendid on ausad, siis on koefitsientide omavaheline korrutis 1. Spordiennustuste kontekstis koefitsiendid paraku enamasti ausad ei ole. Koefitsiendi definitsiooni on lisatud lisaliige o, et tagada ka sellisel juhul matemaatiline korrektsus. Suurus o on tundmatu väike positiivne suurus, mistõttu võiks koefitsiendi alusel sündmuste A ja A^c toimumise tõenäosususi p^*_A ja $p^*_{A^c}$ hinnata kujul

$$p^*_A = \frac{1}{k^i_A + 1}, \quad p^*_{A^c} = \frac{1}{k^i_{A^c} + 1}.$$

Märkus. *Edaspidi tähistamegi otse koefitsiendilt* $k^i = \dfrac{x^*}{y^*}$ *hinnatud tõenäosust* $p^* = \dfrac{1}{k^i+1} = \dfrac{y^*}{x^*+y^*}$.

Siin tuleb aga märgata, et võib kehtida $p^*_{A^c} + p^*_A \neq 1$, kuigi $A \cup A^c$ määrab kogu elementaarsündmuste ruumi. Hoomamaks sellise matemaatilise piirangu puudumise olulisust, peame mõisteid käsitlema spordiennustuste kontekstis.

•

Kui mängija sõlmib kihlveo, panustades mingi sündmuse toimumisele, siis selle sündmuse realiseerumise korral ootab ta teatud suuruses rahalist kasu. Inglise tüüpi koefitsient määrab sellisel juhul võidetava summa suuruse: kui koefitsient on $k^i = x^*/y^*$ ja panus b ühikut, siis kihlveo võitmise korral võidab mängija $k^i \times b$ ühikut, st iga panustatud ühiku pealt teenib mängija k^i täiendavat ühikut kasumit. Seega võib koefitsienti interpreteerida ka järgnevalt: iga sündmuse toimumisele panustatud y^* ühiku pealt teenib mängija selle sündmuse toimumise korral x^* ühikut. Eeldades ausaid koefitsiente, siis ülaltoodud täringu veeretamise näites võidaksime 2 silma tulekul iga panustatud ühiku pealt 5 ühikut; nädalast juhusliku päeva tõmbamisel tööpäevale panustades võidaksime iga panustatud 5 ühiku pealt aga vaid 2 ühikut. Oluline on rõhutada, et võidetud summas ei sisaldu esialgu tehtud panus b, st mängija kogusumma pärast panuse võitmist on $(k^i+1)b$. Selline lahendus on mitmes mõttes loogiline:

1. Võidetav summa on üks-üheses pöördvõrdelises vastavuses sündmuse toimumise tõenäosusega, st mida väiksem on tõenäosus sündmuse toimumiseks, seda suurem on potentsiaalselt võidetav summa.

2. Kui koefitsient k^i on aus ja peegeldab sündmuse toimumise tegelikku tõenäosust p, siis keskmiselt jääb mängija sündmusele panustades nulli, st panuse b korral on oodatav kasum

$$pk^ib - (1-p)b = p\frac{1-p}{p}b - (1-p)b = 0.$$

Koefitsientide kirjapanekul on inglise tüüpi esituse kasutamine ebamugav. Seetõttu eelistatakse pigem kümnendesitust, mis annab sama informatsiooni edasi lihtsamal ja kompaktsemal kujul.

Definitsioon 1.2. *Me ütleme, et koefitsient on euroopa tüüpi, kui see avaldub inglise tüüpi koefitsiendi kaudu* $k^e = k^i + 1$ *ja on esitatud kümnendarvuna.*

Märkame, et lisaks erinevusele kirjapaneku metoodikas sisaldab euroopa tüüpi koefitsient erinevalt inglise tüüpi koefitsiendist ka ühikulist panust. Lisaliige muudab koefitsiendi interpretatsiooni, st k^e määrab mängija kogusumma pärast kihlveo võitmist, mitte enam ainuüksi võidetud kasumi: koefitsiendi k^e ja panuse b korral oleks mängija kogusumma pärast kihlveo võitmist $k^e b$.

Näide 1.4. *Olgu inglise tüüpi koefitsient* $k^i = 5/1$*, see on sama mis euroopa tüüpi koefitsient* $k^e = 6$*; inglise tüüpi koefitsient* $k^i = 2/5$ *on sama mis euroopa tüüpi koefitsient* $k^e = 1{,}4$*.*

Lisaks eelmainitutele eksisteerib koefitsientide esitamiseks veel *ameerika* notatsioon, ent seegi on lihtsasti teisendatav nii euroopa kui inglise tüüpi koefitsiendiks, mistõttu siin seda kuju ei avata. Oluline on märgata, et koefitsiendid peidavad endas ühesugust informatsiooni – mis on tõenäosuse hinnang, et sündmus leiab aset ja millise summa mängija võidab, kui tema kihlvedu on edukas –, erinevus seisneb vaid kirjapanekus. Käesolevas töös kasutatakse nii inglise kui euroopa tüüpi esitusi.

1.3 Kihlveokontor

Nagu juba eelnevalt mainitud, siis spordikihlvedude puhul pole enamasti tegemist ausate koefitsientidega. Traditsiooniliselt eksisteerib üksus või entiteet, kes omakasu eesmärgil mängureeglid endale soodsalt sätestab. Nimetame seda üksust edaspidi *kihlveokontoriks*, *kihlvedude vahendajaks* või lühidalt lihtsalt *vahendajaks*[4].

Vahendaja ehk kihlveokontor võimaldab mängijatel sporditulemustele panustada. Selleks määrab vahendaja sündmustele koefitsiendid ja nõustub koefitsientidega määratud tingimustel mängijate panuseid vastu võtma. Kui mängija peaks kihlveo võitma, siis vahendaja tasub talle koefitsiendiga määratud summa ulatuses; kui mängija kaotab, siis saab kihlveokontor panuse endale. Omavahel seotud kihlveod defineeritakse järgnevalt ühise mõistena.

[4]Töös mõeldakse kõigi kolme mõiste all üksust, mida inglise keeles nimetatakse kui *bookmaker* või *bookie*, kuigi sellega võimalikud vahendusvormid ei piirdu.

Definitsioon 1.3. *Kihlvedude kogumiks nimetatakse ühte või mitut spordivõistlust hõlmavat kihlvedude hulka, millel on kokku $N \geq 2$ erinevat lõpptulemust ehk toimuda võivat sündmust ehk kihlveo võimalust A_i, $i = 1,...,N$, $N \in \mathbb{N}$. Seejuures on kõigile N sündmusele võimalik kihlveokontoris panustada.*

Magistritöös mõeldakse kogumi all ühele korvpallikohtumisele pakutud kihlvedusid $N = 2$ võimaliku panustamisvariandiga. Teistel spordivõistlustel võib võimalikke lõpptulemusi ka rohkem olla, nt suusavõistluse võib võita palju erinevaid sportlasi. Spordivõistlusi on võimalik ka üheks kogumiks kombineerida, näiteks kahe jalgpallikohtumise vaatlemisel ühise kihlvedude kogumina on tegu kogumiga, mil on $N = 3^2 = 9$ erinevat toimuda võivat sündmust. Paljusid võimalikke lõpptulemusi kätkevatel kogumitel on üldjuhul riskantsem panustada ning töös neid ei käsitleta.

Kihlveokontor tahab teenida kasumit. Kui ta pakuks spordisündmustele ausaid koefitsiente, siis nii nagu eespool näidatud jääks vahendaja koos mängijaga keskmiselt nulli. Selleks et kihlveokontor oma tegevuselt teenida võiks, vähendab ta koefitsientide suurust – määrab $o > 0$ –, muutes koefitsiendid niimoodi ebaausaks.

Näide 1.5. *Vaatleme jällegi sündmust A, et ausa täringu veeretamisel saadakse 2 silma. Ausad koefitsiendid selle sündmuse toimumisele on endiselt $5/1$ ja $1/5$. Kui mängija panustaks 1 ühiku sündmusele, et täringuga veeretatakse 2 silma, siis õnnestumise korral peaks vahendaja maksma $0{,}2$ ühikut iga panustatud ühiku kohta. Koefitsiente alandatakse väljamakstavate summade vähendamiseks: näiteks võib $5/1$ ja $1/5$ asemel pakkuda koefitsiente $9/2$ ja $1/6$, millele vastaksid lisaliikmed o_A ja o_{A^c} suurustega $1/66$ ja $1/42$. Modifitseeritud koefitsientide pealt tehtavad väljamaksed oleksid nüüd $4{,}5$ ja $1/6$ ühikut iga panustatud ühiku kohta.*

Märkame, et näites toodud uute koefitsientide korrutis on $3/4 < 1$. Tavaliselt teisendatakse koefitsiendid selle nähtuse uurimiseks tõenäosusteks. Koefitsientide pealt leitud tõenäosuste summaks saame

$$\frac{2}{2+9} + \frac{6}{6+1} = 1 + \frac{1}{66} + \frac{1}{42} = \frac{80}{77} \approx 1{,}04.$$

Tulemus on suurem ühest, mis ei ole matemaatiliselt korrektne, aga võimaldab vahendajal kihlveo pealt teenida. Defineerime esitatud nähtuse nüüd ka formaalselt.

Definitsioon 1.4. *Kihlveokontori liigprotsendiks*[5] *r nimetatakse suurust, mille võrra mängu sündmustele A_i seatud koefitsientide $k^e_{A_i}$ pealt arvutatud tõenäosuste summa on suurem ühest, st*

$$r = \sum_{i=1}^{N} \frac{1}{k^e_{A_i}} - 1 = \sum_{i=1}^{N} o_{A_i},$$

kus o_{A_i} on sündmusele A_i seatud koefitsiendi lisaliige.

Paneme tähele, et liigprotsent saab definitsiooni järgi olla negatiivne, st kihlveokontor võib eksida ning pakkuda kihlvedude kogumit, mille koefitsientide pealt arvutatud tõenäosused summeeruvad ühest väiksemaks suuruseks. Samas on see pigem erandlik olukord ja me eeldame, et seda ei juhtu. Rohkem võib huvi pakkuda see, millise panuste jaotuse korral teenib vahendaja kihlvedude kogumilt keskmiselt enim tulu. Vaatleme jätkuvalt kogumit, millel on 2 erinevat kihlveo võimalust ehk toimuda võivat sündmust. Olgu x osakaal kõikidest panustest, mis on pandud esimesele sündmusele ning olgu selle sündmuse toimumise tegelik tõenäosus p. Olgu vahendaja pandud koefitsiendid esimesele ja teisele sündmusele vastavalt k^i_1 ja k^i_2. Sellisel juhul tuleks lahendada järgmine optimiseerimisülesanne:

$$\max_x \left\{ p((1-x) - xk^i_1) + (1-p)(x - (1-x)k^i_2) \right\},$$

mis pärast konstantide eemaldamist lihtsustub kujule

$$\max_x \left\{ (1-p)xk^e_2 - pxk^e_1 \right\}.$$

Optimiseeritav funktsioon on monotoonselt kas kasvav või kahanev, seega on lahendiks $x = 1$ või $x = 0$ vastavalt sellele, kas $(1-p)k^e_2$ on suurem või väiksem kui pk^e_1. Tegelik tõenäosus p on kihlveokontorile aga tundmatu. Seega ei ole kihlveokontor huvitatud panuste saamisest vaid suurima tõenäousega toimuvale sündmusele: keskmiselt oleks võit küll suurim, aga kui toimuma peaks vale sündmus, siis oleks ka kaotus väga suur. Järgnev lause näitab, et teatud tingimustel võib kihlveokontor teenida riskivabalt, toimuvast sündmusest sõltumata.

Lause 1.1. *Olgu kogumil 2 võimalikku teineteist välistavat sündmust: A ja B. Olgu nende toimumisele seatud koefitsiendid vastavalt k^i_A ja k^i_B. Olgu kihlveokontori*

[5] Ingl *overround*

liigprotsendi suurus $r \geq 0$ ning olgu koefitsiendi k^i_A pealt leitud tõenäosus p^*_A. Siis kehtivad järgmised tulemused.

1. *Vahendaja teenib riskivabalt parajasti siis, kui sündmusele A tehtud panuste osakaal on $x_A \in [p^*_A - r, p^*_A]$.*

2. *Vahendaja teenib mängu tulemusest sõltumata konstantse suuruse $v = \dfrac{r}{1+r}$ parajasti siis, kui sündmusele A tehtud panuste osakaal on $x_A = \dfrac{p^*_A}{1+r}$.*

Tõestus. Selleks et mängult riskivabalt teenida, peab kihlveokontor mängu igale sündmusele saama piisavalt panuseid, et sündmuse toimumise korral oleks vastandsündmusele tehtud panuste abil võimalik väljamaksed katta.

Sündmuse A toimumise korral peavad sündmusele B tehtud panused ületama kihlveokontorile tekkinud nõude suurust:

$$x_A k^i_A \leq 1 - x_A \iff x_A \leq \frac{1}{k^i_A + 1} = p^*_A.$$

Ka sündmuse B toimumise korral peavad sündmusele A tehtud panused ületama kihlveokontorile tekkinud nõude suurust:

$$\begin{aligned}(1 - x_A)k^i_B \leq x_A \iff x_A &\geq \frac{k^i_B}{k^i_B + 1} \\ &= 1 - \frac{1}{k^i_B + 1} = 1 + \frac{1}{k^e_A} - \left(\frac{1}{k^e_A} + \frac{1}{k^e_B}\right) \\ &= 1 + p^*_A - (1 + r) \\ &= p^*_A - r\end{aligned}$$

Sellega on lause esimene osa tõestatud. Teise osa tõestuseks peame x_A suhtes lahendama võrrandi:

$$x_A - (1 - x_A)k^i_B = 1 - x_A - x_A k^i_A, \tag{1.2}$$

mille lahendiks on

$$x_A = \frac{1 + k^i_B}{2 + k^i_B + k^i_A} = \frac{k^e_B}{k^e_A + k^e_B}.$$

Korrutades lugejat ja nimetajat suurusega k_A^e ja teisendades saame

$$x_A = \frac{k_A^e k_B^e}{k_A^e(k_A^e + k_B^e)} = \frac{(k_A^e)^{-1}}{\frac{k_A^e + k_B^e}{k_A^e k_B^e}} = \frac{(k_A^e)^{-1}}{\frac{1}{k_A^e} + \frac{1}{k_B^e}}$$
$$= \frac{p_A^*}{1+r}.$$

Leidmaks võidu suurust v, asendame x_A võrrandi (1.2) paremasse poolde (võib ka vasemasse) ning pärast mõningast avaldamist saame

$$j = 1 - \frac{p_A^*}{1+r} - \frac{p_A^*}{1+r}k_A^i = \frac{1 + r - \left(k_A^e\right)^{-1}(1 + k_A^i)}{1+r} = \frac{r}{1+r},$$

millega on ka lause teine osa tõestatud. □

Järeldus 1.1. *Kehtigu lause* (1.1) *eeldused. Kui koefitsiendi k_B^i pealt leitud tõenäosus on p_B^*, siis peavad paika järgmised tulemused.*

1. *Vahendaja teenib riskivabalt parajasti siis, kui sündmusele B tehtud panuste osakaal on $x_B \in [p_B^* - r, p_B^*]$.*
2. *Vahendaja teenib mängu tulemusest sõltumata konstantse suuruse $v = \frac{r}{1+r}$ parajasti siis, kui sündmusele B tehtud panuste osakaal on $x_B = \frac{p_B^*}{1+r}$.*

Tõestus. Esimene osa kehtib, sest

$$x_B \leq p_B^* \iff x_A = 1 - x_B \geq p_A^* + p_B^* - r - p_B^* = p_A^* - r$$

ja

$$x_B \geq p_B^* - r \iff x_A \leq p_A^* + p_B^* - r - p_B^* + r = p_A^*.$$

Teine osa kehtib, sest

$$x_B = \frac{p_B^*}{1+r} \iff x_A = 1 - x_B = \frac{p_A^*}{1+r},$$

millest v avaldub samamoodi nagu lauses (1.1). □

Tulemus peaks lugejat veenma, et kihlveokontori eesmärk ei pruugi olla mängutulemuste võimalikult täpne ennustamine. Riskivabalt kihlvedudelt teenimine ei sõltu vahendaja jaoks sündmuste toimumiste tegelikest

tõenäosustest. Vahendaja riskivabalt teenimise summat v nimetatakse *vaheltkasuks* või *vahendustasuks*[6]. Kui kihlveokontor suudab garanteerida, et mõlemale kogumis pakutavale sündmusele tuleb õiges koguses panuseid, siis ta võidab hoolimata mängu tulemusest vaheltkasu ulatuses. Selle tagamiseks võib vahendaja koefitsiente jooksvalt muuta: kui ühele poolele tuleb panuseid ebaproportsionaalselt, siis koefitsiente korrigeeritakse. See loob mängijale täiendavaid võimalusi kihlvedudest väärtuse leidmiseks. Muidugi on kihlveokontor huvitatud võimalikult suurest vaheltkasust. Seda hoiab kontrolli all tihe konkurents – mängijal on paljude vahendajate vahel võimalik valida see, kes pakub parimat hinda.

Näide 1.6. *Vaatleme kihlvedude kogumit sündmustega A ja B. Olgu ühe kihlveokontori koefitsiendid sündmustele vastavalt* $k^e_{1A} = 1{,}85$ *ja* $k^e_{1B} = 1{,}95$ *ja teise kihlveokontori koefitsiendid sündmustele vastavalt* $k^e_{2A} = 1{,}95$ *ja* $k^e_{2B} = 1{,}85$*. Kui mängija soovib panustada sündmuse A toimumisele, siis peaks ta valima teise kihlveokontori koefitsiendi* $k^e_{2A} = 1{,}95 > 1{,}85 = k^e_{1A}$*, sest nii saab ta võidu korral suurema kasu. Vastupidi, kui mängija soovib panustada sündmuse B toimumisele, siis peaks ta valima esimese kihlveokontori koefitsiendi* $k^e_{1B} = 1{,}95 > 1{,}85 = k^e_{1A}$*.*

1.4 EV-kontseptsioon

Eelnevalt leidsime lause (1.1) raames, et sobivas koguses panuseid saades võidab kihlveokontor hoolimata tegeliku sündmuse toimumisest vaheltkasu v suuruse summa. Uurime olukorda nüüd ka mängija perspektiivist lähtuvalt. Võrdlemaks mängija edu tegelikus panustamissituatsioonis, vaatleme oodatava tulu suurust kõigepealt eeldusel, et panustatakse kogumi juhuslikule sündmusele. Olgu kihlvedude kogum endiselt $N = 2$ võimaliku teineteist välistava sündmusega A ja B, mille koefitsiendid on k^i_A ja k^i_B ning tegelikud tõenäosused on p_A ja p_B. Kui mängija panustab juhuslikult, siis oodatav tulu on

$$\begin{aligned} E &= \frac{1}{2}(p_A k^i_A - (1 - p_A)) + \frac{1}{2}(p_B k^i_B - (1 - p_B)) \\ &= \frac{p_A k^e_A + p_B k^e_B}{2} - 1, \end{aligned}$$

[6]Ingl *juice* või *vigorish*

mis on positiivne vaid siis, kui $p_A k_A^e + p_B k_B^e > 2$. Üldiselt sõltub mängija oodatav tulu sündmuste tegelikest tõenäosustest. Siiski näeme ka patoloogilist juhtu: kui nii $k_A^e < 2$ ja $k_B^e < 2$, siis on juhuslikul panustamisel mängijal keskmiselt võimatu võita, hoolimata sellest, mis on sündmuste toimumise tegelikud tõenäosused. Osutub, et teatud eeldustel saame mängija oodatavat tulu paremini kvantifitseerida.

Lause 1.2. *Olgu kihlvedude kogumil 2 võimalikku teineteist välistavat sündmust A ja B koefitsientidega vastavalt k_A^e ja k_B^e. Olgu koefitsientide pealt leitud tõenäosused vastavalt p_A^* ja p_B^* ning olgu kihlveokontori liigprotsent $r \geq 0$. Eeldame, et vahendaja on oma liigprotsendi jaotanud tõenäosuste vahel proportsionaalselt tõenäosuste suurusega ja et vastava liigprotsendi osa mahaarvestamisel koefitsientide pealt leitud tõenäosustest saame sündmuste toimumise tegelikud tõenäosused p_A ja p_B. Siis on juhuslikul panustamisel mängija oodatav kulu võrdne kihlveokontori vaheltkasuga $v = \dfrac{r}{1+r}$.*

Tõestus. Leiame kõigepealt sündmuste toimumise tegelikud tõenäosused. Vastavalt lause eeldustele saame

$$o_A = \frac{p_A^*}{p_A^* + p_B^*} r, \quad o_B = \frac{p_B^*}{p_A^* + p_B^*} r,$$

millest

$$p_A = p_A^* - o_A = p_A^* - \frac{r}{1+r} p_A^*, \quad p_B = p_B^* - o_B = p_B^* - \frac{r}{1+r} p_B^*.$$

Mängija oodatav tulu avaldub nüüd

$$\begin{aligned}
E &= \frac{1}{2}(p_A k_A^i - p_B) + \frac{1}{2}(p_B k_B^i - p_A) \\
&= \frac{1}{2}(p_A(k_A^i - 1) + p_B(k_B^i - 1)) \\
&= \frac{1}{2}\left(1 - \frac{1}{1+r}\right)\left(\frac{k_A^i - 1}{k_A^e} + \frac{k_B^i - 1}{k_B^e}\right) \\
&= \frac{1}{2(1+r)}\left(\frac{k_A^e - 2}{k_A^e} + \frac{k_B^e - 2}{k_B^e}\right) \\
&= -\frac{1}{2(1+r)}\left(-2 + \frac{2}{k_A^e} + \frac{2}{k_B^e}\right) \\
&= -\frac{1}{1+r}(1 + r - 1) = -\frac{r}{1+r} = -v,
\end{aligned}$$

millega ongi lause tõestatud. □

Uurime nüüd, millistes situatsioonides peaks mängija kihlvedusid sõlmima, millistes mitte. Osutub, et sugugi ei ole tarvis tingimata panustada sündmusele, mis toimub suurima tõenäosusega.

Definitsioon 1.5. *Vaatleme panustamist sündmusele A koefitsiendiga k^i, st funktsiooni*

$$f(\xi,x) = \begin{cases} k^i x & kui\ \xi = 1 \\ -x & kui\ \xi = 0 \end{cases},$$

kus x on panuse suurus ja ξ on 1, kui A toimub, 0 muidu. Me ütleme, et kihlveos leidub väärtus, kui sündmusele panustades jääme keskmiselt kasumisse, st $E[f(\xi,x)] > 0$.

Omadus. *Kihlveos leidub mängija jaoks väärtus, kui sündmuse toimumise tegelik tõenäosus $p > \frac{1}{k^e} = p_A^*$.*

Tõestus. $E[f(\xi,x)] = pk^i x - (1-p)x = (p(k^i+1)-1)x$, aga $p(k^i+1) > 1$ vastavalt eeldusele, nii et $E[f(\xi,x)] > 0$. □

Tegelikult ei ole tarvis kihlveokontori pakutud koefitsiendi põhist tõenäosust arvutada. Kui tegelik tõenäosus on teada, piisab väärtuse olemasolu tuvastamiseks kontrollimaks, kas $pk^e > 1$. Ülaltoodud omadus pakub aga paremat interpretatsiooni: kui vahendaja sündmuse toimumise tegelikku tõenäosust alahindab, siis leidub kihlveos väärtus.

Me ütleme, et mängijal on *eelis*[7], kui ta leiab kihlveost väärtuse. Eelise suurus $e = pk^e - 1$. Mida suurem on eelis, seda rohkem võib mängija loota kihlveost keskmiselt võita, st keskmiselt võidab mängija oma eelise suuruse. Mängija peaks sõlmima ainult selliseid kihlvedusid, kust ta on leidnud väärtuse. Väärtuse kontseptsioon tähendab, et kihlvedusid sõlmides ei tarvitse panustada poolele, mis suurema tõenäosusega võidab, sest nii võib osa väärtust lauda jääda.

Näide 1.7. *Vaatleme kihlvedude kogumit sündmustega A ja B. Olgu kihlveokontori koefitsiendid sündmustele vastavalt $k_A^e = 5$ ja $k_B^e = 1{,}2$ ning tegelikud tõenäosused vastavalt $p_A = 0{,}22$ ja $p_B = 0{,}78$. Kui panustaksime sündmuse B toimumisele, kaotaksime iga panustatud ühiku pealt* 6,4%. *Panustades aga sündmuse A toimumisele, võidaksime iga ühiku pealt keskmiselt* 10% *kasumit.*

[7] Ingl *edge*

Panustamine väärtuse põhimõttest lähtuvalt on kesksel kohal ka edasises töös. Seetõttu ei piisa korvpallimängude ennustamisel üksnes võitja klassifitseerimisest, vaid oluline on hinnata mõlema meeskonna võidu tõenäosust. Kui saadavad hinnangud on tõele lähemal kui kihlveokontori poolt pakutud koefitsientidest arvata võiks, siis on lootust pikas perspektiivis edukas olla.

Tegelikult ei piirdu spordiennustuste tegemine ainult hea modelleerimisega. Väga tähtis on panustamisel ka investeeringu suuruse määramine, et ühelt poolt maksimiseerida kasumit ja teiselt poolt minimiseerida riski. Teatud juhtudel on võimalik panustada ka riskivabalt: kui leiduvad kihlveokontorid, mille pakutavate koefitsientide kombinatsioon katab kõik mängu sündmused nii, et koefitsientidelt tuletatud tõenäosuste summa on väiksem kui 1, siis on võimalik igale sündmusele õiges proportsioonis panustades kindlat kasumit teenida. Viimast nähtust nimetatakse *arbitraažiks*. Lisaks võib panustamisel lähtuda majandusteooriast: turu käitumisest, koefitsientide liikumisest, turu (in)efektiivsusest. Neid ja paljusid teisigi kontseptsioone ning võimalikke lähenemissuundi spordiennustustesse selles magistritöös ei käsitleta. Huvitatud lugeja võib tutvuda raamatutega [F$^+$10] ja [Buc03]. Esimene annab hea ülevaate erinevatest kihlveoliikidest, teises on Monte-Carlo meetoditega põhjalikult analüüsitud erinevaid panustamisskeeme.

Peatükk 2

Matemaatiline taust

Siin peatükis esitatakse terminoloogia, notatsioon ja matemaatilised kontseptsioonid, mida hilisemas töös kasutatakse. Kirjapanek püüab olla minimalistlik ja on seega ilma tõestusteta. Põhjalikuma tõenäosus- ja klassifitseerimisteoreetilise ning juhuslike protsesside alase käsitluse võib leida vastavalt raamatutest [Bil95], [HTF09] ja [Law95]. Eesti keeles on siin peatükis esitatud tulemusi käsitletud loengukonspekti vormis allikates [Lem12], [Lem13] ja [Kä11].

2.1 Tõenäosusteooria

Olgu $(\Omega, \mathcal{F}, \mathbf{P})$ tõenäosusruum, kus Ω on elementaarsündmuste ruum, $\mathcal{F}$ on σ-algebra ja $\mathbf{P}$ on tõenäosusmõõt. Hulga $\mathcal{F}$ elemente nimetame *sündmusteks* – need on hulgad, millele saame omistada tõenäosust. Edaspidi tegeleme palju suurustega, mille kohta ei ole teada nende kindlat väärtust, σ-algebra võimaldab meil nende väärtuste esinemise tõenäosust mõõta.

Me nimetame *juhuslikuks vektoriks* $\mathcal{F}$-mõõtuvat d-mõõtmelist funktsiooni $X : \Omega \to \mathbb{R}^d$, st originaalid $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ $\forall B \in \mathcal{B}(\mathbb{R})$ korral, kus $\mathcal{B}(\mathbb{R})$ on kõigi reaaltelje lahtiste hulkade poolt tekitatud Boreli σ-algebra. Kui $d = 1$, siis $\mathcal{F}$-mõõtuvat funktsiooni $X : \Omega \to \mathbb{R}$ nimetatakse *juhuslikuks suuruseks*. Definitsioon loob meile formaalse võimaluse mõõta vastavaid reaalteljel määratud hulki B mõõdu $\mathbf{P}$ abil, mis on defineeritud tõenäosusruumis. Selleks peame leidma suurima sellise hulga $A \in \mathcal{F}$, mille korral $X(A) \subset B$, ja saame arvutada $\mathbf{P}(A)$, mille väärtus omistataksegi hulgale B. Sisuliselt transformeeritakse reaalteljel defineeritud hulk B σ-algebrasse $\mathcal{F}$, kus see ära mõõdetakse. Protsessi võib vaadelda

kui mõõdu $\mathbf{P}X^{-1}$ rakendamist hulgale B, st

$$\mathbf{P}X^{-1}(B) = \mathbf{P}(\omega : X(\omega) \in B). \tag{2.1}$$

Mõõtu $\mathbf{P}X^{-1}$ nimetatakse juhusliku vektori *jaotuseks* ja tähistatakse P_X.

Juhusliku vektori jaotus on oluline, sest selle abil on võimalik X-i käitumist iseloomustada. Samas on see defineeritud iga Boreli hulga B jaoks, mis muudab mõiste abstraktseks. Ühise standardi juhuslike vektorite kirjeldamiseks loob *jaotusfunktsioon*: kui P_X on tõenäosusmõõt σ-algebral $\mathcal{B}(\mathbb{R})$, siis mõõdu P_X jaotusfunktsiooniks nimetatakse järgmist funktsiooni:

$$F : \mathbb{R}^d \to [0,1],$$

$$F(x_1, ..., x_d) = P_X((-\infty, x_1], ..., (-\infty, x_d]).$$

Seose 2.1 abil avaldub jaotuse P_X jaotusfunktsioon tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$ defineeritud mõõdu $\mathbf{P}$ kaudu kujul

$$F(x_1, ..., x_d) = \mathbf{P}(X_1 \leq x_1, ..., X_d \leq x_d),$$

mida nimetatakse ka juhusliku vektori X jaotusfunktsiooniks.

Näide 2.1. *Ülaltoodu loob meile korrektse raamistiku juhusliku vektori kirjeldamiseks. Vaadeldes näiteks mündiviset ja juhuslikku suurust*

$$X = \begin{cases} 1 & \textit{kui tuleb kull} \\ 0 & \textit{muidu} \end{cases},$$

siis nüüd saame leida tõenäosuse, et X-i väärtus on kas 1, 0, 1 või 0 või pole kumbki. Ausa mündi korral on X-i jaotus ja jaotusfunktsioon alljärgnevad:

$X(\omega)$	P_X
0	*0,5*
1	*0,5*

, $\qquad F(x) = \begin{cases} 0 & x < 0 \\ 0{,}5 & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}.$

Tuntumad jaotused on näiteks normaaljaotus, eksponentjaotus, Poissoni jaotus ja binoomjaotus. Tihtipeale me juhuslike suuruste tegelikku jaotusfunktsiooni aga ei tea, sellisel juhul saame jaotusfunktsiooni hinnata andmete pealt. Olgu $X_1, ..., X_n$ sõltumatud ja sama jaotusega (ssj) juhuslikud vektorid jaotusfunktsiooniga $F(x)$, siis *empiiriliseks jaotusfunktsiooniks*

nimetatakse funktsiooni

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \le x},$$

kus $I_{X_i \le x}$ on indikaatorfunktsioon, st

$$I_{X_i \le x} = \begin{cases} 1 & X_i(\omega) \le x \\ 0 & X_i(\omega) > x \end{cases}.$$

Sisuliselt on empiirilise jaotusfunktsiooni tähendus lihtne – iga x-i korral on $F_n(x)$ osakaal realisatsioonidest x_1,...,x_n, mis on x-ist väiksemad –, ent tulemus on võimas, sest Glivenko-Cantelli teoreemi kohaselt empriiline jaotusfunktsioon koondub ühtlaselt tegelikuks jaotusfunktsiooniks ja on seega heaks hinnanguks funktsioonile $F(x)$:

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \text{ p.k.} \tag{2.2}$$

Jaotuse kirjeldamisel on olulise tähtsusega statistikuteks *keskväärtus* EX ja *dispersioon* DX:

$$EX = \int_\Omega X d\mathbf{P}, \tag{2.3}$$

$$DX = E(X - EX)^2.$$

Tähtis on ka juhuslike vektorite sõltumatuse defineerimine. Intuitiivselt tähendab see seda, et ühe vaatluse all oleva suvalise juhusliku vektori väärtuse teadmine ei mõjuta kuidagi teiste juhuslike vektorite väärtust, ent formaalse kirjapaneku kohaselt on juhuslikud vektorid $X_1, ..., X_n$ sõltumatud parajasti siis, kui

$$F(x_1, ..., x_n) = F(x_1) \times ... \times F(x_n).$$

Viimaseid tulemusi teades saame kirja panna järgmise keskse tulemuse. Kui $X_1, ..., X_n$ on ssj juhuslikud vektorid keskväärtusega μ ja lõpliku dispersiooniga σ^2, siis *tsentraalse piirteoreemi* kohaselt toimub valimimahu kasvades jaotuse järgi koondumine

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \tag{2.4}$$

ehk

$$\frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

kusjuures *suurte arvude seaduse* kohaselt

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mu \text{ p.k.} \tag{2.5}$$

Seega võiks vähemalt suure valimimahu n korral juhusliku vektori X tegelikku keskmist hinnata valimikeskmise abil ilma eriliste süümepiinadeta. Viimased tulemused on olulised kogu statistikas ning on läbivalt kasutusel ka klassifitseerimisteoorias.

2.2 Klassifitseerimisteooria

Klassifitseerimisteooria on valdkond, mis tegeleb objektide kuuluvuse määratlemisega. Näiteks võivad objektideks olla hulk iirise taimi, mille kohta on kogutud järgmised andmed: kroonlehe pikkus ja laius ning tupplehe pikkus ja laius. Iga iiris on ka mingit kindlat liiki: *Setosa*, *Versicolor* või *Virginica*. Huvi pakkuvaks ülesandeks võib olla uute iiriste liigi ennustamine, kui teada on vaid mõõdetud tunnuste väärtused, kuid mitte liigiline kuuluvus. Sellist probleemipüstitust nimetatakse *klassifitseerimisülesandeks*. Defineerime seonduvad mõisted nüüd ka formaalselt.

Olgu $X = (X_1,...,X_d) \in \mathbb{R}^d$ d-dimensionaalne juhuslik vektor tõenäosusruumil $(\Omega, \mathcal{F}, \mathbf{P})$. Klassifitseerimisteoorias ütleme, et X on objekt, mida kirjeldavad tunnused X_i, $i = 1,...,d$. Iga objekt kuulub mingisse klassi, kusjuures kõikvõimalike klasside hulka tähistame $\mathcal{Y} = \{0, 1, ..., k-1\}$ (selline definitsioon võib nõuda klass-tunnuse kodeerimist: iiriste näites võiks *Setosa* olla kodeeritud 0-iks *Versicolor* 1-ks ja *Virginica* 2-ks). Nüüd saame defineerida *klassifitseerija* järgmise funktsioonina:

$$g : \mathbb{R}^d \to \mathcal{Y}. \tag{2.6}$$

On loomulik eeldada, et uued objektid, mille kuuluvust klassifitseerija abil määrata soovitakse, on juhuslikud. Lihtne on aga näha, et samuti ei saa fikseerituna käsitleda objekti klassi, sest see ei ole tunnuste poolt üheselt

määratud. Näiteks võivad nii *Virginica* kui *Versicolori* teatud isendid omada sama suuri kroon- ja tupplehti. Seega on klass juhuslik ja me tegeleme klassifitseerimisteoorias $(d+1)$-dimensionaalse juhusliku vektoriga (X,Y). Samuti järeldub, et eeskirjaga (2.6) esitatud determineeritud klassifitseerijad paratamatult teevad vigu. Klassifitseerijate headust on võimalik siiski mõõta, selleks defineerime *kaofunktsiooni* ja *riski* mõiste.

Kaofunktsiooniks nimetame funktsiooni

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+,$$

mis kirjeldab tekkivat kahju, kui lahterdada klassi i kuuluv objekt klassi j. Kaofunktsioon ei pea tingimata olema sümmeetriline, st klassi j kuuluva objekti klassifitseerimine klassi i võib tekitada erinevat kahju kui klassi i kuuluva objekti klassifitseerimine klassi j või klassi m kuuluva objekti klassifitseerimine klassi n. Käesolevas töös vaadeldakse aga ainult sümmeetrilist kaofunktsiooni, mis defineeritakse järgmiselt:

$$L(i,j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}.$$

Kaofunktsiooni abil esitub ka klassifitseerija *risk*. See on keskmine kahju, mida klassifitseerija teeb:

$$R(g) = E[L(Y, g(X))].$$

Sümmeetrilise kaofunktsiooni kasutamine võimaldab riski lihtsasti mõõdetavaks teha. Nimelt kehtib (2.3) abil

$$R(g) = \int_\Omega L(Y, g(X)) d\mathbf{P} = \int_\Omega I_{Y \neq g(X)} d\mathbf{P} = \mathbf{P}(Y \neq g(X)), \qquad (2.7)$$

st sümmeetrilise kaofunktsiooni korral on klassifitseerija risk klassifitseerimisvea tegemise tõenäosus.

Kui meil oleks (X,Y) jaotus teada, siis võiksime iga probleemi jaoks leida parima, st minimaalse riskiga klassifitseerija, mida nimetataks *Bayesi klassifitseerijaks*. Reaalsuses me funktsiooni $F(x,y)$ enamasti aga ei tea, mistõttu peaksime klassifitseerija hindama ssj valimi $\mathcal{D}_n = \{(x_1,y_1),...,(x_n,y_n)\}$ põhjal. Tehisõppe-alases kirjanduses nimetatakse sellist hindamisprotsessi

klassifitseerija treenimiseks[8], kusjuures valimit $\mathcal{D}_n$ nimetatakse *treeningandmeteks*. Valimil treenitud klassifitseerija esitub järgmiselt:

$$g_n : \mathcal{D}_n \times \mathbb{R}^d \to \mathcal{Y}. \tag{2.8}$$

Muidugi soovime endiselt leida võimalikult head klassifitseerijat. Siin võime võtta aluseks tulemuse (2.2) ning minimiseerida empiirilist riski. Seega eelistame tulemuse (2.7) kohaselt sellist klassifitseerijat, mis teeb treeningandmetel kõige vähem vigu. Tulemuste (2.4) ja (2.5) kohaselt võiks treeningvigade osakaalu kasutada ka riski hinnanguna.

Kuigi toodud teooria kohaselt võiks parima klassifitseerija valida treeningvea põhjal, siis tuleks märgata võimalikku probleemi. Kui klassifitseerija treenimiseks ja selle headuse testimiseks kasutatakse samu andmeid, siis saadud viga ei ole X-i jaotuse suhtes representatiivne. Täpsemalt öeldes viga alahinnatakse mingi nihke võrra, sest klassifitseerija on optimeeritud konkreetse valimi suhtes. Väga kompleksseid mudeleid kasutades võiksime treeningvea koguni nulli viia, kuid see oleks selge ülesobitamine. Meid ei huvita, et klassifitseerija maksimaalselt hästi ühe konkreetse juhusliku andmekomplekti peale sobituks, vaid et tema kirjeldusvõime oleks võimalikult hea üle kogu X-i jaotuse. Seetõttu peaksime klassifitseerijate headust mõõtma üle mingi teise ssj valimi, mida nimetame *testandmeteks*. Tüüpiliselt demonstreeritakse treeningandmetel riski alahindamise nähtust järgmises näites kirjeldatud lihtsa klassifitseerija abil, mis valib vaatlustele klassi treeningandmete enamuse põhjal.

Näide 2.2. *Vaatleme klassifitseerijat, mis valib klassi enamuse põhjal, st* $g(X) = I_{p_A \geq p_B}$*, kus* p_A *ja* p_B *on klasside A ja B osakaalud treeningandmete hulgas. Oletame seejuures, et objektid pärinevad jaotusest, kus klassid A ja B jagunevad võrdselt. Testandmete klassifitseerimisel eksime seega keskmiselt 50% juhtudest. Kui me hindaksime riski aga nende samade andmete peal, millel oma klassifitseerija treenisime, siis valimi suuruse* $n = 1$ *korral ei eksiks me kunagi,* $n = 2$ *korral paneksime alati vähemalt 1 täppi,* $n = 3$ *korral oleks suurim viga* $1/3$ *jne. Valimimahu kasvades läheneme küll tegelikule veale* 0,5*, ent päris kohale ei jõua kunagi. Treeningandmete peal riski hinnates võime seda alahinnata.*

Siiamaani oleme defineerinud ja käsitlenud klassifitseerijat traditsiooniliselt, st kui funktsiooni, mis seab objektidele vastavusse klassi. Tegelikult

[8]Protsessi nimetatakse ka *õpetajaga treenimiseks*, sest iga objekti kohta on tema klassiline kuuluvus teada

võib huvi pakkuda ka suvalise väljundi saamine lõigust [0,1], mis sisuliselt tähendaks tõenäosust, et objekt kuulub konkreetsesse klassi. Ülaltoodud teooria jääks samaks, kui defineerida 0,5-st suuremad väljundid klassi 1 ja ülejäänud klassi 0 (kahe klassi korral), ent klassi kuulumise tõenäosuse hinnang võimaldab hiljem tuvastada spordiennustustes väärtust.

2.3 Juhuslikud protsessid

Olgu t aeg ja X juhuslik suurus. *Juhuslikuks protsessiks* nimetatakse juhuslike suuruste jada $\{X_t : t \geq 0\}$. Sõltuvalt sellest, kas aeg t on diskreetne või pidev, eristatakse ka *diskreetse* ja *pideva ajaga* juhuslikku protsessi. Juhusliku suuruse X_t väärtust ajahetkel t nimetatakse *olekuks*; kõigi väärtuste hulka, mida juhuslik protsess mingil ajahetkel omada võib, nimetatakse *olekuruumiks*.

Tihti käsitletakse juhuslike protsesside teoorias *Markovi ahelaid*. Need on protsessid, mille tulevik sõltub igal ajahetkel vaid olevikust, aga mitte minevikus esinenud protsessi olekutest. Formaalselt, protsess $\{X_t : t \geq 0\}$ on Markovi ahel, kui kehtib

$$\mathbf{P}(X_{t+s} = j | X_t = i, X_u = x_u, 0 \leq u < t) = \mathbf{P}(X_{t+s} = j | X_t = i),$$

kus x_u on protsessi olek ajahetkel u.

Leidub nii diskreetse kui pideva ajaga Markovi ahelaid/protsesse. Tüüpiliseks pideva ajaga juhuslikuks protsessiks, mis on ühtlasi ka Markovi protsess, on *Poissoni protsess*. Juhuslikku protsessi $\{N_t : t \geq 0\}$ nimetatakse Poissoni protsessiks intensiivsusega λ, kui kehtivad järgmised tingimused:

(a) $N_0 = 0$;
(b) $\mathbf{P}(N_t = k) = \dfrac{\exp(-\lambda t(\lambda t)^k)}{k!}$, $k \geq 0$;
(c) $N_{t_2} - N_{t_1}$ ja $N_{t_4} - N_{t_3}$ on sõltumatud iga $t_4 > t_3 > t_2 > t_1 \geq 0$ korral,
(d) $N_t - N_s$ ja N_{t-s}, $t > s$, on samast (Poissoni) jaotusest.

Seega on Poissoni protsess loendav protsess, mille juurdekasvud on sõltumatud ja statsionaarsed. Saab näidata, et Poissoni protsessis on sündmustevahelised ajad eksponentjaotusest parameetriga $\dfrac{1}{\lambda}$. Kui intensiivsus λ püsib kogu protsessi vältel muutumatuna, siis on tegu *homogeense*, vastasel juhul *mittehomogeense* protsessiga.

Peatükk 3

Andmed

Magistritöös valminud analüüs on teostatud allikatest [NMV14] ja [Ltd14] pärit andmete põhjal. Andmetega toimetamisel on lähtutud Eesti Vabariigi autoriõiguse seaduse ([Tea14]) peatükist VIII[1], mis lubab avalikke andmebaase õppe- ja teadusliku uurimistöö eesmärgil kasutada (vt lisa B).

Analüüsiks oluliste andmete kogumiseks on programmeerimiskeeles *Python* implementeeritud *veebirobot ehk -ämblik*[9]. Ämbliku ehitamisel on kasutatud *Selenium WebDriveri* liidest, mis võimaldab programmikoodis veebibrauserit kontrollida: sellele käsklusi saata ja tulemusi vastu võtta. Rakendust leiab raamita veebilehitseja *PhantomJS*, mis erineb tuntud brauseritest nagu Mozilla Firefox ja Google Chrome selle poolest, et töötab tagaplaanil, kasutajale veebilehti kujutamata, ja peaks seetõttu võimaldama ämblikul kiiremini tegutseda.

Tüüpiliselt ei saa või pole mugav tooreid andmeid koheselt analüüsiks kasutada: need tuleb eelnevalt korrastada, sobivalt esitada või organiseerida ning nende valiidsuses ehk korrektsuses peab veenduma. Andmetega töötamise hõlbustamiseks on nendest moodustatud relatsiooniline andmebaas, mille haldamiseks kasutatakse andmebaasi juhtimissüsteemi *SQLite*. Viimast eelistatakse selle kasutusmugavuse tõttu, sest erinevalt paljudest teistest juhtimissüsteemidest ei vaja SQLite severit ega seadistamist. Kogu andmebaas sisaldub ühes failis, mida on teistesse protsessidesse lihtne kaasata. Lihtsus toob paratamatult kaasa teatud piiranguid, nt ei ole SQLite'is võimalik kirjutada protseduure ehk funktsioone. Vastav otstarve on realiseeritud programmeerimiskeeles *R*, kus teostatakse ka andmete analüüs.

Kasutada olevaid andmeid ja nendevahelisi seoseid illustreerib lisas C

[9] Ingl *bot, web spider, web crawler*

toodud andmebaasi skeem. Analüüsi jaoks olulistest andmetest anname ülevaate alljärgnevates peatükkides.

3.1 NBA mängud

Analüüsis kasutatakse andmeid korvpalliliiga NBA mängudelt 13 erineval hooajal, alates 2000-01 kuni 2012-13. Iga NBA hooaja mängud võib jaotada kaheks osaks: põhihooaja mängudeks ja *playoff*-mängudeks. Ühe tervikliku põhihooaja jooksul mängib iga meeskond 82 mängu, millest pooled kodustaadionil ja pooled võõrsil, kusjuures kõik meeskonnad kohtuvad omavahel vähemalt 2 korda. Käesolevas magistritöös vaadeldakse vaid põhihooaja mänge, mida on kõigi nimetatud hooaegade peale kokku 15585[10].

3.1.1 Koondandmed[11]

Iga korvpallimäng on mõlema meeskonna kohta teada järgmised kokkuvõtlikud andmed:

- FGA - väljakult sooritatud visete arv,
- FGM - väljakult tabatud visete arv,
- 3FGA - sooritatud 3-punkti visete arv,
- 3FGM - tabatud 3-punkti visete arv,
- FTA - sooritatud vabavisete arv,
- FTM - tabatud vabavisete arv,
- OREB - ründelauast püütud pallid arv,
- DREB - kaitselauast püütud pallid arv,
- AST - resultatiivsete söötude arv,
- TOV - pallikaotuste arv,
- STL - vaheltlõigete arv,
- BLK - blokeeritud vastasmeeskonna visete arv,
- PF - tehtud vigade arv,
- PTS - visatud punktide arv,
- MIN - mänguminutite arv.

[10]Kuni hooajani 2003-04 osales korvpalliliigas NBA 29 meeskonda, alates hooajast 2004-05 aga 30 meeskonda. See teeks kokku $4 \times 1189 + 9 \times 1230 = 15826$ mängu, ent hooaeg 2011-12 oli lühendatud (toimus vaid 990 mängu) ja hooajal 2012-13 jäi 1 mäng ära.

[11]Ingl *Box Scores*

Korvpallimängudes peab alati selguma võitja, viiki esineda ei saa. Kui kaks meeskonda on normaalaja lõpuks visanud võrdse arvu punkte, siis mängitakse 5-minutilisi lisaaegu kuni võitja selgumiseni. Seega tuleb arvestada, et mängud võivad olla erineva pikkusega, mistõttu on tunnused võrreldavuse huvides tarvis normeerida. Ülaltoodud tunnuste pealt on võimalik tuletada ka mitmeid teisi tunnuseid, mis võivad korvpallitulemuste kirjeldamisel olulised olla, nt meeskonna võitude arv või keskmine visatud punktide arv viimasest x mängust.

3.1.2 Sündmus-sündmus andmed[12]

Koondandmed annavad palju informatsiooni, mida saab ennustamises kasutada, ent need on siiski vaid mängu kokkuvõtlikud statistikud, mis ei pruugi piisavalt hästi edastada mängu kulgu. Ühte korvpallikohtumist võib vaadelda kui hulka sündmusi ning detailne informatsioon nende sündmuste kohta võib aidata avastada seoseid, mida koondandmetest leida ei õnnestuks. Näiteks võib sündmustepõhistest andmetest leida järgmist informatsiooni:

- meeskondade keskmine rünnakute lõpetamise kiirus,
- detailne informatsioon sooritatud visete tüüpide kohta,
- mängu dünaamika muutus mängu lõppfaasis,
- igal ajahetkel väljakul olevad mängijad.

Sündmus-sündmus andmed on allikas [NMV14] esitatud tekstina. Olulise info kaevandamiseks relatsioonilistesse tabelitesse on Pythonis iga mängu kohta realiseeritud simulatsioon, mis teksti kujul andmetest paljude regulaaravaldiste toel olulise informatsiooni eraldab. Näiteks saab mängusituatsioonist

Thompson 24’ 3PT Jump Shot (3 PTS) (Barnes 1 AST)

informatsiooni esiteks selle kohta, et Thompson on 24’ kauguselt tabanud 3-punkti viske ja teiseks, et Barnes on andnud selleks resultatiivse söödu. Korvpallimängijatele on toorandmetes enamasti viidatud perekonnanime

[12]Ingl *Play-by-play data*

abil: kui ühes mängus on ühes meeskonnas olnud 2 sama perenimega korvpallurit, st õiget mängijat ei ole võimalik tuvastada, siis on andmed esitatud meeskonna täpsusega (alati on teada vähemalt see, millise meeskonna mängijaga sündmus toimus). Eraldatud informatsioonist saab detailse ülevaate lisas (C) toodud andmebaasi skeemist.

3.2 Mängude koefitsiendid

Koefitsiendid on saadaval 5 hooaja kohta: 2008-09 kuni 2012-13. Iga korvpallimängu mõlema meeskonna kohta on teada kuni 10 erineva kihlveokontori *alustavad* ja *sulgevad* koefitsiendid. Esimesel juhul on tegemist hindadega, millega vahendaja panuste vastuvõtmist alustab, teisel juhul aga hindadega, millega panuste vastuvõtmine lõpeb. Seega võib alustavat koefitsienti käsitleda kui kihlveokontori ennustust mängu tulemusele, sulgevas koefitsiendis kajastub lisaks vahendaja arvamusele ka mängijate vastavasisuline hinnang. Teatud juhtudel on alustava ja sulgeva koefitsiendi asemel vaid üksainus – sellist koefitsienti käsitletakse sõltuvalt vajadusele mõlemat pidi.

Magistritöös eeldame, et mängijal ei ole võimalik kõikide vahendajate juures korraga panustada. See on loomulik eeldus mitmel põhjusel: paljude kihlveokontorite juures panustamisfinantside omamine nõuab mängijalt suures koguses ressursse, finantside vahetamine vahendajate vahel võtab aega ning ei pruugi olla tasuta, kõik kihlveokontorid ei ole võrdselt usaldusväärsed ega paku parimaid hindu. Eeldame, et mängijal on võimalik panustada tabelis 3.1 esitatud 5 kihlveokontori juures, mida allikas [Spo14] on hinnanud skaalal F kuni A+ vähemalt B-ga.

Tabel 3.1: Usaldusväärsemad kihlveokontorid, nende poolt erinevatele mängudele pakutud koefitsientide arv ja viimastele lisatud keskmine liigprotsendi suurus.

Kihlveokontor	Hinnang	Koefe mängudele	Keskmine liigprotsent
Pinnacle Sports	A+	4989	2,42%
Bet365	A+	5257	4,14%
William Hill	A	3379	4,14%
MarathonBet	B	1217	1,57%
TitanBet	B	3811	4,64%

Tabelist ilmneb, et paljude mängude korral ei ole teada kõikide vahendajate koefitsiendid. Suurel määral see analüüsi käiku ei mõjuta, kesisem valik piirab üksnes väärtuse leidmise võimalusi. Mäng loetakse koefitsienti omavaks, kui sellele on teada vähemalt 1 vahendaja koefitsiendid. Perioodil 2008-09 kuni 2012-13 on selliseid mänge kokku 5331. Ühtlasi näeme tabelist vahendajate keskmist liigprotsenti, mida oma pakutud koefitsientidele lisatakse: Pinnacle Sports ja MarathonBet lisavad koefitsientidele keskmiselt vähem üleliigseid protsente[13], vastavalt 1,60% ja 2,44%, ülejäänud vahendajate keskmine liigprotsent on üle 4%. Kui arvestada ainult igale mängule pandud parimaid koefitsiente, siis keskmine liigprotsent üle kõikide mängude on 1,79%. See on suurem kui MarathonBeti liigprotsent, kuid selle vahendaja kohta on teada vaid ühe hooaja mängude koefitsiendid, mistõttu keskmine liigprotsent on rohkem mõjutatud teiste kihlveokontorite koefitsientidest.

Üldiselt võiks rohkem huvi pakkuda panustamine kihlveokontorite alustavatele koefitsientidele, sest need tähistavad vahendaja esialgset pakkumist, mis võiks olla ebatäpsem kui sulgevad koefitsiendid, sest viimased on kujunenud paljude osapoolte arvamuse pealt. Samas ei saa püstitada eeldust, et kõik vahendajad alustasid koefitsientide pakkumist samal ajahetkel, seega uurime edaspidi sulgevaid koefitsiente.

[13]Ingl *reduced juice books*

Peatükk 4

Teoreetiline käsitlus

Käesolevas peatükis anname põgusa ülevaate viisidest, kuidas korvpallitulemuste ennustamisele on võimalik läheneda. Samuti kirjeldame meetodeid, millel hilisem analüüs tugineb.

Spordiennustusteks kasutatavad mudelid ja meetodid peavad lõpuks väljastama samal kujul informatsiooni – kes (ja millise tõenäosusega) võidab. Seega on eesmärk määrata mängu klassiline kuuluvus ning kõiki vastavaid süsteeme saame käsitleda klassifitseerijatena kujul (2.6). Kuigi eesmärk on sama, siis selleni jõudmiseks on palju erinevaid meetodeid. Kaks põhimõttelist lähenemist on reitingusüsteemid ja tehisõppe algoritmid.

4.1 Reitingusüsteemid

Reitingusüsteemi võib ette kujutada tabelina, kus mingi mõõdiku ehk *reitingu* alusel on järjestatud hulk objekte. Korvpallialases käsitluses on objektideks enamasti meeskonnad ja modelleerija põhiülesandeks jääb reitingu koostamine. Reitingut uuendatakse iga mäng ja mängu võitja valitakse vastavalt reitingu suurusele. Tuntud reitingusüsteem on malest pärit ELO-süsteem, mille põhimõtteid on üle kantud ka teistele spordialadele.

ELO meetodi korral meeskondade reitingud esmalt algväärtustatakse. Edaspidi muutuvad need iga mäng sõltuvalt sellest, kuivõrd oli mängu tulemus ootuspärane. Olgu korvpallimeeskonnad i ja j, olgu $A = \{i \text{ võitis } j\}$ ning tähistagu $p(A|\cdot)$ sündmuse A toimumise tinglikku tõenäosust. Siis arvutatakse meeskonna i uus ELO-reiting r_i^{uus} vana reitingu r_i^{vana} abil

$$r_i^{uus} = r_i^{vana} + K(I_A - p(A|r_i^{vana}, r_j^{vana})),$$

kus K on normeeriv konstant ja suurust $p(A|\cdot)$ võib hinnata nt logistilise funktsiooni abil argumendist $x_{ij} = const(r_i^{vana} - r_j^{vana})$:

$$L(x_{ij}) = \frac{1}{1 + e^{-x_{ij}}}.$$

Siit ilmneb ka süsteemi võlu: meeskondade reitingute muutus sõltub mängu tulemuse tõenäolisusest, st tugevad meeskonnad, kelle võit on ootuspärane, võivad reitingus tõusta vähem ja langeda rohkem kui meeskonnad, kelle võit on vähem tõenäoline.

Lisaks ELO-süsteemile on populaarsed ka Massey meetod ja Markovi meetod. Massey reitingusüsteem põhineb lineaaralgebral: meeskondade reitingud leitakse teatud lineaarvõrrandisüsteemi lahendamise teel. Markovi reitingusüsteem tugineb juhuslike protsesside teooriale: meeskondade vahel defineeritakse üleminekutõenäosused, misjärel teostatakse juhuslik ekslemine kuni tasakaaluoleku saabumiseni, kusjuures meeskonnad järjestatakse tasakaalutõenäosuste järgi.

Reitingusüsteemid on küll populaarsed ja põhinevad mitmekülgsel matemaatikal, ent nendest võib olla keeruline mängu võitjate tõenäosusi tuletada. Käesoleva töö analüüsis reitingusüsteeme ei puudutata, mistõttu pikemalt nende olemust siinkohal ei avata. Toodud meetoditest ja veel paljudest teistest süsteemidest saab matemaatilise ülevaate allikast [LM12]. Korvpallitulemuste (NCAA[14]) ennustamiseks on Markovi ahelal põhinevat reitingusüsteemi kasutatud artiklis [KS06], kus iga meeskond on defineeritud kui üks olek (olekuruumi moodustavad kõik korvpallimeeskonnad). Olekute vahel on leitud üleminekutõenäosused, misjärel on ahelal teostatud juhuslik ekslemine kuni tasakaaluoleku saabumiseni. Meeskondade järjestus ongi leitud vastavate tasakaalutõenäosuste põhjal.

4.2 Tehisõpe

Masin- või tehisõpe hõlmab endas meetodeid, millega andmete pealt õppida. Sporditulemuste ennustamisel kasutatakse selleks peatükis (2.2) mainitud õpetajaga õppimist. Erinevalt reitingusüsteemidest ei põhine õppimine meeskondade järjestamisel, vaid keskendub rohkem peaeesmärgile

[14] *National Collegiate Athletic Association*

ehk prognoosimisele. Võimalikke masinõppe algoritme, mida klassifitseerijana (2.6) kasutada, on väga palju. Peatükis 5.2 kasutatakse NBA korvpallimängude ennustamiseks logistilist regressiooni ja AdaBoosti. Esimene neist on lineaarne klassifitseerija, ja et lineaarsed klassifitseerijad on tehisõppes väga olulise tähtsusega – nende erinevatel vormidel põhinevad ka keerulisemad meetodid –, anname nende näitel klassifitseerimisteooriast lühikese ülevaate. Järgnevas eeldame endiselt, et erinevate klasside arv on kaks. Samuti lähtume tulemuste esitamisel eeldusest, et meil on kasutada treeningvalim $\mathcal{D}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$, kus $\forall i : x_i \in \mathbb{R}^d$ on tunnuste vektor ja y_i on vastav klass.

•

Lineaarne klassifitseerija püüab objekti tunnused eraldada hüpertasandi abil: hüpertasandi ühele poole jäävad objektid klassifitseeritakse ühte klassi, teisele poole jäävad objektid klassifitseeritakse teise klassi. Formaalselt: olgu $x \in \mathbb{R}^d$ tunnuste vektor ja $(w^T, w_0) \in \mathbb{R}^{d+1}$ kaalude vektor, siis esitub lineaarne klassifitseerija kujul

$$g(x) = \begin{cases} 1 & w^T x + w_0 \geq 0.5 \\ 0 & w^T x + w_0 < 0.5 \end{cases}$$

Ükski mõistlik klassifitseerija ei saa sõltuda klasside kodeeringust, seetõttu eeldame hetkeks, et klassid on defineeritud kui +1 ja -1. Siis saame lineaarse klassifitseerija kirjutada lihtsamalt:

$$g(x) = \text{sign}(w^T x + w_0). \tag{4.1}$$

Meenutame, et klassifitseerimisel soovime minimiseerida riski. Seega tahame funktsioonis (4.1) leida sellised kaalud (w^T, w_0), millele vastav klassifitseerija oleks teatud mõttes parim. Matemaatiliselt on keeruline minimiseerida treeningvigade arvu, st lahendada ülesannet

$$\min_{w^T, w_0} \sum_{i=1}^{n} I_{g_n(x_i) \neq y_i},$$

mistõttu kasutatakse alternatiivseid lahendusi. Üheks võimaluseks on leida kaalud nii nagu lineaarse regressiooni korral, minimiseerides jääkide

ruutude summat[15]

$$\min_{w^T, w_0} \sum_{i=1}^{n} (y_i - g_n(x_i))^2.$$

Samas on klassifitseerimine ja regressioon põhimõtteliselt võrdlemisi erinevad ülesanded, sest klassid on nominaalsed. Jääkide ruutude summa alusel minimiseerimine käsitleb klasse pidevatena ning selliselt leitud lahend on mõjutatud erinditest. Seetõttu ei pruugi vastavalt leitud klassifitseerija suuta objekte eraldada, kuigi tegelikkuses on need lineaarselt eraldatavad (vt joonisel (4.1) osa (b)).

Alternatiivne meetod püüab minimiseerida valesti klassifitseeritud objektide kaugust hüpertasandist, st vaadelda ülesannet

$$\min_{w^T, w_0} \sum_{i=1}^{n} y_i g_n(x_i) I_{y_i g_n(x_i) < 0},$$

mille lahendamiseks saab kasutada numbrilisi meetodeid. Tüüpiliselt on nendeks *gradientmeetodite* nime kandvad algoritmide variatsioonid, mille korral juhuslikust algväärtusest alustades püütakse iteratiivselt liikuda globaalse miinimumi poole. Selleks leitakse igal sammul valesti klassifitseeritud punktides tuletis

$$\frac{d\left(\sum_{i=1}^{n} y_i g_n(x_i) I_{y_i g_n(x_i) < 0}\right)}{d(w^T, w_0)},$$

ja liigutakse selle vastassuunas ehk risti vastupidi funktsiooni maksimaalsele kasvamissuunale. Saab näidata, et kui klassid on lineaarselt eraldatavad, siis antud algoritm koondub lõpliku arvu sammude jooksul [HTF09, lk 131]. Selliselt leitud lineaarset klassifitseerijat nimetatakse *pertseptroniks*[16], mille kombineerimisel põhinevad nt populaarsed närvivõrkude klassifitseerimisalgoritmid.

Paneme tähele, et gradientmeetodite algoritmid alustavad juhuslikust punktist, mistõttu on koondumise korral ka lõpplahend juhuslik. Optimaalne oleks valida selline klasse lineaarselt eraldav hüpertasand, mis oleks mõlema klassi lähimast objektist sama kaugel[17]. Sellel ideel põhinevad nt

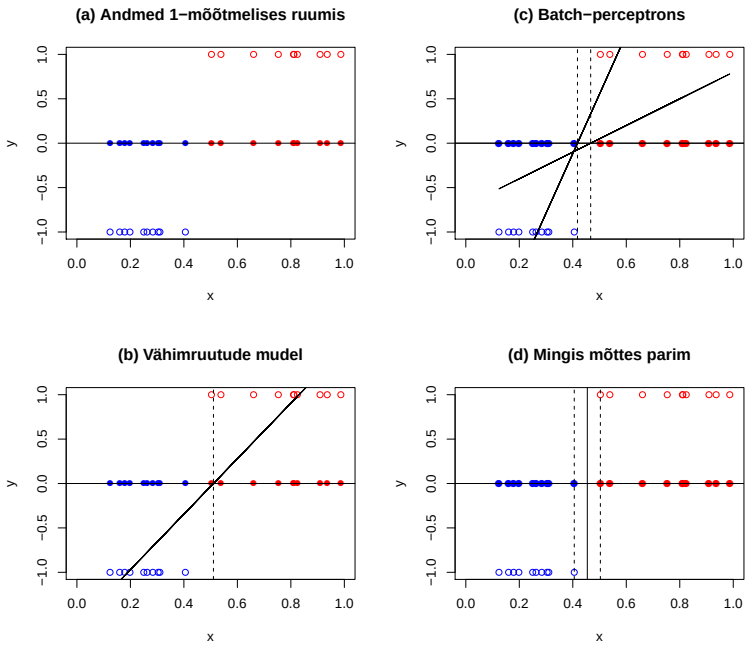
[15] Ka jääkide ruutude summat minimiseeriv klassifitseerija ei sõltu klasside kodeeringust [Lem13].

[16] Ingl *perceptron*

[17] Sellist kaugust nimetatakse ingl *margin*

väga populaarsed tugivektormasinad.

Vaatleme ülal esitatud teooriat lihtsal ühedimensionaalsel juhul. Joonisel (4.1) on näidatud erinevate lineaarsete klassifitseerijate tööd olukorras, kus klassid on lineaarselt eraldatavad. Selleks on ühtlasest jaotusest $U(0,1)$ genereeritud 20 juhuslikku suurust ning jaotatud need klassidesse vastavalt sellele, kas väärtused on suuremad või väiksemad kui 0,5. Joonise osast (b) näeme, et vähimruutude meetodiga leitud lineaarne klassifitseerija eksib ka sellise lihtsa näite korral ühe punkti klassifitseerimisel. Osal (c) on implementeeritud *gradient descent* algoritm ning näidatud leitud klassifitseerijad 2 erineva algväärtuse korral: mõlemal juhul suudab leitud pertseptron klassid eraldada. Osa (d) näitab lineaarset klassifitseerijat, mis võiks sellise treeningvalimi korral parim olla.



Joonis 4.1: (a) Ühtlasest jaotusest $U(0,1)$ genereeritud andmed: punase klassi (+1) moodustavad väärtused, mis on suuremad kui 0,5, sinise (-1) ülejäänud. (b) Vähimruutude mudel eksib ühe vaatluse klassifitseerimisel, kuigi klassid on lineaarselt eraldatavad. (c) Kaks pertseptroni, mille leidmiseks on *gradient descent* algoritmi alustatud erinevate algväärtuste korral. Mõlemad suudavad klassid eraldada. (d) Antud olukorras teatud mõttes parim lineaarne klassifitseerija, mis lisaks treeningvea minimiseerimisele maksimiseerib lähima punkti kauguse eraldavast hüpertasandist.

Võimalikke meetodeid on palju ning huvitatud lugejat suunatakse taas allikatele [HTF09] ja [Lem13]. Paljusid tehisõppe meetodeid on NBA korvpallimängude ennustamisel kasutatud artiklis [MGKK10] ja magistritöös [Pur13]. Esimesel juhul saavutati parimaid tulemusi Naïve Bayesi klassifitseerijaga, mis põhineb Bayesi valemi rakendamisel – korrektne võitja suudeti ennustada 67% mängudele. Teisel juhul on paremaid tulemusi saavutatud tugivektormasinatega, millega on mõnel hooajal korrektselt ennustatud ka üle 70% mängudest. Suur osa tööst [Pur13] on keskendunud oluliste tunnuste leidmisele ja defineerimisele, samas kui artiklis [MGKK10] on tunnustena kasutatud vaid kokkuvõtlikke statistikuid. Kummaski allikas ei ole uuritud meetoditega panustamise kasumlikkust.

4.2.1 Logistiline regressioon

Suurus $w^T x + w_0$ võib võtta ka negatiivseid väärtusi, mistõttu tõenäosuse hinnangute saamiseks ei pruugi ülaltoodud meetoditega leitud lineaarsed klassifitseerija hästi sobida. Logistiline regressioon on selles osas parem, sest sobitab lineaarse funktsiooni klassitõenäosuste suhte logaritmile, mis tagab hinnangute tõenäosusteoreetilise korrektsuse. Vaatleme klasse 1 ja 0 ning olgu $p_1(x) = p(1|x)$ ja $p_0(x) = p(0|x)$ tinglikud tõenäosused, et sisendi x korral on objekti klass vastavalt 1 ja 0. Logistilise regressiooni mudel on siis

$$\ln\left(\frac{p_1(x)}{p_0(x)}\right) = w^T x + w_0, \tag{4.2}$$

millest

$$p_1(x) = \frac{e^{w^T x + w_0}}{1 + e^{w^T x + w_0}}, \qquad p_0(x) = 1 - p_1(x)$$

ja valimi peal treenitud klassifitseerija (2.8) on

$$g_n(x) = \begin{cases} 1 & \hat{p}_1(x) \geq \hat{p}_0(x) \\ 0 & \hat{p}_1(x) < \hat{p}_0(x) \end{cases} = \begin{cases} 1 & \hat{w}^T x + \hat{w}_0 \geq 0 \\ 0 & \hat{w}^T x + \hat{w}_0 < 0 \end{cases}, \tag{4.3}$$

millest järeldub, et logistiline regressioon on samuti lineaarne klassifitseerija. Siin leitakse hinnangud $(\hat{w}^T, \hat{w}_0)$ enamasti logaritmilist tinglikku tõepärafunktsiooni maksimiseerides, st lahendades ülesannet

$$\max_{w, w_0} \ln\left(\prod_{i=1}^{n} \hat{p}_1^{y_i}(x_i)\hat{p}_2^{1-y_i}(x_i)\right),$$

mis lihtsustub kujule

$$\max_{w,w_0} \sum_{i=1}^{n} \left[y_i \ln \hat{p}_1(x_i) + (1 - y_i) \ln \hat{p}_2(x_i) \right].$$

Lineaarse mudeli kasutamine on antud juhul õigustatud, sest ei ole põhjust arvata, et korvpallimängude võitja jaotus järgib väga spetsiifilist struktuuri (võrdle nt malelaua jaotusest genereeritud andmetega), mille korral lineaarsed mudelid hästi töötada ei saa. Kui treeningvalim on representatiivne, siis ei kaasne lihtsa lineaarse mudeliga ka suurt ülesobitamise ohtu, st treeningvalimil leitud mudel peaks hästi sobima ka testandmetele. Logistiline regression on antud juhul sobilik lineaarne meetod, sest modelleerib otse tõenäosusi.

4.2.2 AdaBoost

AdaBoost[18] on komitee-meetod[19], mis klassifitseerib paljude teiste (enamasti lihtsate) klassifitseerijate abil, agregeerides nende prognoosid. Täpsemalt, olgu g^m nõrk (või lihtne) klassifitseerija, mis on parem kui juhuslik pakkumine, st suudab korrektselt klassifitseerida üle 50% vaatlustest. Olgu α_m selle klassifitseerija kaal. Siis avaldub AdaBoosti korral funktsioon (2.6) kujul (klassid on kodeeritud kui +1 ja -1)

$$g(x) = \text{sign} \left(\sum_{m=1}^{M} \alpha_m g^m(x) \right). \tag{4.4}$$

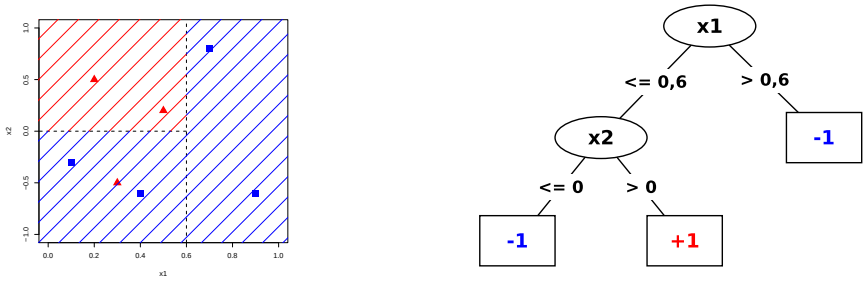
Klassifitseerija g^m pärineb eelnevalt määratletud klassifitseerijate hulgast, mis tihtipeale koosneb *otsustuspuudest*[20]. Viimased on meetodid, mis klassifitseerivad objekte reeglite abil. Iga reegel määrab tunnuste ruumis tüki, kuhu kuuluvad vaatlused klassifitseeritakse enamushääletuse teel. Seejuures on kõik tükid lõikumatud, nii et otsustuspuud võib esitada kujul

$$g^m(x) = \sum_{t=1}^{T} y_t I_{x \in R_t}, \tag{4.5}$$

kus T on reeglite arv, R_1,...,R_T on vastavad tükid ning y_t on klass. Reegleid

[18] *Adaptive Boosting*
[19] Ingl *ensemble method*
[20] Ingl *decision trees*

moodustatakse rekursiivselt treeningandmeid lõigates, kusjuures iga lõikamine peab mingis mõttes optimaalne olema. Joonisel (4.2) on näidatud lihtne 3 reegliga otsustuspuu.



Joonis 4.2: Otsustuspuu 2 lõikamise ja 3 reegliga, mis eksib ühe treeningpunkti klassifitseerimisel. Puu reeglid on järgmised:

$x_1 \leq 0{,}6$ ja $x_2 \leq 0 \implies -1$, $x_1 \leq 0{,}6$ ja $x_2 > 0 \implies +1$, $x_1 > 0{,}6 \implies -1$.

Tõenäosuste saamiseks võib valemis (4.5) diskreetse klassi väärtuse asemel väljastada osakaalu, et vastavasse tükki kuuluvate objektide klass on +1. Valemis (4.4) saab neid tõenäosuseid siis kombineerida ja normeerida.

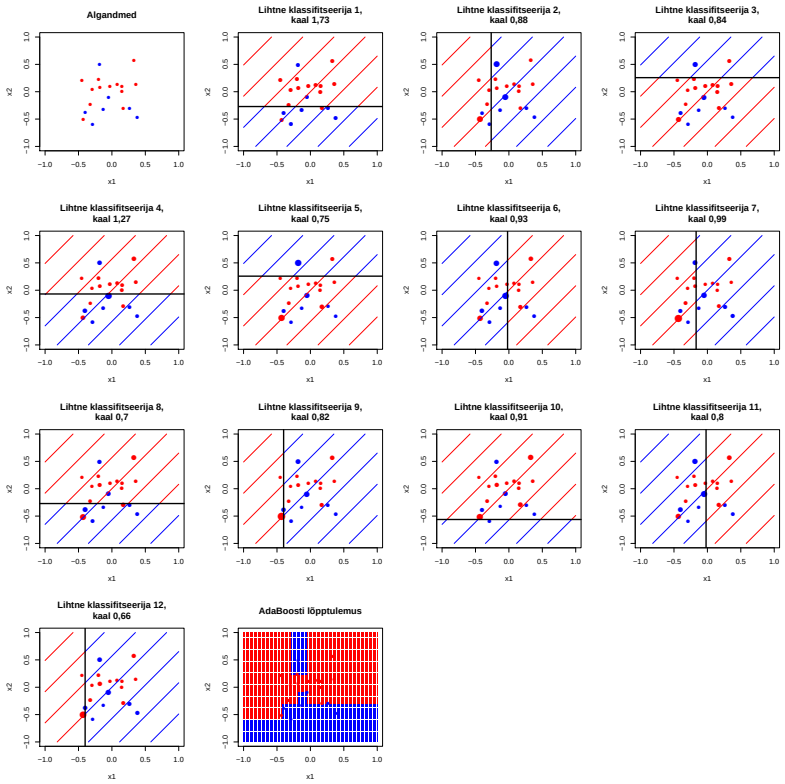
AdaBoosti korral tegeleb iga lihtne klassifitseerija küll samade, kuid erinevalt kaalutud vaatlustega. Täpsemalt, olgu β_i^m vaatluse (x_i, y_i) kaal, $i = 1, ..., n$. Siis püüab g^m minimiseerida kaalutud riski ehk empiirilise riski korral avaldub

$$g^m = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \beta_i^m I_{g(x_i) \neq y_i},$$

kus $\mathcal{G}$ on lihtsate klassifitseerijate hulk. Meetod töötab iteratiivselt, kusjuures tulemuses (4.4) toodud M on kogu iteratsioonide arv ja m tähistab seega konkreetset iteratsiooni. Esmalt antakse igale valimiobjektile võrdne kaal, st $\beta_i^1 = 1/n \ \forall i$. Seejärel treenitakse vaatlusi nõrga klassifitseerijaga g^1, misjärel kaalutakse vaatlused ümber vastavalt sellele, kas objektile omistatud klass oli korrektne: valesti klassifitseeritud vaatluste kaal suureneb, õigesti klassifitseeritud vaatluste kaal väheneb. Kaalude selline muutmine tähendab, et järgmisel iteratsioonisammul keskendutakse keeruliste objektide klassifitseerimisele rohkem. Nõrga klassifitseerija kaal α_1 sõltub sellest, milline oli vastav kaalutud risk: hästi töötav klassifitseerija saab suurema kaalu kui kehvem klassifitseerija. Vaatluste ümberkaalumist ja nõrkade meetoditega klassifitseerimist jätkatakse kuni määratud iteratsioonisammude lõpuni või kuni mõne muu tingimuse täitumiseni. Nt on

sobilik meetodi töö lõpetada, kui saavutatud komitee suudab kõik treeningandmed korrektselt klassifitseerida. AdaBoost on seega algoritmiline meetod; pseudokood on toodud algoritmis (1) [CJM06], mille väljundi abil klassifitseeritakse tulemuses (4.4) näidatud viisil.

AdaBoosti algoritmi tööd on visualiseeritud joonisel (4.3). Selleks on algoritmi rakendatud kahemõõtmelisest normaaljaotusest genereeritud 20 punktile, millele on klassid leitud juhuslikult. Veendume, et AdaBoost on võimeline treeningvalimi piisavalt suure hulga iteratsioonide korral "meelde jätma".



Joonis 4.3: AdaBoost rakendatuna kahemõõtmelisest normaaljaotusest genereeritud juhuslikult klassifitseeritud andmetele. Lihtsa klassifitseerijana on kasutatud otsustuspuud 1 lõikamise ja 2 reegliga (joonistel tähistatud musta sirgega): igal sammul on proovitud kõiki $2 \times 20 = 40$ erinevat tulemust andvat klassifitseerijat, mille hulgast on valitud parim. Punktide suurus näitab igal sammul objekti kaalu, viirutatud ala värv näitab, millisesse klassi lihtne klassifitseerija andmed igal iteratsioonil klassifitseeris. Lõpptulemus on lihtsate klassifitseerijate kombinatsioon. Näeme, et AdaBoost eraldab klassid antud andmetel 12 sammuga.

Algoritm 1 AdaBoost

1: $\beta_i^1 = \frac{1}{n}, \quad i = 1,...,n$ ▷ Algväärtustatud kaalud
2: **Iga** m = 1 kuni M:
3: $\hat{y}_i \leftarrow g^m(x_i), \quad i = 1,...,n$ ▷ Parim nõrk klassifitseerija
4: $\epsilon \leftarrow \sum_{i=1}^n \beta_i (y_i \neq \hat{y}_i)$ ▷ Kaalutud klassifitseerimisviga
5: $\alpha_m \leftarrow \ln\left(\frac{1-\epsilon}{\epsilon}\right)$ ▷ Nõrga klassifitseerija kaal
6: $\beta_i^{m+1} \leftarrow \exp(\alpha_m (y_i \neq \hat{y}_i))$ ▷ Uued vaatluste kaalud
7: $\beta_i^{m+1} \leftarrow \frac{\beta_i^{m+1}}{\sum_{i=1}^n \beta_i^{m+1}}$ ▷ Kaalude normeerimine
8: **Tagasta** $\{(\alpha_1, g^1),...,(\alpha_M, g^M)\}$

Ülaltoodust võib tekkida küsimus, miks mitte kasutada paljude lihtsate klassifitseerijate kombineerimise asemel ühte suurt puud. Sellel on erinevaid põhjendusi: puud moodustatakse ahnel viisil (iga lõikamispunkt valitakse selliselt, mis konkreetsel hetkel mingis mõttes parima klasside eralduvuse annab), mistõttu ei pruugi tulemus olla optimaalne; samuti on suure puu korral ülesobitamise oht. AdaBoostil seevastu on palju häid teoreetilisi omadusi. Näiteks minimiseerib see ahnel viisil empiirilist eksponentsiaalset riski ([Lem13])

$$\sum_{i=1}^n \exp(-y_i f(x_i)) = \sum_{i=1}^n \exp(-y_i(\alpha_1 g^1(x_i) + ... + \alpha_M g^M(x_i))) \tag{4.6}$$

üle kõikide $\alpha_1, ..., \alpha_M$ ja $g^1, ..., g^M$. Samuti annab AdaBoosti algoritm mittelineaarse klassifitseerija ning ülesobitamise oht ei suurene iteratsioonide arvu kasvades ehk mudeli komplekssuse suurenemisel üleliia kiiresti. Täpsemaid teoreetilisi tulemusi koos tõestustega saab vaadata allikatest [Lem13], [HTF09] ja [Sõ13].

4.3 Mängude modelleerimine

Korvpallimängu tulemus sõltub paljude korvpallimängijate kombineeritud sooritusest. Seega võivad võitja otsustamisel saada määravaks peened, abstraktsed detailid: kuidas tiimikaaslased omavahel interakteeruvad või vastasmeeskonna mängijate sooritusele mõjuvad. Tehisõppe algoritmid

töötavad aga veidi robustsemal tasandil. Need keskenduvad küll peaeesmärgile – mängu lõpptulemuse võimalikult täpsele hindamisele –, kuid neisse on keeruline põimida informatsiooni mängijate ja sündmuste tasandil. Mängu käigu modelleerimine võib teatud põhjustel olla aga kasulik ja isegi soovitav.

- Saame arvesse võtta spetsiifilist infot, mis peitub sündmus-sündmus andmetes.
- Kui eksisteerib töötav simuleerimisprotsess, siis on seda lihtsasti võimalik käivitada suvalisest mängu hetkest (vajab vaid sobivat mängu (alg)oleku initsialiseerimist) – see võimaldab kihlveokontoris panustada mängu kestel, selle jaoks eraldi mudelit treenimata.

Korvpallimängude simuleerimisel on võimalik tugineda juhuslike protsesside teooriale, nt käsitleda simulatsioone pideva või diskreetse ajaga Markovi ahelatena, kus toimub tõenäosustel põhinev liikumine eelnevalt defineeritud olekute vahel. Samas on korvpall väga dünaamiline mäng, kus adekvaatse täpsusega simulatsioonide ehitamine on keeruline. Juba varasemalt mainitud artiklis [KS06] aga ka artiklis [ŠV12] on siiski Markovi ahelatega katsetatud ka korvpallitulemuste ennustamisel. Viimases on kasutatud pallivaldamistel põhinevat mudelit, kus erinevad olekud moodustavad kõikvõimalikud situatsioonid, kus pallivaldamine kandub ühelt meeskonnalt teisele, nt vaheltlõike või 2p viske tabamise korral. Üleminekutõenäosused on arvutatud ajalooliste andmete põhjal, misjärel on leitud protsessi tasakaaluolekud. Viimaste pealt on leitud meeskondade ühe pallivaldamise jooksul keskmiselt visatud punktide arv (simulatsiooni jooksul sellel muutuda ei lubata, seega on protsess homogeenne), mängu lõppskoori saamiseks on seda korrutatud pallivaldamiste koguarvu hinnanguga. Mudelit on testitud kahel hooajal ning mängu võitjat on suudetud korrektselt klassifitseerida 69% ümber.

Magistritöös proovitakse korvpallimängude tulemusi simuleerida teistsuguse lähenemise abil. Korvpallis on korraga võimalik visata 3 erinevas väärtuses korve: vabavisked annavad 1p, kaugvisked 3p ja ülejäänud 2p. Eeldame, et igat tüüpi korvid tekivad üksteisest sõltumatute Poissoni protsesside kohaselt. Seega saame kolm erinevat protsessi $\{N_t^1 : t \geq 0\}$, $\{N_t^2 : t \geq 0\}$ ja $\{N_t^3 : t \geq 0\}$, mis ütlevad, kui palju on ajahetkeks t visatud vastavalt 1p korve, 2p korve ja 3p korve. Iga mängu korral arvutatakse ajaloolistelt andmetelt meeskondade poolt keskmiselt ühes sekundis visatud

erinevat tüüpi korvide arv ja neid keskmisi käsitletakse vastavate Poissoni protsesside intensiivsustena $\lambda_1, \lambda_2, \lambda_3$. Korvide arvu teisendamiseks meeskonna skooriks kombineerime saadud Poissoni protsesside käigus loetud korvid sobivalt kokku, nii et ajahetkel t avaldub meeskonna visatud punktide arv X_t kujul

$$X_t = N_t^1 + 2N_t^2 + 3N_t^3.$$

Meeskonna visatud korve simuleeritakse vastavatest Poissoni protsessidest kogu mänguaja jooksul (sekundites on see $48 \times 60 = 2880$) ja leitakse meeskonna lõppskoor X_{2880}. Sama tehakse mõlema meeskonna korral ja mängu võitjaks osutub see, kes viskas lõppskoori kohaselt rohkem punkte. Kui mäng jääb viiki, siis jätkatakse protsessidest simuleerimist $5 \times 60 = 300$ sekundi ehk 5-minutiliste lisaaegade kaupa kuni võitja selgumiseni. Kogu lähenemine eeldab, et korvidevaheline aeg on eksponentjaotusest, mis tegelikult ei kehti, kuid nagu hiljem näeme, siis selline lihtne mudel töötab siiski üsna hästi.

Peatükk 5

Analüüs

Käesolevas peatükis rakendame eelpool esitatud teooriat praktikas ja püüame NBA korvpallimängude tulemusi ennustades kihlveokontorit võita. Edasises uurime tihti, kui palju mängija mingit klassifitseerijat g(x) kasutades reaalses panustamissituatsioonis võitnud või kaotanud oleks. Meenutame, et testandmetena kasutatavate mängude kohta, mida on maksimaalselt 5331, on teada ka vähemalt ühe tabelis (3.1) toodud kihlveokontori sulgevad koefitsiendid. Eeldame, et mängija valib vastavate kihlveokontorite hulgast alati parima koefitsiendi ja panustab alati 1 ühiku mängule, millel klassifitseerija g(x) põhjal leidub väärtus. Nagu eelnevalt nägime, siis klassifitseerija poolt pakutav klass leitakse mingit reaalarvulist skoori teatud kohas poolitades. Väärtuse leidmisel oleme huvitatud tõenäosustest ja seda skoori võibki mõista tingliku tõenäosusena $p_1^g = p^g(1|x)$ (vajadusel saab skoori lõiku [0, 1] teisendada), et antud sisendi x korral kuulub objekt klassifitseerija g põhjal klassi 1. Ühtlasi saame sellest, et $p_0^g = p^g(0|x) = 1 - p_1^g$. Seejuures on kõikide klassifitseerijate skoorid hinnatud kodumeeskonna suhtes, nii et klassi 1 võib mõista kodumeeskonna võiduna, klassi 0 aga võõrsil meeskonna võiduna. Kokkuvõttes on ühe mängu pealt saadav tulu

$$T(g) = \begin{cases} k_1^i & p_1^g k_1^e > 1 \text{ ja } y_i = 1 \text{ ja } p_1^g k_1^e \geq p_0^g k_0^e \\ k_0^i & p_0^g k_0^e > 1 \text{ ja } y_i = 0 \text{ ja } p_0^g k_0^e > p_1^g k_1^e \\ 0 & p_1^g k_1^e \leq 1 \text{ ja } p_0^g k_0^e \leq 1 \\ -1 & \text{muidu} \end{cases}. \tag{5.1}$$

Tingimus $p_1^g k_1^e \geq p_0^g k_0^e$ tagab, et väärtuse leidmisel nii kodu- kui võõrsilmeeskonna mängult panustatakse kõrgema väärtusega mängule, võrdse

väärtuse korral aga kodumeeskonnale (viimane olukord on ebatõenäoline, kuid on lisatud matemaatilise korrektsuse tõttu). Sellist panustamismudelit kasutame, et keskenduda klassifitseerija headuse hindamisele, jättes vaatluse alt välja arbitraaži võimalused.

Kõik mängud, mille kohta andmed eksisteerivad, on ajaliselt järjestatavad. Seega saame skeemi (5.1) abil simuleerida mängudele panustamist nende toimumise järjekorras. See võimaldab esitada aegridasid, millelt saame infot trendi ja variatsiooni kohta, mis oleks mängijat tabanud, kui ta vastavat mudelit g reaalselt kasutanud oleks.

5.1 Baasmudelid

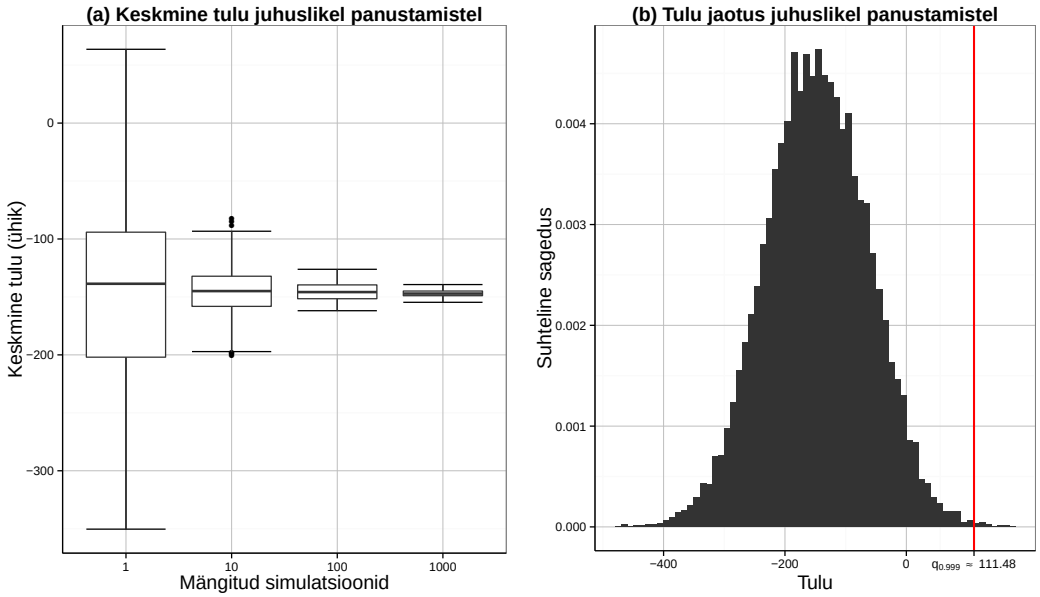
Loome nüüd esimesed lihtsad mudelid NBA korvpallimängude tulemuste ennustamiseks. Nende abil ei looda me mängude võitjaid kuigi edukalt klassifitseerida, pigem võimaldavad need valutut sissejuhatust modelleerimisprotsessi. Lihtsate mudelite abil saadud tulemused võtame baasinfoks edasisele analüüsile ja võrdlusobjektiks keerukamate meetoditega leitud tulemustele.

Vaatleme esmalt juhuslikku mudelit, mille korral ennustatakse korvpallimängu võitjat ausa mündiviske teel, st $g(x) = p_1^g = 1$, kui tuleb kull, ja $g(x) = p_1^g = 0$, kui tuleb kiri. Selline mudel eksib keskmiselt pooltel ennustustel, st juhusliku mudeli risk on 0,5. Huvitav on aga näha, kuidas mõjub juhuslikule mudelile tuginev panustamine mängija kontojäägile. Selleks teostame panustamist 5331 testmängule vastavalt skeemile (5.1). Teeme seda 1, 10, 100 ja 1000 korda ning leiame vastavad mängija keskmised lõpptulud. Iga simulatsioonide arvu korral leiame keskmise lõpptulu 100 korda ja võtame omakorda vastavad keskmised. Tulemuste varieeruvust ilmestab joonis (5.1) (a). Märkame, et lõpptulu koondub simulatsioonide arvu suurenedes kiiresti ja juba 1000 simulatsiooni pealt arvutatud keskmine varieerub võrdlemisi vähe. Seega võib mängija 5331 mängule juhuslikult panustades oodata keskmiselt u 145 ühiku suurust kahju, mis moodustab kogu panustatud summast 2,72%. Seda on veidi rohkem kui võiks järeldada lause (1.2) põhjal: arvestades, et keskmine liigprotsent üle kõikide mängude oli 1,79%, saaksime

$$v = \frac{0{,}0179}{1 + 0{,}0179} \approx 0{,}0176.$$

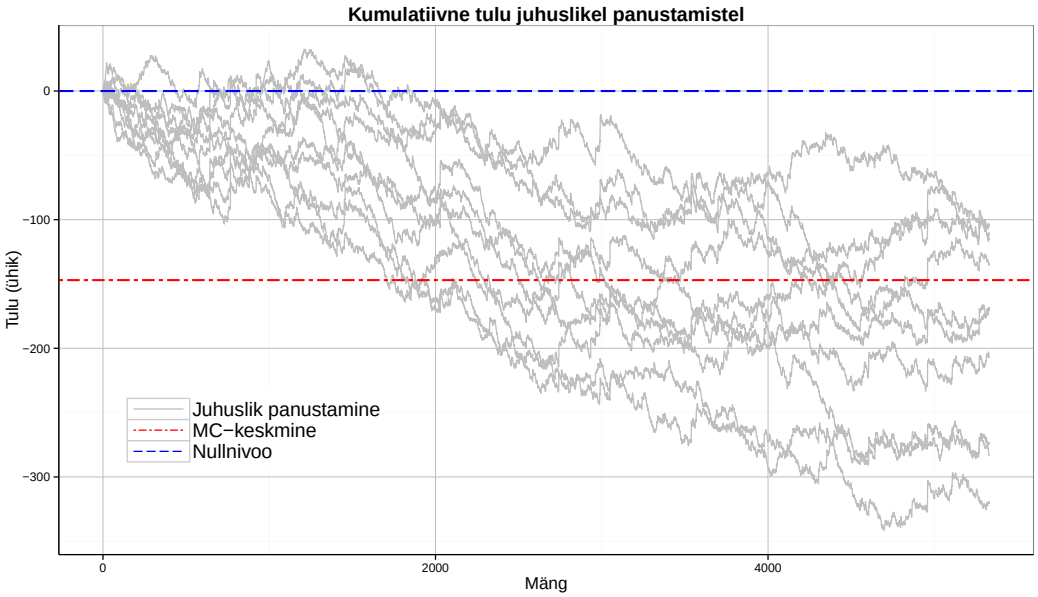
Samas põhineb lause (1.2) eeldustel, et vahendajad jaotavad liigprotsendi koefitsientide vahel proportsionaalselt vastavate tõenäosuste suurusele ja et vastavad tõenäosused on tegelikult ka õiged. Kumbki eeldus ei pruugi kehtida.

Joonisel (5.1) (b) on leitud lõpptulu jaotuse histogramm 10000 simulatsiooni pealt. Jooniselt näeme ka 99,9% kvantiili, mille väärtuse võime võtta aluseks mudelite statistilise olulisuse hindamisel. See on motiveeritud joonisest (5.1) (a), kus juba 1000 simulatsiooni pealt leitud lõpptulu keskmine andis vähe varieeruvaid tulemusi – sellest tulenevalt eeldame, et 10000 simulatsiooni pealt leitud 0,1% täiendkvantiil ei kõigu suurtes piirides.



Joonis 5.1: (a) Keskmise lõpptulu koondumine. Juba 1000 simulatsiooni pealt arvutatud lõpptulu keskmine varieerub väikestes piirides, mistõttu võime keskmise kahju suuruse üsna julgelt u 145 ühiku peale hinnata. (b) Lõpptulu jaotuse histogramm 10000 simulatsiooni pealt. Histogramm meenutab väga normaaljaotust. Vertikaaljoonega on tähistatud 99,9%-kvantiil $q_{0,999}$, mille väärtus on 111,48.

Joonisel (5.2) on esitatud mõned võimalikud kontojäägi aegread, kui mängija oleks panustanud juhusliku klassifitseerija abil. Märkame, et need on selge langustrendiga, st mängija võib juhusliku panustamise korral oodata stabiilset kahjumi kuhjumist. Lõppkulu poolest on aegread 145 ühiku ümbruses, kusjuures kõik kõverad on pärast 5331 mängu nullnivoost allpool ehk kahjumis.



Joonis 5.2: Kumulatiivne kontojääk juhuslikel panustamistel. Monte-Carlo keskmine on leitud 100 väärtuse hulgast, millest igaüks on omakorda 1000 simulatsiooni keskmine (vt joonist (5.1) (a)).

Püüame nüüd ennustustesse ka veidi täpsust lisada. Selleks kasutame järgmiseid lihtsaid lähenemisi:

- Klassifitseerime mängu võitja vastavat sellele, kes meeskondade viimastes omavahelistes kohtumistes edukam on olnud. Kui meeskondade A ja B viimasest 5 mängust on 3 korral võitnud esimene, siis klassifitseeritakse meeskond A mängu võitjaks, kusjuures tõenäosus, mida panustamisel kasutatakse, hinnatakse 0,6-ks.
- Kasutame eelmises punktis toodud lähenemist, ent vaatleme vaid selliseid mänge, kus meeskond A mängis kodus ja B võõrsil (vastupidised jätame arvestamata). Kui meeskond A võõrustab meeskonda B ja viimasest 5 meeskonna A kodustaadionil peetud mängust on 3 korral B kaotanud, siis klassifitseeritakse meeskond A mängu võitjaks, kusjuures tõenäosus, mida panustamisel kasutatakse, hinnatakse 0,6-ks.
- Klassifitseerime mängu võitjaks alati kodumeeskonna.

Esimesed 2 lähenemist ei vaja mudeli treenimist. Kolmandal juhul hinnatakse treeningandmetel tõenäosus, et kodumeeskond võidab. Seda kasutatakse testandmetel panustamissituatsioonis. Koefitsientide olemasolu

võimaldab andmed loomulikul viisil treening- ja testandmeteks jaotada: treeningandmed moodustavad sellised korvpallikohtumised, millele koefitsiente ei ole; testandmed moodustavad sellised mängud, millele on vähemalt ühe kihlveokontori sulgevad koefitsiendid teada. Tabelis (5.1) näeme ülaltoodud mudelitega saavutatud tulemusi.

Tabel 5.1: Erineval arvul viimaseid mänge arvesse võtvate lihtsate mudelite klassifitseerimisvead ja nendega panustamisel saavutatud lõpptulud pärast 5331 mängu.

Meetod	Mänge	Testviga	Lõpptulu
Kõik viimased mängud	1	0,414	177,02
	3	0,403	40,59
	5	0,410	−57,25
	7	0,411	−32,22
	10	0,405	−28,47
Viimased kodumängud	1	0,398	40,48
	3	0,399	−222,68
	5	0,409	−221,74
Koduvõidud		0,398	−278,71

Kõige väiksem testrisk ehk testmängudel tehtud klassifitseerimisviga on koduvõitude mudelil ja viimast kodumängu arvestaval mudelil. Samas saavutatakse suurima testriskiga mudelil (viimase omavahelise mängu põhjal klassifitseerimine) suurim lõpptulu, 177,02 ühikut, mis moodustab 3,3% kogu investeeritud 5331 ühikust. Niivõrd suur lõpptulu võib joonise (5.1) (b) kohaselt juhusliku panustamise korral esineda väga väikese tõenäosusega, mistõttu mudel võib suuta tuvastada väärtust. Samas näeme, et mudel, mis klassifitseerib paremini, ei pruugi seetõttu veel panustes väärtust leida.

5.2 Keerukamad mudelid

5.2.1 Tunnuste valik

Andmetest õppimisel on tunnuste valik väga oluline: me tahame leida võimalikult vähe tunnuseid (et mudel oleks lihtne), mis annavad võimalikult hea klassifitseerimistulemuse (väikse riski). Seega võib tunnuste valikut

vaadelda kui optimiseerimisprobleemi. Tihti ei ole võimalik seda ülesannet jõuga lahendada – proovida kõiki variante ja valida parim tunnuste komplekt –, sest kui tunnuste hulk on $\mathcal{X}$, siis kõikvõimalikke tunnuste komplekte on $2^{|\mathcal{X}|}$, mis kasvab tunnuste arvu kasvades eksponentsiaalselt. Seega on tunnuste valik keeruline probleem, mille lahendamiseks kasutatakse erinevaid heuristikuid.

Uute tunnuste loomisele ja tunnuste valikule NBA ennustamises on pühendatud suurem osa magistritööst [Pur13]. Viidatud töös on parima tunnuste komplekti valikuks kasutatud *ahnet* algoritmi[21] ja *geneetilist* algoritmi, kusjuures paremaid tulemusi on raporteeritud just esimese poolt. Ahne meetod on tunnuste valikul populaarne heuristik, mis tühjast tunnuste komplektist alustades lisab sinna iteratiivselt tunnuseid, mis vastaval sammul annab testandmetel parima ennustustäpsuse, tehes seda senikaua, kuni ennustustäpsus paraneb. Alustada võib ka suurimast võimalikust tunnuste hulgast ja eemaldada sealt tunnuseid senikaua, kuni klassifitseerija risk testandmetel väheneb. Ahne algoritm valib igal sammul vastaval hetkel parimat kirjeldusvõimet omava tunnuse, liikudes nii enamasti lokaalse optimumi poole. Meetod ei pruugi avastada tunnuseid, mis individuaalselt ei ole head, kuid kombinatsioonis teiste tunnustega on. Seetõttu on käesolevas magistritöös lisaks ahnele algoritmile implementeeritud ka *libalõõmutamise*[22] heuristiku arhitektuuri järgiv meetod, mis lubab lahendil lokaalsetest optimumidest välja hüpata.

Libalõõmutamise algoritm matkib terase lõõmutamise kontrollitud jahtumise protsessi, kus kuumutatud terase temperatuuri aeglasel vähendamisel liiguvad osakesed teatud tasakaaluolekusse. Tunnuste valiku kontekstis eeldab meetod mingi algoleku ehk lahendi olemasolu, millest alustades liigutakse naaberlahenditesse, püüdes otsida globaalset optimumi. Olgu alglahendiks kõikide tunnuste hulga $\mathcal{X}$ juhuslik mittetühi alamhulk $\mathcal{H} \subseteq \mathcal{X}, \mathcal{H} \neq \emptyset$. Meetod püüab hulka $\mathcal{H}$ lisada või sealt eemaldada tunnuseid, mis lahendit parandavad, leides nii pidevalt uusi tunnuste komplekte $\mathcal{H}_{uus}$. Seejuures kirjeldab lahendi ehk tunnuste hulga headust vastavate tunnuste peal treenitud klassifitseerija risk R testandmetel. Teatud tõenäosusega on võimalik, et hulka $\mathcal{H}$ lisatakse või sealt eemaldatakse tunnus, mis

[21] Ingl *greedy algorithm*
[22] Ingl *simulated annealing*

lahendit kehvemaks muudab. Sellist olukorda nimetame hüppeks, kusjuures hüpe toimub tõenäosusega, mis on võrdeline suurusega

$$\exp\left(\frac{R - R_{\text{uus}}}{\text{temperatuur}}\right).$$

Hüpped võimaldavad lahendil väljuda lokaalsetest optimumidest, et otsida globaalset optimumi n-ö teistest suundadest. Hüppe toimumise tõenäosus sõltub uue lahendi headusest võrreldes esialgse lahendiga, ent igal juhul see väheneb ajas, seega meetod koondub lõpuks mingis (lokaalses) optimumis. Tunnuste valikuks implementeeritud libalõõmutamise meetodi pseudokood on esitatud algoritmis (2). Meetodi kasutamiseks on tarvis

Algoritm 2 Libalõõmutamine tunnuste valikuks

1: $\mathcal{H} \leftarrow$ juhuslik mittetühi tunnuste komplekt $\mathcal{H} \subseteq \mathcal{X}, \mathcal{H} \neq \emptyset$
2: $\mathcal{H}_{\text{parim}} \leftarrow \mathcal{H}$
3: **Kuni** temperatuur > 0
4: X $\leftarrow$ lihtsa juhusliku valikuga $X \in \mathcal{X}$, nii et $\mathcal{H} \neq \{X\}$
5: **Kui** $X \in \mathcal{H}$, **siis**
6: $\mathcal{H}_{\text{uus}} \leftarrow \mathcal{H} \backslash \{X\}$
7: **muidu**
8: $\mathcal{H}_{\text{uus}} \leftarrow \mathcal{H} \cup \{X\}$
9: **Kui** $R_{\text{uus}} < R$ **või** hüppemoment **siis**
10: $\mathcal{H} \leftarrow \mathcal{H}_{\text{uus}}$
11: **Kui** $R < R_{\text{parim}}$ **siis**
12: $\mathcal{H}_{\text{parim}} \leftarrow \mathcal{H}$
13: $\mathcal{H} \leftarrow \mathcal{H}_{\text{uus}}$
14: temperatuur $\leftarrow$ temperatuur $-$ jahtumise samm
15: **Tagasta** $\mathcal{H}_{\text{parim}}$

valida 2 parameetrit: algtemperatuur ja jahtumise samm. Antud implementatsioonis on algtemperatuuriks valitud 0,05 ja jahtumise sammuks 0,0001. Lahendi ehk tunnuste hulga headust mõõdetakse vastavate tunnuste peal treenitud klassifitseerija poolt tehtud klassifitseerimisvea tegemise tõenäosusega testandmetel, seega toimuvad hüpped niisugusest algtemperatuurist alustades adekvaatse tõenäosusega. Meetod lõpetab töö, kui temperatuur jõuab nulli, tagastades parima leitud lahendi ehk tunnuste komplekti.

5.2.2 Tulemused

Kasutame nüüd eelmises peatükis kirjeldatud logistilist regressiooni ja AdaBoosti NBA korvpallimängude tulemuste ennustamiseks. Selleks kasutame R-i funktsioone *glm* ja *ada* (paketist *ada*). Igale mängule vastavad tunnused moodustame meeskondade teatud arvu eelnevate mängude põhjal. Eristame ka siin kahte lähenemist: valime meeskonna eelnevaid mänge nii kõikide kui ka vaid vastaval väljakul (kodus või võõrsil) selle meeskonna poolt eelnevalt mängitud mängude hulgast. Esimesel juhul vaatleme vaid selliseid mänge, kus mõlema meeskonna kohta on teada vähemalt 10 eelneva mängu tulemused, teisel juhul selliseid, kus mõlemad meeskonnad on eelnevalt mänginud vähemalt 5 mängu. Testandmed moodustavad sellised mängud, mille kohta on koefitsiendid teada, treeningandmed ülejäänud.

Igas mängus osaleb 2 meeskonda, seetõttu moodustub iga vaatlus mõlema meeskonna vastavatest tunnustest. Viimastena kasutatakse peaasjalikult peatükis (3.1.1) esitatud koondandmeid (normeeritud kujul); täiendava tunnusena on moodustatud meeskonna eelnevate mängude keskmine visatud punktide ja sisselastud punktide vahe (tähistame DIFF), mis allikas [Pur13] parimat individuaalset kirjeldusvõimet omas.

Mõlema meeskonna tunnuste ühendamisel saame kokku 30 tunnust. Parimate tunnuste leidmiseks kasutame eelmises alapunktis kirjeldatud algoritme: tühjast tunnuste hulgast alustavat ahnet meetodit, maksimaalse suurusega tunnuste hulgast alustavad ahnet meetodit ja libalõõmutamise printsiibil põhinevat meetodit. Kõik algoritmid kasutavad headuse kriteeriumina logistilise regressiooniga või AdaBoostiga testandmetel tehtud klassifitseerimisviga. Kõikidel juhtudel saavutati parim tunnuste komplekt edaspidise ahne meetodiga, kuigi libalõõmutamise algoritm andis väga lähedasi tulemusi.

Vähemalt 10 ja maksimaalselt 30 viimase mängu põhjal leitud tunnuste korral valis tühjast hulgast alustav logistilise regressiooniga testviga hindav ahne algoritm välja 5 olulist tunnust – A_DIFF, B_DIFF, B_3FGM, B_OREB, B_TOV (A - kodumeeskond, B - võõrsil meeskond) –, mille korral saavutati logistilise regressiooniga testandmetel ennustamistäpsuseks 68,9% (libalõõmutamise algoritmi abil leiti parimal juhul 11 tunnust, mis saavutasid testandmetel klassifitseerimistäpsuseks 68,5%; maksimaalsest tunnuste

hulgast alustav ahne algoritm ei andnud võrreldavaid tulemusi). AdaBoosti meetodi korral leiti sobivad tunnused ahnest algoritmist nii, et headuse kriteeriumina kasutati AdaBoosti klassifitseerijaga tehtud klassifitseerimisviga testandmetel. Nõrkade klassifitseerijatena rakendati otsustuspuid maksimaalse kõrgusega 2 ning lubati maksimaalselt 50 iteratsiooni: ennustustäpsuseks testandmetel saavutati nii 67,6%, mis jääb logistilise regressiooni tulemusele selgelt alla. Kirjeldatud protseduur läbiti nii logistilise regressiooni kui AdaBoosti korral ka teiste tunnuste konstrueerimise variantidega. Resultaadid on esitatud tabelites (5.2) ja (5.3), kus lõpptulu on leitud jällegi skeemi (5.1) alusel testandmetel (vastavaid mänge on sedapuhku 5319) panustades.

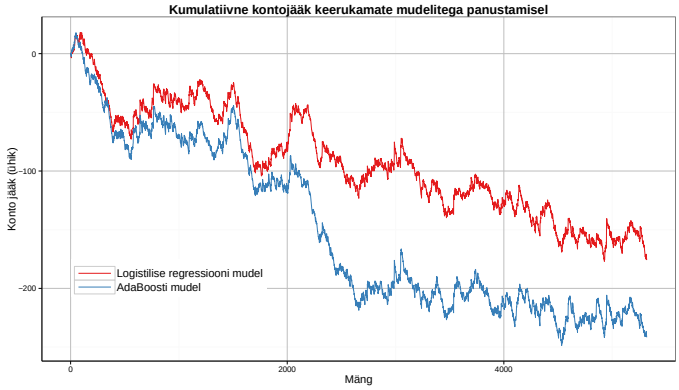
Tabel 5.2: Logistilise regressiooniga klassifitseerimise tulemused.

Variant	Mänge	Testviga	Lõpptulu
Kõik viimased mängud	20	0,314	−211,44
	30	0,311	−175,86
Viimased kodumängud	10	0,328	−160,56
	15	0,320	−110,64

Tabel 5.3: AdaBoostiga klassifitseerimise tulemused.

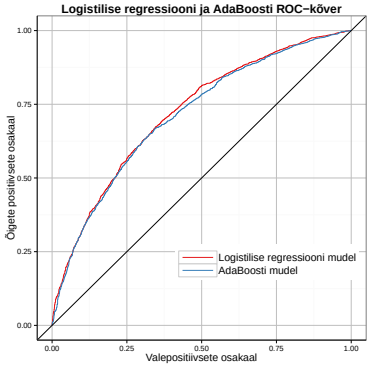
Variant	Mänge	Testviga	Lõpptulu
Kõik viimased mängud	20	0,331	−218,76
	30	0,324	−241,67
Viimased kodumängud	10	0,333	−183,01
	15	0,331	−227,92

Meetodid ennustavad küll üsna täpselt, kuid ei ole kuigi tulusad. Väikseim risk testandmetel on 31,1%, mis saavutati konstrueerides tunnuseid vähemalt 10 ja maksimaalselt 30 viimase mängu põhjal (tunnuste valiku protsess on kirjeldatud ülal) ja kasutades logistilist regressiooni. Samas oleks mängija selle mudeli alusel panustades kaotanud 175,86 ühikut ehk 3,3% kogu panustatud rahasummast. Langustrendi visualiseerib ilmekalt joonis (5.3), kus nimetatud tunnuste konstrueerimise variandi korral on esitatud logistilise regressiooni ja AdaBoosti klassifitseerijaga reprodutseeritud rahavood. Logistilise regressiooni abil leitud tulemused on seejuures veidi paremad kui AdaBoostiga saadud tulemused.



Joonis 5.3: Mängija kumulatiivne kontojääk parima logistilise regressiooni ja AdaBoosti klassifitseerijaga panustamise korral. Mõlema meetodi korral oleks mängija testandmete põhjal kahjumisse jäänud.

Klassifitseerijate headust saab võrrelda ka ROC[23]-kõverate abil. Parima logistilise regressiooni ja AdaBoosti mudeli jaoks on vastavad ROC-kõverad esitatud joonisel (5.4).



Joonis 5.4: Parima logistilise regressiooni ja AdaBoosti klassifitseerija ROC-kõverad. Mõlemad klassifitseerijad on väga sarnased, ent logistilise regressiooni ROC-kõvera alune pindala on veidi suurem.

ROC-kõverate mõistmiseks meenutame veelkord, et klassifitseerija binaarne tulemus põhineb mingi reaalarvulise skoori lõikamisel. Hea klassifitseerija on selline, mis annab kõrge skoori ühte klassi – nt klassi 1 – kuuluvatele objektidele ja madala skoori klassi 0 kuuluvatele objektidele, nii et sobivast kohast skoori lõigates saaks klassid võimalikult hästi eraldada. ROC-kõverad näitavad iga lõikamiskoha suhtes, kui suur osa õigesti klassi

[23]Ingl *receiver operating characteristic*

1 klassifitseeritud objekti (n-ö õiget positiivset) esineb valesti klassi 1 klassifitseeritud objekti (n-ö valepositiivse) kohta. Iga ROC-kõvera murdepunkt tähistab ühte sellist lõikepunkti, seega saab klassifitseerijate ROC-kõverate aluse pindala suuruse põhjal klassifitseerijaid võrrelda. Jooniselt (5.4) näeme, et parima logistilise regressiooni mudeli ROC-kõvera alune pindala on veidi suurem kui parima AdaBoosti klassifitseerija ROC-kõvera alune pindala, seega on esimene meetod klassifitseerimise mõttes veidi parem.

Mudelite ennustustäpsuse headust aitab hoomata see, et kihlveokontorite koefitsientide alusel klassifitseerimine andnuks täpsuseks testandmetel 69,3% (ka kihlveokontorite poolt favoriidiks peetud meeskondadele panustamine osutus kahjumlikuks), mis on vähem kui poole protsendi võrra parem kui parimal logistilise regressiooni mudelil. Hooaegade kaupa täpsuse võrdlus on esitatud tabelis (5.4).

Tabel 5.4: Parima logistilise regressiooni mudeli ja kihlveokontorite koefitsientide alusel klassifitseerimise täpsuse võrdlus testandmetel.

Aasta	Mänge	Logistiline regressioon	Kihlveokontorite koefitsiendid
2008	676	0,700	0,706
2009	1228	0,696	0,698
2010	1206	0,692	0,692
2011	990	0,683	0,682
2012	1219	0,677	0,692
Kokku	5319	0,686	0,693

Logistilise regressiooni mudeli risk testandmetel on aastatel 2008-2011 väga sarnane kihlveokontorite koefitsientide alusel klassifitseerimise – võitjaks on valitud väiksema koefitsiendiga meeskond – riskiga, vaid 2012. aastal on risk suurem. Sellest hoolimata on mudel kahjumlik, kuigi panustati väärtuse printsiibist lähtuvalt. Märksa kehvem mudel peatükis (5.1) saavutas oluliselt suurema tulususe.

5.3 Mudeli valik tulususe põhjal

Siiamaani oleme mudelit valides püüdnud minimiseerida klassifitseerimisvigade arvu. Selline lähenemine on õigustatud, sest kui klassifitseerija on hea, siis on mängijal lootust ka tulu saada. Kui klassifitseerija on 100%

täpne – ka skoori poolest, st kui panustes peaks leiduma väärtus, siis mängija suudaks selle antud klassifitseerija abil leida –, siis ei saaks mängija panuseid kaotada. Siiski oleme eelnevalt näinud, et klassifitseerimisvigade arv ei ole liiga tugevas korrelatsioonis mängija kasumiga, seetõttu võiks klassifitseerija leidmisel otse kasumit maksimiseerida.

Olgu kodu- ja võõrsilmeeskonna võidud kodeeritud vastavalt +1 ja -1. Kui valimis olevate mängude kohta on teada meeskondade inglise tüüpi koefitsiendid – vastavalt k_+ ja k_- –, st kasutada on treeningvalim

$$\mathcal{D}_n = \{(x_1, k_+(1), k_-(1), y_1), ..., (x_n, k_+(n), k_-(n), y_n)\},$$

siis soovime klassifitseerija leidmisel maksimiseerida treeningandmetel kasumit, st leida klassifitseerija $g_n \in \mathcal{G}$ nii, et

$$g_n = \arg\max_{g \in \mathcal{G}} \sum_{i=1}^{n} U_i(y_i, g(x_i)),$$

kus

$$U_i(y_i, g(x_i)) = \begin{cases} k_+(i) & y_i = g(x_i) = +1 \\ k_-(i) & y_i = g(x_i) = -1 \\ -1 & y_i \neq g(x_i) \end{cases}.$$

Uurime kõigepealt, kas mõni kasutada olevast 30 tunnusest suudab individuaalselt treenida kasumipõhist mudelit selliselt, et me testandmetel ka kasumit näeme. Klassifitseerijate hulgana $\mathcal{G}$ kasutame siin otsustuspuid, sest nende korral suudame antud ülesannet jõuga lahendada. Täpsemalt, olgu X mingi tunnus ja olgu mingis puu tipus t treeningandmete arv n_t. Olgu tunnuse X erinevate väärtuste hulk selles tipus V_X ja $|V_X|$ selle hulga võimsus ehk unikaalsete väärtuste arv. Siis leitakse selles tipus tunnuse X lõikepunkt z_X^t hulgast

$$Z = \begin{cases} V_X & |V_X| <= 100 \\ \left\{ \lambda \max(V_X) + (1-\lambda) \min(V_X) : \lambda = 0, \frac{1}{99}, ..., 1 \right\} & |V_X| > 100 \end{cases}$$

nii, et

$$z_X^t = \arg\max_{z \in Z} \left(\max_{c \in \{1,-1\}} \left(\sum_{i: x_i \leq z}^{n_t} U_i(y_i, c) + \sum_{i: x_i > z}^{n_t} U_i(y_i, -c) \right) \right).$$

Seega valime maksimaalselt 100 lõikepunkti hulgast välja sellise, mis lõikab treeningandmed kasumi mõttes kõige paremini. Mängule panustamine otsustatakse puu lehtedes olevate treeningandmete kasumlikkuse põhjal. Kui x on testmäng, mis mööda puud reegleid pidi alla liikudes jõuab lehte l_x, kus on n_{l_x} treeningmängu, siis vastav klassifitseerija on

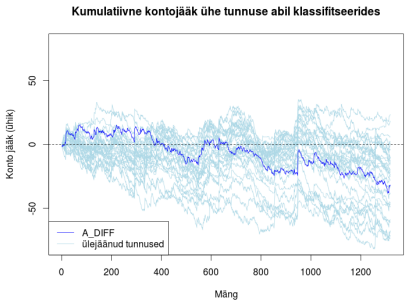
$$g(x) = \arg\max_{c \in \{1,-1\}} \sum_{i=1}^{n_{l_x}} U_i(y_i, c). \tag{5.2}$$

Mängule panustame vaid siis, kui $\sum_{i=1}^{n_{l_x}} U_i(y_i, c) > 0$, seega on ühe mängu pealt saadav tulu

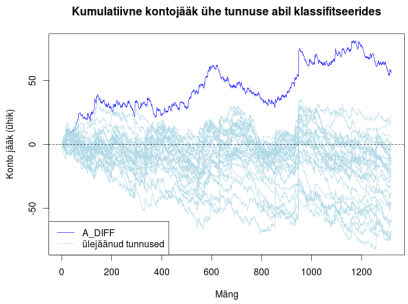
$$T(g(x)) = \begin{cases} k_+ & g(x) = +1 \text{ ja } \sum_{i=1}^{n_{l_x}} U_i(y_i, 1) > 0 \\ k_- & g(x) = -1 \text{ ja } \sum_{i=1}^{n_{l_x}} U_i(y_i, -1) > 0 \\ 0 & \sum_{i=1}^{n_{l_x}} U_i(y_i, g(x)) \leq 0 \\ -1 & \text{muidu} \end{cases}. \tag{5.3}$$

Klassifitseerijat (5.2) ja selle abil panustamise skeemi (5.3) kasutades simuleeriti jällegi reaalseid rahavooge. Kuna kasumit maksimiseeriv klassifitseerimine nõuab koefitsientide olemasolu, siis treeningandmed moodustasid 4000 varasemat mängu, millele koefitsiendid ning vähemalt 10 ja maksimaalselt 30 viimase mängu põhjal leitud tunnused olemas olid, testandmed aga hilisemad 1319. Joonisel (5.5) vastab igale joonele üheainse tunnuse põhjal puu ehitamine – kokku on seega 30 erinevat aegrida. Näeme, et treeningandmete põhjal lahendame me õiget probleemi, st lõpptulu on pea igal juhul positiivne. Paraku ei saa sama öelda testandmete kohta. Seega individuaalselt ei suuda mitte ükski meil kasutada olevatest tunnustest treenida mudelit nii, et selle põhjal testandmetel selge positiivse trendiga rahavoogu näeks.

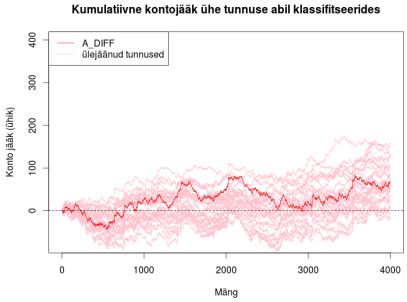
Kuigi ükski tunnus individuaalselt meile soovitud tulemust ei anna, siis võiks proovida kõiki tunnuseid korraga kasutada ehk puud ehitades lõigata andmeid selliselt, mis maksimiseeriksid kasumi üle kõikide tunnuste ja nende lõikepunktide. Eelpool toodud tähistusi kasutades soovime leida



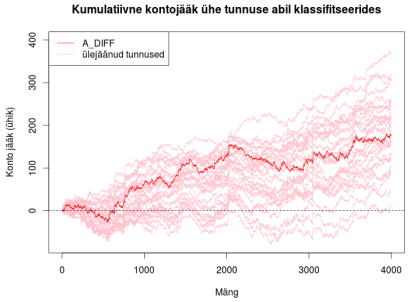
(a) Test, puu kõrgus 2



(b) Test, puu kõrgus 5



(c) Treening, puu kõrgus 2



(d) Treening, puu kõrgus 5

Joonis 5.5: Rahavood test- ja treeningandmetel, kui mudel on treenitud otsustuspuul, mis leiab lõikamised selliselt, mis maksimiseerivad kasumit treeningandmetel. Kõik mudelid on leitud ainult ühte tunnust kasutades, proovides niimoodi kõik 30 erinevat tunnust läbi.

tunnust X ja sellele vastavat lõikepunkti z_X selliselt, mis maksimiseeriksid suurust

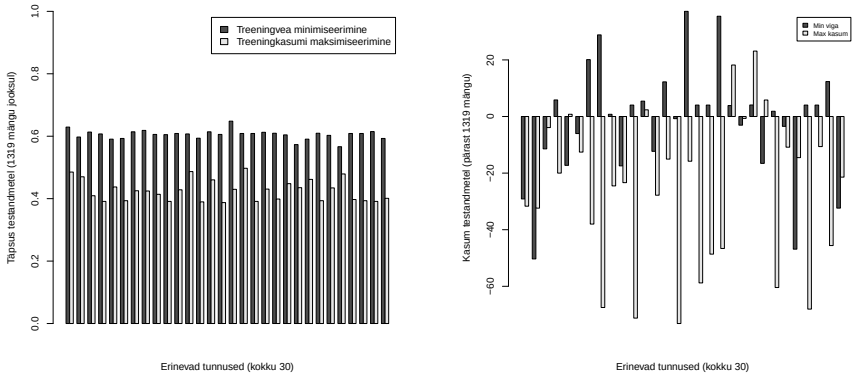
$$\max_{c\in\{1,-1\}}\left(\sum_{i:x_i\leq z}^{n_t} U_i(y_i,c)+\sum_{i:x_i>z}^{n_t} U_i(y_i,-c)\right).$$

Niimoodi ehitatud puudega on rahavooge kontrollitud tabelis (5.5).

Tabel 5.5: Rahavood otsustuspuu kasutamisel, kui kasumit on maksimiseeritud, st parimad lõikamised on valitud üle kõigi 30 tunnuse korraga.

Andmete tüüp	Max puu kõrgus	Testviga	Lõpptulu
Treeningandmed	2	0,482	173,57
	5	0,420	687,41
Testandmed	2	0,565	−38,27
	5	0,532	−35,18

Tabelist näeme, et juba puu kõrguse 2 korral, mis juhul veel ei tohiks karta ülesobitamist, on treeningandmetel panustamine väga kasumlik. Testandmetel panustades saanuks mängija samas kahjumit. Nagu võime näha jooniselt (5.6), siis ühe tunnusega klassifitseerija treenimisel ei anna kasumi maksimiseerimine testandmetel paremaid tulemusi isegi võrreldes treeningvea minimiseerimisega. Sama kehtib ka siis, kui kasutada puu treenimisel kõiki tunnuseid: treeningriski minimiseerimise teel leitud otsustuspuu peal teostatud panustamine andis suurema lõpptulu (kuid siiski negatiivse) kui kasumi maksimiseerimise teel leitud otsustuspuu. Seega need mängude omadused, mis treeningandmete põhjal on viidanud panustamise kasumlikkusele, ei ole tähendanud sama testandmete peal.



(a) Klassifitseerimise täpsused (b) Mudelite lõpptulu

Joonis 5.6: Ühe tunnuse peal treenitud otsustuspuude (max kõrgusega 2) klassifitseerimise täpsused testandmetel ja mängija lõpptulud vastavate puude abil teostatud panustamise korral. Kasumit maksimiseerivad mudelid ei ole lõpptulu poolest paremad. Treeningviga minimiseerivate klassifitseerijate tulemused on väga sarnased, sest enamikel tunnustel ei ole head kirjeldusvõimet ja seega on klassifitseerimine lähedane enamushääletuse tulemusele (kodumeeskonna võite on andmestikus ca 60%).

Mudeli treenimine kasumile orienteeritult püüab sisuliselt avastada kihlveokontori süstemaatilist viga: kui vahendaja pakub mingit tüüpi sündmusele pidevalt liiga suurt koefitsienti, siis peaks mudel selle avastama. Jooniste (5.5), (5.6) ja tabeli (5.5) põhjal saame seega järeldada, et kihlveokontorid ei tee koefitsientide määramisel süstemaatilist viga – vähemalt mitte sellist, mida meie 30 tunnust üles leida võiksid.

5.3.1 Modifitseeritud AdaBoost

Üritame kasumit silmas pidades rakendada ka AdaBoosti algoritmi. Leiame nõrku klassifitseerijaid samamoodi nagu enne – kasutades maksimaalselt kõrgusega 2 otsustuspuid, mis leitakse treeningvigade arvu minimiseerivalt –, ent proovime vaatluste sobiva kaalumise teel kasumit suurendada. Selleks defineerime funktsiooni

$$u_i(y_i) = \begin{cases} k_+(i) & y_i = +1 \\ k_-(i) & y_i = -1 \end{cases}.$$

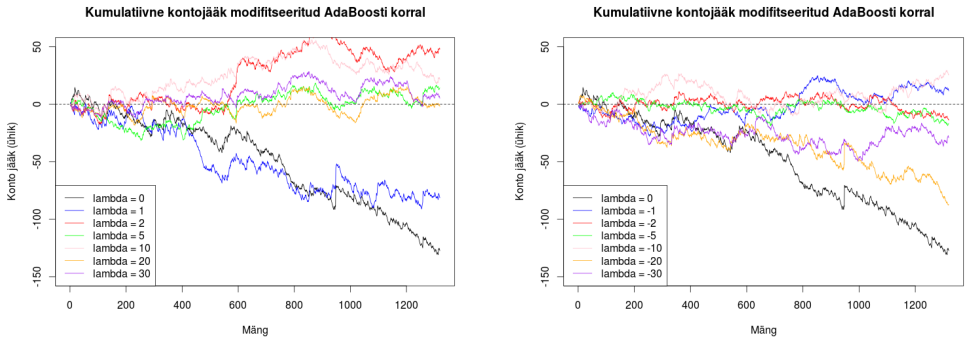
Seega on $u_i(y_i)$ mängija kasum mängu i tulemuse äraarvamisel. Meenutame valemit (4.6), mille kohaselt AdaBoost minimiseerib teataval moel eksponentisaalset riski. Asendame summa (4.6) kaalutud summaga

$$\sum_{i=1}^{n} \exp(\lambda u_i(y_i) - y_i f(x_i)) = \sum_{i=1}^{n} \exp(\lambda u_i(y_i)) \exp(-y_i f(x_i)),$$

kus λ on regulariseerimiskonstant. Kui $\lambda > 0$, siis on suuremate koefitsientidega mängudel rohkem kaalu, kui $\lambda < 0$, siis on väiksemate koefitsientidega mängudel rohkem kaalu, kui $\lambda = 0$, saame valemi (4.6). Modifitseeritud AdaBoosti rakendamiseks valime vaatluste esialgseteks kaaludeks

$$\beta_i^1 = \exp(\lambda u_i(y_i)).$$

Modifitseeritud AdaBoosti on rakendatud mitmete erinevate λ-de korral – nii positiivsete kui negatiivsete –, et uurida vaatluste ümberkaalumisse mõju mängija kasumile. Tulemused on esitatud joonisel (5.7). Joonise osast (a) tundub, justkui aitaks suurema koefitsiendiga mängudele suurema kaalu omistamine (mudelid, mille korral $\lambda > 0$) mängija kasumit tõsta: 6 mudelist 5 korral oleks mängija 1319 mängu jooksul jäänud kasumisse. Samas on keeruline seletada, miks $\lambda = 1$ korral on rahavood langustrendiga, ent $\lambda = 2$ korral tõusutrendiga. Mudelid, mille korral $\lambda < 0$ (joonise osa (b)), ei ole nii kasumlikud olnud. Samas annavad kõik modifitseeritud AdaBoostiga treenitud mudelid ($\lambda \neq 0$) testandmetel suurema kasumi kui originaalne AdaBoost ($\lambda = 0$). Seega näib vaatluste kaalumine koefitsientidest lähtuvalt mängija kasumit tõstvat.



(a) $\lambda \geq 0$ (b) $\lambda \leq 0$

Joonis 5.7: Rahavood testandmetel, kasutades modifitseeritud AdaBoosti.

5.4 Simulatsioonid

Püüame nüüd ennustamiseks kasutada sündmus-sündmus andmeid, mis annavad iga mängu kohta detailsemat informatsiooni. Loogiline on eeldada, et enne iga korvpallikohtumist on teada, kes mängijatest on mängukõlbulikud ja kes mitte (nt vigastuse tõttu). Kasutame seda informatsiooni, et konstrueerida korvide viskamise intensiivsusel tuginev korvpallimängu simulatsioon, mille mitmekordsel rakendamisel saadud keskmiste tulemuste alusel moodustame klassifitseerimiseeskirja.

Vaatleme korvpallimängu ja selles osalevat meeskonda S. Olgu $s_1, ..., s_m$ meeskonna S mängukõlbulikud mängijad selleks kohtumiseks ja olgu mängule eelnevaid meeskonna S mänge n. Tähistame $n^* = \min\{n, \gamma\}$, kus γ on maksimaalne ajalooliste mängude arv, mida vaatleme (eelnevas analüüsis saavutati häid tulemusi juhul $\gamma = 30$). Olgu mängus j mängija s_i poolt mängitud sekundite arv $t_j^{s_i}$, mängija s_i poolt väljakul oldud ajal meeskonna S poolt visatud korvide arv $\phi_j^{s_i}$ ja samal ajal vastasmeeskonna poolt visatud korvide arv $\psi_j^{s_i}$, $i = 1, ..., m$, $j = 1, ..., n^*$. Siis mängule eelnevate meeskonna S poolt osaletud mängudes, kui väljakul on olnud mängijad $s_1, ..., s_m$, avaldame S korvide viskamise intensiivsuse μ_S ja S-ile visatud korvide intensiivsuse ν_S kujul

$$\mu_S = \sum_{i=1}^{m} w_i \mu_{s_i}, \quad \nu_S = \sum_{i=1}^{m} w_i \nu_{s_i},$$

kus

$$w_i = \frac{\sum_{j=1}^{n^*} t_j^{s_i}}{\sum_{i=1}^{m}\sum_{j=1}^{n^*} t_j^{s_i}}, \quad \mu_{s_i} = \frac{\sum_{j=1}^{n^*} \phi_j^{s_i}}{\sum_{j=1}^{n^*} t_j^{s_i}}, \quad \mu_{s_i} = \frac{\sum_{j=1}^{n^*} \psi_j^{s_i}}{\sum_{j=1}^{n^*} t_j^{s_i}}.$$

Lihtsustades saame vastavateks intensiivsusteks

$$\mu_S = \frac{\sum_{j=1}^{n^*}\sum_{i=1}^{m} \phi_j^{s_i}}{\sum_{j=1}^{n^*}\sum_{i=1}^{m} t_j^{s_i}}, \quad \nu_S = \frac{\sum_{j=1}^{n^*}\sum_{i=1}^{m} \psi_j^{s_i}}{\sum_{j=1}^{n^*}\sum_{i=1}^{m} t_j^{s_i}}.$$

Olgu mängu kodumeeskond A ja võõrsil mängiv meeskond B, siis A korvide viskamise intensiivsuse B vastu λ_{AB} ja B korvide viskamise intensiivsuse A vastu λ_{BA} avaldame kujul

$$\lambda_{AB} = \kappa\mu_A + (1-\kappa)\nu_B,$$

$$\lambda_{BA} = \kappa\mu_B + (1-\kappa)\nu_A,$$

kus $\kappa \in [0,1]$. Ühtlasi saame eristada erinevaid mänguperioode ning 1-punkti, 2-punkti ja 3-punkti korve, nii saame mängus osalevate meeskondade korvi viskamise intensiivsused vastavalt

$$\lambda_{AB}^{k,l} = \kappa\mu_A^{k,l} + (1-\kappa)\nu_B^{k,l},$$

$$\lambda_{BA}^{k,l} = \kappa\mu_B^{k,l} + (1-\kappa)\nu_A^{k,l},$$

kus $k = 1,2,3,4$ tähistab perioodi (4. perioodi hulka loetakse ka lisaajad) ja $l = 1,2,3$ tähistab korvi väärtust.

Korvi tüüpide (erinevad väärtused on endiselt 1, 2, 3) kaupa defineeritud intensiivsusi saame käsitleda peatükis (4.3) kirjeldatud Poissoni protsesside $\{N_t^{A,1} : t \geq 0\}$, $\{N_t^{A,2} : t \geq 0\}$, $\{N_t^{A,3} : t \geq 0\}$ ja $\{N_t^{B,1} : t \geq 0\}$, $\{N_t^{B,2} : t \geq 0\}$, $\{N_t^{B,3} : t \geq 0\}$ (A viitab kodumeeskonna protsessidele, B aga võõrsil mängiva meeskonna protsessidele) parameetritena. Homogeensete protsesside korral arvutame iga mäng mõlemale meeskonnale nende viimaste mängude põhjal (maksimaalselt 30, minimaalselt 10) kogu mängu kehtivad intensiivsused $\lambda_{AB}^{\cdot,l}$ ja $\lambda_{BA}^{\cdot,l}$, kus "$\cdot$" viitab sellele, et intensiivsused veerandaegade põhjal ei erine; mittehomogeensete protsesside korral arvutame intensiivsused $\lambda_{AB}^{k,l}$ ja $\lambda_{BA}^{k,l}$ vaid vastavate veerandaegade k põhjal ja muudame protsesside parameetrit vastaval ajal ka simuleerimisel.

Simulatsioone teostame 5319 mängul, mille kohta on teada vähemalt

ühe tabelis (3.1) toodud kihlveokontori sulgevad koefitsiendid ja mille mõlemad meeskonnad on meie andmete põhjal eelnevalt mänginud vähemalt 10 mängu. Simulatsioonid on üles ehitatud peatükis (4.3) kirjeldatud loogika alusel. Iga mäng simuleeritakse eelpool kirjeldatud viisil leitud intensiivsuste abil korvide viskamist; visatud korvide arv korrutatakse vastavate väärtustega (korvi väärtus on kas 1, 2 või 3 punkti) misjärel väärtused summeeritakse mängu skooriks. Seega leitakse kahe meeskonna A ja B lõppskoorid pärast 48 minutit ehk $T = 2880$ sekundit X_T^A ja X_T^B järgmiselt:

$$X_T^A = N_T^{A,1} + 2N_T^{A,2} + 3N_T^{A,3},$$

$$X_T^B = N_T^{B,1} + 2N_T^{B,2} + 3N_T^{B,3}.$$

Kui pärast 48 minutit on skoor viigis, siis jätkatakse erineva väärtusega korvide simuleerimist vastavate intensiivsustega (homogeensete protsesside korral intensiivsustega $\lambda_{AB}^{\cdot,l}$ ja $\lambda_{BA}^{\cdot,l}$; mittehomogeensete protsesside korral intensiivsustega $\lambda_{AB}^{4,l}$ ja $\lambda_{BA}^{4,l}$) 5-minutiliste ehk 300-sekundiliste lisaaegade jooksul kuni võitja selgumiseni. Niimoodi teostatakse iga mängu kohta 1001 simulatsiooni, kusjuures mängu võitja klassifitseeritakse kahel erineval viisil: keskmistatud skoori põhjal ja 1001 simulatsioonist suurema arvu võitude põhjal. Võitude osakaalud määravad ka panustamiseks vajalikud tõenäoused. Parimaid tulemusi saavutasime $\kappa = 0{,}5$ korral, vastavad tulemused on esitatud tabelis (5.6).

Tabel 5.6: 1001 Poissoni liitprotsessi simulatsiooni abil teostatud klassifitseerimise tulemused üle 5319 mängu, kui $\kappa = 0{,}5$.

Protsessi tüüp	Täpsus: keskmine skoor	Täpsus: võitude arv	Lõpptulu
Homogeenne	0,654	0,651	−206,26
Mittehomogeenne	0,644	0,647	−185,1

Mittehomogeenne mudel andis veidi kehvemaid tulemusi, kuid see võib olla tingitud sellest, et 30 mängust jääb 4 erineva perioodi intensiivsuste hindamiseks väheks. Üldiselt näeme aga juba tuttavat efekti: mudelite üsna heast klassifitseerimistäpsusest hoolimata on nendega panustamine kahjumlik. Suures ulatuses näib meetodite suur kahjumlikkus tekkivat sellest, et vastavalt skeemile (5.1) kasutame panustamisel mudelite poolt leitud tinglikke tõenäosusi ehk teatud juhtudel võime panustada meeskonnale,

keda võitjaks ei klassifitseeritud (kui väärtus leiti hoopis kaotajaks klassifitseeritud meeskonna koefitsiendilt). Kui panustaksime alati klassifitseerija poolt võitjaks kuulutatud meeskonnale, siis oleksid tulemused märksa paremad: homogeense mudeli korral oleks lõpptulu sõltuvalt keskmise skoori ja võitude arvu põhjal klassifitseerimisest vastavalt 13,97 ja -16,65, mittehomogeense mudeli korral aga -59,26 ja -75,72. Seega on väärtusel põhinev panustamine väljakutsuv probleem, sest klassikuuluvust kirjeldavate tinglike tõenäosuste korrektne hindamine on keeruline ülesanne.

Viited

[BHP97] Bruce Bukiet, Elliotte Rusty Harold, and José Luis Palacios. A markov chain approach to baseball. *Operations Research*, 45(1):14–23, 1997.

[Bil95] Patrick Billingsley. *Probability and Measure*. John Wiley & Sons, third edition, 1995.

[Buc03] Joseph Buchdahl. *Fixed Odds Sports Betting*. High Stakes, 2003.

[CJM06] Mark Culp, Kjell Johnson, and George Michailidis. ada: An R Package for Stochastic Boosting. *Journal of Statistical Software*, 17(2):9, 2006.

[F+10] Elihu D. Feustel et al. *Conquering Risk: Attacking Vegas and Wall Street*. Elihu Feustel, 2010.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

[KS06] Paul Kvam and JoelŠ Sokol. A logistic regression/markov chain model for ncaa basketball. *Naval research Logistics (NrL)*, 53(8):788–803, 2006.

[Kä11] Meelis Käärik. Juhuslikud protsessid. Loengukonspekt, 2011.

[Law95] G.F. Lawler. *Introduction to Stochastic Processes*. Chapman & Hall/CRC Probability Series. Taylor & Francis, 1995.

[Lem12] Jüri Lember. Tõenäosusteooria II. Loengukonspekt, 2012.

[Lem13] Jüri Lember. Tehisõpe I. Loengukonspekt, 2013.

[LM12] Amy N Langville and Carl Dean Meyer. *Who's# 1?: the science of rating and ranking*. Princeton University Press, 2012.

[Ltd14] LiveSport Media Ltd. Oddsportal. `http://www.oddsportal.com`, märts 2014.

[MGKK10] Dejan Miljkovic, L Gajic, Aleksandar Kovacevic, and Zora Konjovic. The use of data mining for basketball matches outcomes prediction. In *Intelligent Systems and Informatics (SISY), 2010 8th International Symposium on*, pages 309–312. IEEE, 2010.

[NMV14] LLC NBA Media Ventures. NBA. `http://www.nba.com`, märts 2014.

[Pur13] Keshav Puranmalka. Modelling the NBA to make better predictions. Master's thesis, Massachusetts Institute of Technology, 2013.

[Spo14] SportsbookReview. Sportsbookreview. `http://www.sportsbookreview.com/sportsbooks/`, märts 2014.

[ŠV12] Erik Štrumbelj and Petar Vračar. Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting*, 28(2):532–542, 2012.

[Sõ13] Jaak Sõnajalg. Tutvumine adaboostiga. Bakalaureusetöö, Tartu Ülikool, 2013.

[Tea14] Riigi Teataja. Autoriõiguse seadus. `https://www.riigiteataja.ee/akt/810714`, märts 2014.

Lisad

A Programmi kood

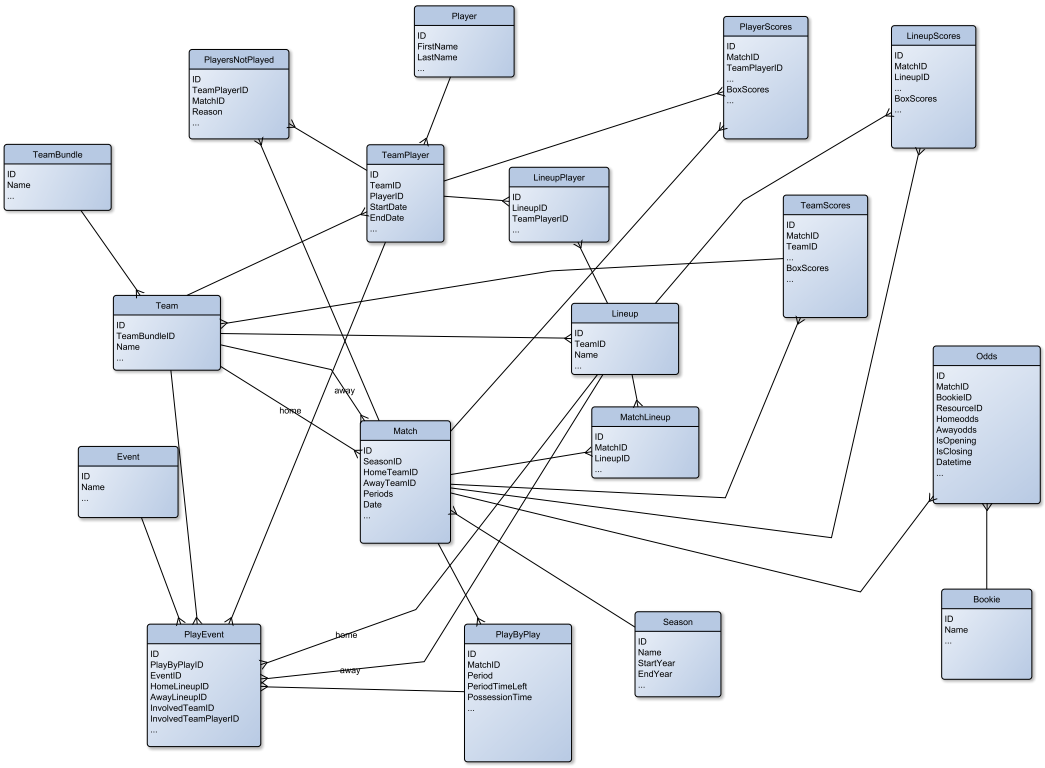
Programmi kood on nähtaval aadressil `https://github.com/K-L/NBA`.

B Autoriõiguse seadus

Järgnevalt on esitatud asjakohased punktid autoriõiguse seadusest [Tea14]: **§ 75^2** kohaselt saame viidatud allikates avaldatud andmete kogu käsitleda andmebaasina, **§ 75^3** alusel kohaldatakse autoriõiguse seaduse peatüki VIII^1 sätteid tulenevalt Berni konventsioonist ka viidatud allikatele ning **§ 75^6** lubab andmete kasutamise õppe- ja teadusliku uurimistöö raames.

§ 75^2. Andmebaasi mõiste
Andmebaas käesoleva peatüki tähenduses on teoste, andmete või muu materjali süstemaatiliselt või metoodiliselt korraldatud kogu, mis on individuaalselt kasutatav elektrooniliste või muude vahendite abil.
§ 75^3. Andmebaasi tegija
(1) Andmebaasi tegija on isik, kes on teinud kas laadilt, väärtuselt või suuruselt olulise investeeringu selle andmebaasi sisuks olevate andmete kogumiseks, omandamiseks, kontrollimiseks, süstematiseerimiseks või kättesaadavaks tegemiseks.
(2) Käesoleva peatüki sätteid kohaldatakse juhul, kui:
3) andmebaasi tuleb kaitsta vastavalt Eesti Vabariigi välislepingule.
§ 75^6. Andmebaasi tegija õiguse piiramine
Üldsusele mis tahes viisil õiguspäraselt avalikustatud andmebaasi õiguspärane kasutaja võib ilma andmebaasi tegija nõusolekuta ja tasu maksmiseta teha väljavõtteid andmebaasi sisu olulisest osast või seda taaskasutada juhul, kui:
2) koos andmebaasi avaldamisallika kohustusliku äranäitamisega tehakse andmebaasist väljavõte illustreeriva materjalina õppe- või teadusliku uurimistöö eesmärkidel, nende eesmärkidega motiveeritud mahus ja tingimusel, et selline kasutamine ei taotle ärilisi eesmärke.

C Andmebaasi skeem



Joonis C.1: Väljavõte andmetest ja nendevahelistest seostest.

Lihtlitsents lõputöö reprodutseerimiseks

Mina, _____Kaido Lepik_____,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
_____
_____“Spordiennustused: kihlveokontoritega konkureerimine NBA-s“_____
_____,
(*lõputöö pealkiri*)

mille juhendaja on _____Jüri Lember_____,
(*juhendaja nimi*)

reprodutseerimiseks ainult säilitamise, sealhulgas digitaalarhiivis DSpace säilitamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni. Lõputöö avaldamine ei ole lubatud.

2. olen teadlik, et punktis 1 nimetatud reprodutseerimise õigus jääb alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **19.05.2014**