TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Laura Ruusmann

Gaussi protsesside usaldusvahemik

Bakalaureusetöö (9 EAP)

Juhendaja: Meelis Kull, PhD

Tartu 2018

Gaussi protsesside usaldusvahemik

Lühikokkuvõte:

Masinõpe on arvutiteaduse valdkond, mis tegeleb arvutisüsteemide oskusega iseseisvalt õppida. Masinõppemeetodeid kasutatakse nii andmete kirjeldamiseks kui ka tunnustele väärtuste ennustamiseks. Kui masinõppemudelit kasutatakse reaalarvulise väärtuse ennustamiseks, siis nimetatakse seda regressiooniks. Praktikas on reaalarvulist väärtust ennustades tihti tarvis arvestada, et vääral ennustusel võivad olla kallid tagajärjed. Väärade ennustuste kahju aitab vähendada see, kui mudel oskab ise hinnata, kui täpne tema ennustus on. Üheks näiteks sellisest hinnangust on tagastada vahemik, kuhu mudel 95% tõenäosusega hindab olevat õige väärtuse. Selline lähenemine on Gaussi protsessidel põhineva regressioonimudeli eriliseks omaduseks ning seda vahemikku nimetatakse usaldusvahemikuks. On oluline, et mudeli hinnang enda täpsuse kohta vastaks tegelikkusele ning et mudel ei hindaks end liiga enesekindlalt.

Masinõppemudelite usaldusväärsuse hindamine on oluline, sest selliste mudelitega tarkvara kätte on tänapäeval usaldatud üha vastutusrikkamate otsuste langetamine. Antud bakalaureusetöö keskendub Gaussi protsessidel põhineva regressioonimudeli enesekindluse uurimisele. Antud töös uuritakse, kui tihti satuvad ennustatavate väärtuste tegelikud väärtused vahemikku, kuhu mudel hindab nende sattumise 95% tõenäosusega.

Mõõtmised 6651 mudelil näitavad, et suurem osa päris märgendeid satuvad usaldusvahemikku oluliselt harvem kui 95% juhtudest ehk et Gaussi protsesside mudel on liigselt enesekindel. Keskmiseks usaldusvahemikku kuulunud osakaaluks on 0,93.

Töö peamine tulemus on, et 73% mõõtetulemustest on madalamad kui võiks olla eelnevalt nimetatud tõenäosuse järgi. Ühtlasi on märkimisväärne see, et kõige väiksemate ja kõige suuremate väärtustega sisendväärtuste puhul on mudel rohkem liigselt enesekindel.

Gaussi protsesside usaldusvahemiku uurimise näol on tegemist millegagi, mida ei ole varem uuritud. Tänu käesolevale tööle on olemas hinnang Gaussi protsesside regressioonimudeli usaldusväärsusele ning selle töö tulemus aitab Gaussi protsesside kasutajatel võtta arvesse antud meetodi liigset enesekindlust.

Võtmesõnad: masinõpe, regressioon, Gaussi protsessid, usaldusvahemik, mudeli enesekindlus

CERCS: P175 informaatika, süsteemiteooria

Confidence of Gaussian Processes

Abstract:
Machine learning is a field in computer science that provides computer systems with the ability to learn independently. Machine learning methods are used for both descriptive and predictive purposes. When a machine learning model is used to predict a real valued number it is called regression. In practice, it is often important in regression to take into account that false predictions might have severe consequences. To avoid such false predictions, it is helpful if the model is able to rate how accurate its prediction is. An example of this is for the model to provide an interval where it predicts the true value with 95% certainty. This approach is unique to Gaussian process regression model and this interval is called confidence interval. It is important that the model rates itself accurately and not overly confidently.

Evaluating confidence of machine learning models is important since software solutions equipped with machine learning algorithms are becoming more common and are being trusted with decisions that require more responsibility. This Bachelor's thesis focuses on the confidence of Gaussian process regression models. This research examines how often are true values contained in the intervals where model predicts them with 95% probability.

Measurement results on 6651 models show that the majority of true labels are included in the confidence interval in less than 95% of cases, which means that Gaussian process regression model is overconfident. Mean ratio of true labels in confidence intervals per model was 0.93.

Main result of the research is that for 73% of the models the confidence interval contained less true labels than was expected by the probability. It is noteworthy that for input values that had smallest or largest values the model was more often overconfident.

Confidence of Gaussian processes has not been researched before and this research provides evaluation on how reliable are Gaussian processes. The results of this thesis enable users of Gaussian processes models to consider overconfidence of models.

Keywords: machine learning, regression, Gaussian processes, confidence interval, model's confidence

CERCS: P175 informatics, systems theory

Sisukord

1 Sissejuhatus **5**

2 Gaussi protsessid **7**
2.1 Definitsioon 7
2.2 Gaussi protsessid masinõppes 8
2.2.1 Ennustatav väärtus 8
2.2.2 Parameetrid 9
2.3 Mudeli enesekindlus 10

3 Enesekindluse uurimine **13**
3.1 Eesmärk 13
3.2 Eksperimendi ülesehitus 14
3.2.1 Tarkvara valik 14
3.2.2 Gaussi protsessid praktikas 15
3.2.3 Andmete kirjeldus 16
3.2.4 Andmete eeltöötlemine 17
3.2.5 Eksperimendi loogika 18
3.2.6 Mõõtmiste tulemuste mõistmine 20
3.3 Tulemused 21
3.3.1 Erinevate tunnuste arvuga treenimine 21
3.3.2 Tulemused vahemike kaupa 26
3.3.3 Erinevate treeningsuurustega treenimine 27
3.3.4 Binoomtesti tulemused 28
3.4 Tulemuste kokkuvõte ja järeldused 33

4 Kokkuvõte **35**

Viidatud kirjandus **38**

Lisad **39**
I. Repositoorium 39
II. Andmestikud 40
III. Litsents 44

1 Sissejuhatus

Masinõppemudelite kasutamine seadmete programmeerimisel on viimasel aastakümnel märkimisväärselt hoogustunud. Robotite tarkvaradele on usaldatud üha vastutusrikkamate otsuste vastu võtmine. On tähtis, et selliste otsuste tegemisel arvesse võetud ennustused oleksid usaldusväärsed. Usaldusväärsuse all on peetud silmas seda, et mudel annab hinnangu ennustuse õigsuse kohta ning et see hinnang on vastavuses sellega, kui õige on ennustus.

Regressiooniülesande puhul on masinõppe abil vastuvõetavaks otsuseks hinnang tundmatule reaalarvule, kasutades selleks vaadeldud andmeid. Selliste hinnangute tegemiseks on üheks võimaluseks kasutada Gaussi protsesside abil loodud regressiooni masinõppemudelit. On oluline hinnata selliste mudelite usaldusväärsust.

Gaussi protsesse kasutatakse tüüpiliselt signaalianalüüsis (Rasmussen, 1999), ent neid on kasutatud ka automaatselt muusika esitusloendite genereerimiseks (Platt jt, 2002), intelligentses monitoorimissüsteemis energiakulu ennustamiseks (Bhinge jt, 2014) ja reaalajas mudelipõhise roboti juhtimises (Nguyen-Tuong ja Peters, 2008). Kõige efektiivsem on Gaussi protsesside kasutamine mittelineaarsete seostega andmete puhul.

Antud uurimus hindab Gaussi protsesside abil loodud mudeli usaldusväärsust läbi selle enesekindluse. Gaussi protsesside abil treenitud regressioonimudeli puhul on ennustamisel olulisel kohal teadmatuse määr ennustatud väärtuse ümber. Seetõttu on iga ennustus varustatud andmetega selle kohta, kui suure tõenäosusega arvab mudel olevat päris tulemuse mingis väärtuste vahemikus. Nimetatud tõenäosuse ja vahemiku abil on võimalik hinnata, kui enesekindel on mudel. Kui mudel hindab läbivalt väga suure tõenäosusega, et päris väärtus kuulub mingisse vahemikku, ent tegelikult see sinna ei kuulu, siis on mudel olnud liiga enesekindel. Mudel võib olla ka parajalt või liiga vähe enesekindel. Mõningate otsuste vastu võtmisel on ennustustes vigade tegemisel tekkiv kahju äärmiselt suur ning sellisel juhul on tarvis omada täpset hinnangut selle kohta, millisel määral võib ennustus olla väär. Kui mudel on liigselt enesekindel, siis ta on oma ennustustes rohkem väär, kui võiks eeldada ning sellisel juhul on kahjulike otsuste tegemine sagedasem.

Varem on läbi viidud uurimusi Gaussi protsesside soorituse kohta, ent need on peamiselt keskendunud Gaussi protsesside jõudlusele, millest siin töös on lähemalt kirjutatud peatükis 3.2.2. Raamatu ”Gaussian Processes for Machine Learning” üks autoritest (Rasmussen, 1999) on hinnanud Gaussi protsesse ja muid mittelineaarseid

regressioonimudeleid. Tegemist on ulatusliku tööga, mis uuris peamiselt Gaussi protsesside kasutamise jõudlust sünteesitud ning päris andmetel. Kuna Gaussi protsesside kasutamise üheks suurimaks miinuseks on mudeli treenimise keerukus, on läbi viidud palju uurimusi selle kohta, kuidas ligikaudsete mudelite treenimise jõudlust parandada (Nguyen-Tuong ja Peters, 2008; Schreiter, Englert jt, 2015; Schreiter, Nguyen-Tuong jt, 2016). Antud bakalaureusetöö valmimise hetkel ei ole läbi viidud uuringut Gaussi protsesside usaldusvahemike täpsuse kohta. Erinevalt jõudluse testimisest, keskendub antud töö usaldusvahemiku hindamisele. Selle tulemused täiendavad varasemaid uurimusi Gaussi protsesside sooritusvõime kohta.

Bakalaureusetöö koosneb viiest peatükist. Sissejuhatusele järgnev peatükk 2 kirjeldab teooriat Gaussi protsesside olemusest ning nende kasutamisest masinõppes. Kolmandas peatükis on sõnastatud konkreetne eesmärk, kirjeldatud Gaussi protsesside enesekindluse uurimise eksperimendi ülesehitust ja tehtud kitsendusi ning seejärel saadud tulemusi ja järeldusi. Neljandas ehk viimases peatükis on kokkuvõte kogu tööst, kaasa arvatud uurimuse järeldused ning võimalused edasiseks uurimiseks. Töö lisadest leiab lingi repositooriumile, mis sisaldab lähtekoodi ja kogutud andmeid, kasutatud andmestike kirjeldusi ning litsentsi.

2 Gaussi protsessid

Järgnev materjal tugineb raamatule "Gaussian Processes for Machine Learning" (Rasmussen ja Williams, 2006). Antud peatükk selgitab Gaussi protsesside olemust masinõppes ning nendega seostuvaid olulisemaid märksõnu ja kontseptsioone.

2.1 Definitsioon

Gaussi protsess on mitmemõõtmelise Gaussi jaotuse ehk normaaljaotuse üldistus lõpmatule arvule mõõtmetele. Et mahutada see üldistus võrdlusesse vähemamõõtmeliste Gaussi jaotustega, siis võib jagada Gaussi jaotused mõõtmete järgi kolmeks. Ühemõõtmeline Gaussi jaotus näitab tõenäosuse jaotust juhuslikul reaalarvulisel suurusel. Lõpliku arvu mõõtmetega mitmemõõtmeline Gaussi jaotus näitab jaotust juhuslikel suurustel, millest saab koostada vektoreid (paare, kolmikuid, nelikuid, jne). Lõpmatu arvu mõõtmetega juhuslikud suurused moodustavad lõpmatu pikkusega vektoreid. Jättes kõrvale matemaatilised detailid, võib lõpmatu pikkusega vektorist mõelda kui funktsioonist $f(x)$ ning et selle vektori iga liige tähistab nimetatud funktsiooni väärtust mõne konkreetse sisendi x korral. Joonisel 1 (a) on toodud näide sellistest funktsioonidest. Öeldakse ka, et Gaussi protsess kirjeldab jaotust üle funktsioonide - see tähendab, funktsioonid on omavahel normaaljaotusega. Sisendi igale koordinaadile vastav juhuslik suurus on seega normaaljaotusega, sest seda kirjeldavad funktsioonid on normaaljaotusega. Need funktsioonid ongi Gaussi protsesside väljundiks. Rasmusseni ja Williamsi raamatus on Gaussi protsess defineeritud järgnevalt:

Definitsioon 1. *Gaussi protsess on lõpmatu kogu juhuslikest suurustest, kusjuures iga lõplik arv neist allub mitmemõõtmelisele Gaussi jaotusele.*

Gaussi protsess on täielikult määratletud selle keskväärtuse ning kovariatsioonifunktsiooniga (*kernel*, ingl k) ning selle kirjaviis on

$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}); k(\mathbf{x}, \mathbf{x'})\big),$$

kus $\mathbf{x}$ on sisendväärtuste vektor, $m(\mathbf{x})$ keskväärtusfunktsioon ning $k(\mathbf{x}, \mathbf{x'})$ kovariatsioonifunktsioon. Kahe muutuja kovariatsioon kirjeldab, kuidas nende väärtused koos muutuvad. Lihtsustatult võib kovariatsioonifunktsiooni mõista kui midagi, mis määrab Gaussi protsessi loodud juhuslike funktsioonide kuju - kas need on sakilised, lainelised

või mingi muu kujuga. Enamasti eeldatakse, et funktsioonide jaotuse keskväärtuseks $m(\mathbf{x})$ on 0. Selle väärtusega arvestamine on äärmiselt oluline, kui on soov kasutada Gaussi protsesse masinõppes.

2.2 Gaussi protsessid masinõppes

Gaussi protsesse kasutatakse masinõppes mitmete ülesannete lahendamiseks. Antud uurimus keskendub Gaussi protsesside abil juhendatud õpet kasutades loodud regressiooni masinõppemudeli uurimisele. Juhendatud õppe abil loodud masinõppemudelist võib mõelda kui sisendi ja väljundi vastavusse seadmisest ning selle vastandiks on juhendamata õpe, kus puuduvad teadaolevad väljundtunnused. Regressiooniülesande puhul on väljundiks pidevad reaalarvulised väärtused.

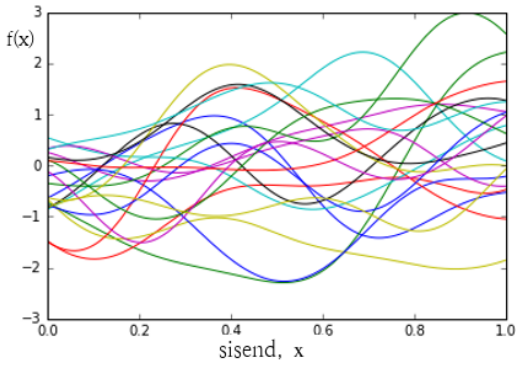
2.2.1 Ennustatav väärtus

Gaussi protsessid tagastavad regressiooniülesande lahenduseks ehk ennustuseks ühe reaalarvu asemel terve jaotuse. Tagastatavat jaotust kirjeldavad selle keskväärtus ning usaldusvahemik, kusjuures keskväärtus on see, mida loetakse ennustatud väärtuseks ning usaldusvahemik kirjeldab teadmatuse määra antud ennustuses.

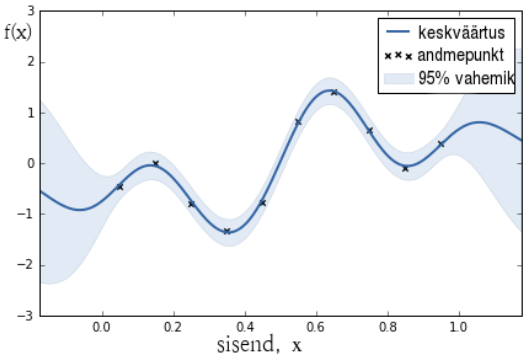
Gaussi protsesside mudeli treenimisel kasutatakse Bayesi statistika lähenemist. Bayesi statistika kohta on kirjutatud (Traat ja Lepik, 2013), et see tähendab, et otsuste tegemiseks kasutatakse lisaks vaatlusandmetele ka eelinformatsiooni. Esialgu luuakse enne andmestiku nägemist Gaussi protsesside abil juhuslikud funktsioonid eeljaotuseks (*prior*, ingl k) ning seejärel kombineeritakse neid Bayes'i teoreemi kasutades treenimiseks kasutatud andmetega, et leida järeljaotus (*posterior*, ingl k).

Valdavalt kasutatakse regressioonimudelite täpsuse hindamiseks ruutjuurt ruutkeskmisest veast (*root mean squared error - RMSE*, ingl k) ning see on väärtus, mida regressioonimudelite õppimisel üritatakse minimeerida (Martino jt, 2017). Gaussi protsesside puhul toimub sobivaima regressioonijoone ehk järeljaotuse õppimine maksimeerides marginaalset tõepära (*marginal likelihood*, ingl k). Otsitakse vastust küsimusele, milline regressioonijoon muudaks juba vaadeldud andmed kõige tõenäolisemaks. Kuigi võib jääda mulje, et marginaalse tõepära maksimeerimine jätab tähelepanuta mudeli tehtud vea minimeerimise, siis tegelikult on need tegevused ekvivalentsete tulemustega (Bailey, 2016). Kusjuures, olles ära valinud eeljaotuse, on oluline jälgida, et ka ennustatava tunnuse keskväärtus oleks sama ning kui ei ole, on soovitatav see selliseks töödelda.

Joonisel 1 (a) on kujutatud, milline võiks välja näha eeljaotus ning on visualiseeritud lõplik mudel järeljaotusena (b).



(a) Gaussi protsessid eeljaotusena. Joonis kujutab 20 juhuslikku funktsiooni keskväärtusega 0.



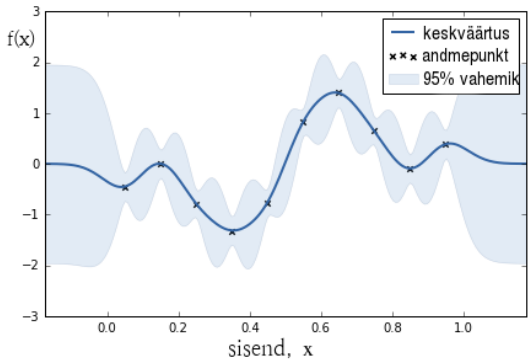
(b) Gaussi protsessid järeljaotusena. Musta värvi täpid on treeningandmed, sinine joon on regressioonijoon ning helesinine ala näitab usaldusvahemikku tõenäosusel 0,95.

Joonis 1. Gaussi protsesside eeljaotus ja järeljaotus.

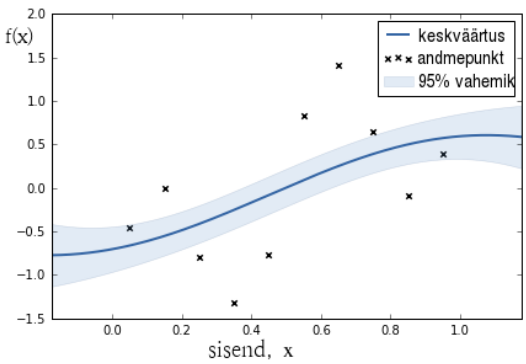
2.2.2 Parameetrid

Gaussi protsessid on mitteparameetrilised. Parameetrilised mudelid eeldavad ennustatava tunnuse ja ennustamiseks kasutatavate tunnuste vahele mingit kindlat seost. Sellisteks mudeliteks on näiteks lineaarsed mudelid, mis eeldavad, et andmed on lineaarfunktsiooniga kirjeldatavad. Gaussi protsessid taolisi eeldusi ei tee ning seetõttu ei pea muretsema selle pärast, kas mudelil on võimalik sobituda andmestikuga.

Gaussi protsessidel on võimalik määrata hüperparameetreid, mis kirjeldavad kovariatsioonifunktsiooni. Neid parameetreid on üldjuhul kaks: pikkusskaala ℓ (*length-scale*, ingl k) ning dispersioon v. Pikkusskaala näitab, kui lähestikku peavad andmepunktid olema, et nende vahel oleks arvestatav korrelatsioon ning dispersioon kirjeldab andmestiku mürasust. Joonisel 2 on kujutatud muus mõttes identse, aga ainult erineva pikkusskaalaga kovariatsioonifunktsiooniga treenitud mudelid.



Väikese pikkusskaalaga, $\ell = 0.06$



Suure pikkusskaalaga, $\ell = 1.0$

Joonis 2. Erinevate pikkusskaaladega kovariatsioonifunktsioonid. Musta värvi punktid on treeningandmed, sinine joon on regressioonijoon ning helesinine ala näitab usaldusvahemikku tõenäosusel 0,95. Samadel andmetel treenitud paraja pikkusskaalaga mudel on joonisel 1 (b).

Liiga väikese pikkusskaala puhul on näha, et mudeli usaldusvahemik on kitsas kõikide andmepunktide juures, sest iga vaadeldud andmepunkti jaoks on teised andmepunktid liiga kaugel, et nende vahele kindlat seost luua ehk teisisõnu, pikkusskaala on lühem kui andmetevaheline kaugus. Samas liiga suure pikkusskaalaga funktsiooni puhul võtab mudel arvesse andmepunktist väga kaugel asuvaid punkte ning tulemuseks saadud mudel on pigem üldine.

2.3 Mudeli enesekindlus

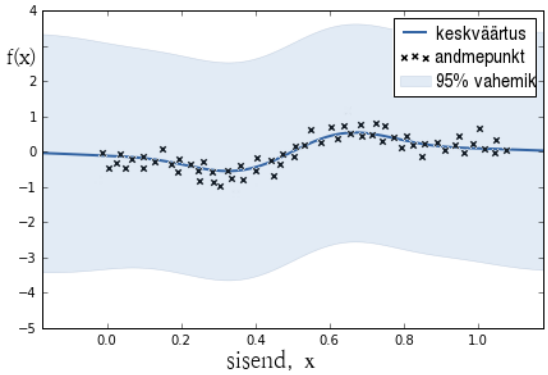
Gaussi protsesside regressioonimudel ennustab tunnuse väärtuse, ent sellega käib alati kaasas ka usaldusvahemik. Erinevalt klassikalisest statistikast, kus usaldusintervalli puhul parameeter on fikseeritud ning hinnatakse erinevuse olulisust sellest parameetrist, siis Bayesi statistikas tähistab usaldusvahemik intervalli, kuhu kuulub mittefikseeritud parameeter mingi tõenäosusega ning mida oleks korrektsem nimetada tõenäosusintervalliks (*credible interval*, ingl k) (Traat ja Lepik, 2013). Ent kuna Rasmussen ja Williams kasutavad oma raamatus selle kirjeldamiseks usaldusvahemiku terminit, on ka käesolevas uurimuses otsustatud seda terminoloogiat järgida.

Usaldusvahemik on hea indikaator mõistmaks seda, kui enesekindlalt mudel min-

git väärtust tunnusele ennustab. Näiteks tõenäosusega 0,95 usaldusvahemik tähendab seda, et mudel hindab, et 95% tegelikest ennustatavatest väärtustest ehk märgenditest langevad usaldusvahemikku. Kui tegelike märgendite, mis kuuluvad usaldusvahemikku, osakaal on sama suur, kui usaldusvahemikku kirjeldav tõenäosus, võib väita, et mudel on parajalt enesekindel. Kui tegelike märgendite osakaal erineb tõenäosusest, et nad kuuluvad usaldusvahemikku, on tegemist kas liigselt enesekindla või liiga vähe enesekindla mudeliga.

Joonisel 2 kujutatud suure pikkusskaalaga mudel on hea näide mudelist, mis on liiga enesekindel, kui leppida kokku, et joonisel kujutatud andmepunktid on ennustatavad väärtused, mida mudel ei ole treenimiseks kasutanud. Usaldusvahemikku kujutav helesinine ala peaks sisaldama 95% nendest ennustatavatest väärtustest, samas on näha, et sinna kuulub 10-st andmepunktist ainult kaks ehk 20%.

Näite liiga vähe enesekindlast mudelist leiab jooniselt 3. Erinevalt liiga enesekindlast mudelist on näha, et kõik andmepunktid kuuluvad ennustatud vahemikku, kusjuures usaldusvahemik on kordades laiem kui vahemik, kuhu kõik väärtused kuuluvad. Selline mudel võiks olla enesekindlam ehk antud juhul kitsama usaldusvahemikuga.



Joonis 3. Näidis liiga vähe enesekindlast mudelist.

Erinevus üle- ja alasobitamisest Masinõppes nimetatakse mudeli ülesobitamiseks seda, kui mudeli ennustused on treeningandmestikul oluliselt täpsemad kui testandmestikul. Mudeli treenimisel on see liigselt sobitatud mürale treeningandmestikus. Seevastu alasobitatud mudel on liiga üldine, et tabada andmetevahelist seost.

On oluline mõista erinevust mudelitel, mis ei ole parajalt enesekindlad ja mudelitel,

mis on üle- või alasobitatud. Hinnang mudeli enesekindluse kohta ei pruugi kuidagi seostuda tema üle- või alasobituvusega. Joonisel 2 kujutatud suure pikkusskaalaga mudel tundub olevat alasobitatud - nimelt joon on väga lauge ega arvesta kõiki andmestikus peituvaid nüansse. Samas, võib-olla need nüansid on liigne müra ning võib-olla on see täpselt sobiv joon andmete kirjeldamiseks, ilma et see oleks liigselt sobitatud treeningandmete. Küll aga mõlemal juhul peaks olema usaldusvahemik oluliselt laiem. Nii ülekui ka alasobitatud mudelid võivad olla liigselt enesekindlad.

3 Enesekindluse uurimine

3.1 Eesmärk

Uurimuse eesmärgiks on anda hinnang Gaussi protsesside abil loodud regressioonimudeli enesekindlusele ning leida põhjendusi sellele hinnangule.

Kahes kategoorias uurimine Mudelite enesekindluse hindamine toimub kahes kategoorias.

Üheks kategooriaks on uurida, kui suur hulk tegelikest märgenditest langeb neile ennustatud usaldusvahemikku ning teha seda mitmete mudelite peal.

Teine kategooria on selle täiendus - uurida, kuidas sõltub mudeli enesekindlus sellest, mis väärtuste vahemikku kuulub sisendtunnus. Selle illustreerimiseks on hea vaadata joonist 1 (b) - nimelt on näha, et kui sisendi väärtus on väiksem kui 0.0 või suurem kui 1.0, siis mudeli usaldusvahemik muutub ekstrapoleerimise tõttu väga suureks, näidates selle ebakindlust. Selle nähtuse uurimiseks on jaotatud sisendtunnus väärtuste järgi vahemikeks, millest iga puhul leitakse osakaal märgenditest, mis kuuluvad usaldusvahemikku. Kasutatakse vaid ühe ja reaalarvulise sisendtunnusega treenitud mudelitel.

Kitsendused Kuigi selline lähenemine annaks kõige põhjapanevama tulemuse, siis paraku on teostamatu treenida testimiseks kõik võimalikud mudelid kõikidel võimalikel andmestikel ja seetõttu on tehtud antud uurimuses kitsendusi.

Uurimus keskendub regressiooniülesannetele, mis on treenitud kas ühe, kahe või kõikide andmestikus olevate tunnuste peal. Väljundiks on alati üks reaalarvuline väärtus. Ühel andmestikul saab lahendada mitmeid regressiooniülesandeid ja seetõttu on piirdutud kolme andmestikuga, millel treenida mitmeid mudeleid erinevate sisendtunnuste ja väljundtunnuste kombinatsioonidega. Regressiooniülesannete puhul on olnud eesmärgiks treenida võimalikult palju mudeleid, et nende enesekindlust hinnata. Sellise eesmärgi täitmiseks on vähe oluline, et lahendatavad regressiooniülesanded oleksid semantiliselt loogilised. See tähendab, et nii mõnigi mudel võib olla selline, mille treenimiseks keegi põhjust ei näe, ent antud töös vaadeldakse neid regressiooniülesandeid ikkagi, sest käesolevas töös aitavad ka rakendustes mittevajalikel ülesannetel saadud tulemused teha üldjäreldusi.

Lisaks kombinatsioonidele treeninguks kasutatud tunnuste hulgas, on varieeritud

treeningandmete hulga suurust ning valitud suurusteks $\{1000, 2500, 5000\}$ ning antud suuruste valiku tagamaid on avatud peatükis 3.2.2.

Kõikidel treenitud mudelitel on kasutatud sama kovariatsioonifunktsiooni, ent selle parameetreid on optimeeritud vastavalt regressiooniülesandele. Kõikide mudelite puhul on väljundtunnustele eeldatud sama keskväärtust, milleks on 0. See tähendab, et eeljaotus luuakse sellise keskväärtusfunktsiooniga, mis on konstantselt 0. Tegemist on Gaussi protsesside puhul tavapärase tegevusega (Rasmussen ja Williams, 2006). Optimeerimist on alustatud korduvalt ning kuni 10 korda juhuslike väärtustega, et vältida lokaalsesse miinimumi sattumist (Murray, 2016).

Töö keskendub osale usaldusvahemikust, mis omab 95% tõenäosust, et päris märgend kuulub selle piiridesse.

Sisendi väärtuste vahemike kaupa (eelnevas paragrahvis kirjeldatud teises uurimiskategoorias) mõõdetakse enesekindlust vaid ühelt sisendtunnuselt ennustades, kui sisendtunnus on reaalarvuline. Sellisel juhul on testväärtused jaotatud kümneks vahemikuks.

Antud uurimuse eesmärk ei ole katsetada erinevate Gaussi protsesside implementatsioonide jõudluse ega enesekindluse erinevusi. Uurimusse on valitud Gaussi protsesside tarkvaraline implementatsioon, mis võimaldab protsesse uurida vastavalt eelnevalt välja toodud kitsendustele.

3.2 Eksperimendi ülesehitus

3.2.1 Tarkvara valik

Gaussi protsesside usaldusvahemiku täpsuse mõõtmiseks kirjutati skript programmeerimiskeeles Python 3. Terve skript ning saadud tulemused on saadaval lisas I. Mudelite treenimiseks kasutati Pythoni masinõppe raamistikku nimega scikit-learn v0.19.1. Antud teek võimaldab Gaussi protsesside abil regressioonimudeli treenimist kasutades klassi `sklearn.gaussian_process.GaussianProcessRegressor` (Pedregosa jt, 2011). Antud klassi abil loodud mudelilt on võimalik ennustades küsida lisaks jaotuse keskväärtusele ka selle standardhälbe, millest on võimalik leida usaldusvahemik (Pedregosa jt, 2011).

Alternatiivseks valikuks oli teek GPy (GPy, 2012). GPy suureks plussiks oleks võimalus kasutada treenimiseks hõredat mudelit. Hõredatest mudelitest on pikemalt kirjutatud peatükis 3.2.2. Ent kuna kõnealune teek omab suurt puudujääki, sest ei toeta mitme sisendtunnusega mudeli treenimist, otsustati scikit-learni kasuks.

3.2.2 Gaussi protsessid praktikas

Järgnev alapeatükk kirjeldab väljakutseid ja valikuid Gaussi protsesside praktikas kasutamisel.

Treenimise keerukus Teadlased nendivad (Hensman jt, 2013), et paraku on Gaussi protsesside mudeli treenimine keerukusega $\mathcal{O}(n^3)$, sealjuures mälulise keerukusega $\mathcal{O}(n^2)$, kus n on andmestiku suurus ning et Gaussi protsesside puhul peetakse suureks andmestikuks juba seda, kui see sisaldab paar tuhat andmepunkti ning isegi selliste andmestike peal treenimiseks on tarvis rakendada ligikaudseid meetodeid. Hensman jt kirjeldavad, kuidas levinuim nendest meetoditest on hõreda (*sparse*, ingl k) mudeli treenimine. Harilikult on hõreda Gaussi protsesside mudeli treenimine keerukusega $\mathcal{O}(nm^2)$ ja ruumilise keerukusega $\mathcal{O}(nm)$, kus m on kasutaja valitud parameeter kirjeldamaks treenimisel kasutatavate punktide arvu. Antud uurimus keskendub tavalisele Gaussi protsesside treenimise meetodile. Gaussi protsesside skaleerimine suurtele andmestikele on valdkond, mis vajab endiselt aktiivset uurimistööd (Gal jt, 2014).

Kovariatsioonifunktsiooni optimeerimine Antud uurimuses kasutatakse kõikides regressiooniülesannetes sama kovariatsioonifunktsiooni. Tegemist on kahe kovariatsioonifunktsiooni liitmisel saadud funktsiooniga.

Üheks osaks sellest on radiaalsete baasfunktsioonide meetod (*radial basis function - RBF*, ingl k), tuntud ka Gaussi *kernel*-ina või ruuteksponentsiaalse kovariatsioonifunktsioonina (*squared exponential kernel*, ingl k) (Duvenaud, 2014), mis on ühtlasi scikit-learn teegi puhul vaikimisi kasutatav kovariatsioonifunktsioon.

Et eksperiment viiakse läbi päris andmetel, võib eeldada, et andmestik sisaldab müra ehk esineb kasutut lisainformatsiooni. Küll aga ei ole teada, millisel määral andmestik sellist informatsiooni sisaldab. Seetõttu on RBF kovariatsioonifunktsioon kombineeritud scikit-learn *kernel*-iga `sklearn.gaussian_process.kernels.WhiteKernel`, mille kohta seisab dokumentatsioonis, et see kirjeldab sisendi müra komponenti. Kovariatsioonifunktsioonide kombineerimine on tavapärane tegevus Gaussi protsesside puhul ning kasutatavad kovariatsioonifunktsioonid valitakse vastavalt sellele, mida on andmestiku kohta teada.

Ei saa eeldada, et vaikimisi väärtused kovariatsioonifunktsiooni parameetritele oleksid sobilikud konkreetsele mudelile ning seetõttu on neid tarvis optimeerida. Optimee-

ritavad hüperparameetrid on kirjeldatud varasemas peatükis 2.2.2. Scikit-learn teegis kasutatakse optimeerimiseks L-BFGS-B algoritmi (Pedregosa jt, 2011).

3.2.3 Andmete kirjeldus

Eksperimendi läbiviimiseks valiti välja kolm andmestikku, mis on sobilikud regressiooniülesannete lahendamiseks. Usaldusvahemiku uurimiseks on vaja lahendada mitmeid regressiooniülesandeid ja seetõttu on tarvis, et andmestik sisaldaks reaalarvulisi väärtusi, mida ennustada. Mida enam on tunnuseid, seda rohkem saab sellel andmestikul regressiooniülesandeid lahendada. Seetõttu valiti andmestikke selle järgi, et need sisaldaksid reaalarvulisi tunnuseid ning et tunnuste koguarv oleks vähemalt 10. Enam kui kümnest tunnusest saab kombineerida kolmel andmestikul rohkelt regressiooniülesandeid. Pööra-ti tähelepanu ka andmestiku suurusele ning eelistati andmestikke, mille suurus ületab 10000 andmerida. Andmestikud on pärit UCI Machine Learning Repository (Dheeru ja Karra Taniskidou, 2017). Andmestikud on kirjeldatud kolmes järgnevas paragrahvis.

Laevade jõuseadmete seisundipõhise hoolduse andmestik. Valitud andmestik kirjeldab laevade jõuseadmete seisundipõhist tehnilist hooldust. Edaspidi on seda andmestikku nimetatud laevaandmestikuks. Andmestiku autorite sõnul (Coraddu jt, 2016) on tegemist simuleeritud andmetega, mis on kooskõlas päriselt võimaliku laeva näitudega. Andmestikus on 11934 andmerida ning kokku 18 tunnust, millest kõik on reaalarvuliste väärtustega. Andmete simuleerimiseks on kasutatud lisaks laeva kiirusele ka turbiini ja kompressori ajas kulumise koefitsiente. Tunnused ning nende kirjeldused on toodud lisas II tabelis 3. Et tegemist on tehnilise sõnavaraga, on nende tunnuste kirjeldused originaalkeeles toodud välja lisas II tabelis 4.

Jalgratta jagamise andmestik. Jalgratta jagamise andmestik (Fanaee-T ja Gama, 2013) sisaldab automatiseeritud rattalaenutuse kogutud andmeid rataste laenutuste ning kohalike ilmastiku- ja keskkonnaolude kohta aastatel 2011-2012 Ameerika Ühendriikides Washingtonis. Edaspidi on seda andmestikku nimetatud rattaandmestikuks. Nendel andmetel saab ennustada rattalaenutuse hulka ilmastiku- ja keskkonnaolude põhjal. Andmestik 17379 andmerida, millest iga rida kirjeldab ühte tundi. Andmestikus on 17 tunnust, millest 14 on sisendiks ja 3 väljundiks, ning nende kirjeldused on toodud lisas II tabelis 5.

Veini kvaliteedi andmestik. Veini kvaliteedi andmestik (Cortez jt, 2009) (edaspidi veiniandmestik) sisaldab andmeid valge Portugali "Vinho Verde" veini keemiliste näitude ning nende kvaliteedi hinnangu kohta nullist kümneni süsteemis. Andmestikus on 11 reaalarvulist atribuuti, mille põhjal saab ennustada 12-ndat ehk hinnangut veini kvaliteedile. Valge veini andmestikus on 4898 andmerida. Antud andmestiku tunnused eestikeelsete selgitustega leiab lisast II tabelist 6.

3.2.4 Andmete eeltöötlemine

Sisendtunnuste töötlemine Laevaandmestik sisaldab kompressori ja turbiini kuluvuskoefitsiente, mida koos kiirusega (`v`, atribuutide lühendid siin ja edaspidi originaalses kirjapildis) on kasutatud andmete simuleerimiseks.

Andmestikuga kaasa tulevas kasutamist tutvustavas tekstis puudub kirjeldus selle kohta, millised võiksid olla andmestikul ennustatavad tunnused ning seetõttu on siin eksperimendis kasutatud väljundtunnustena kõiki reaalarvulisi tunnuseid. Andmestiku hilisemas täiendatud väljaandes on täpsustatud, et on soovitud ennustada jõuseadme kuluvuse seisu, mida kirjeldavad varasemalt mainitud kuluvuskoefitsiendid (ja uuendatud versioonis ka teistsugused koefitsiendid) (Cipollini jt, 2016), seega on otsustatud lisaks kõikidele reaalarvulistele väärtustele ennustada ka koefitsiente, kasutamata neid sisendtunnustena. Koefitsiendid on jäetud sisendtunnuste hulgast välja põhjusel, et tegemist on andmete simuleerimiseks kasutatud väärtustega ning päris laeval selliseid näite ei mõõdeta. Sellegipoolest, kuna terve andmestik on simuleeritud just selleks, et ennustada laeva jõuseadme kuluvustaset, siis on neid kasutatud väljundtunnustena. Väljundtunnustena on kasutatud ka kõiki teisi reaalarvulisi tunnuseid. Seega osad tunnused on kasutatud nii sisendi kui ka väljundina. Sisendtunnuste hulgast on jäetud välja `P1` ja `T1`, sest nende väärtus on konstantne. Selles andmestikus on kokku kasutatud 14 tunnust nii sisendi kui ka väljundina ning kahte ainult väljundina.

Jalgrattalaenutuse andmestikust on otsustatud lugeda ebaoluliseks tunnust `instant` ehk kirje indeksit. Tegemist on metatunnusega, mis kirjeldab andmeid ja mitte nende sisu. Samuti on ebaoluliseks peetud kuupäeva kirjeldavat tunnust `dteday`. Nii aasta kui ka kuu on toodud eraldi tunnusena.

Rattaandmestiku kasutamine on mõeldud nii, et saab tunnuseid `cnt`, `casual` ja `registered` ehk laenutatud rataste arvu ülejäänud 12 põhjal ennustada. Regressiooniülesandeks sobib ka ennustada registreeritud kasutajate laenutuste arvult (`registered`)

kogu laenutuste arvu ja vastupidi.

Standardiseerimine Gaussi protsessid eeldavad, et ennustatava väärtuse keskväärtus on 0 - nimelt luuakse eeljaotus, mis vastab sellisele eeldusele. Kuna antud andmestike puhul ei ole ükski ennustatav väärtus sellise keskväärtusega, on soovitatud väljundtunnused eelnevalt standardiseerida (Murray, 2016). Andmete standardiseerimiseks on kasutatud masinõppe abivahendite teegi mlxtend (Raschka, 2016) meetodit `mlxtend.preprocessing.standardize()`, mis teostab veeru-põhist standardiseerimist NumPy massiivi peal, määrates selle keskväärtuseks 0 ning standardhälbeks 1.

3.2.5 Eksperimendi loogika

Usaldusvahemiku testimiseks on kavandatud algoritm. Algoritm on implementeeritud Pythoni programmina. Eksperimendi ülesehitus ning selleks kasutatav skript eeldab varasemaid teadmisi selle kohta, mis on Gaussi protsessid ning millised on nende rakendamise parimad praktikad, mis on välja toodud varasemas peatükis 2.

Tunnuste valimine. Kõigepealt valitakse tunnused, millel treenida ning tunnus, mida ennustada. Eksperimendis on otsustatud katsetada ennustamist ühe, kahe või kõikide tunnuste põhjal. Seega tunnuste valimise punktis toimub erinevate kombinatsioonide läbi proovimine.

Andmestiku jaotamine test- ja treeningandmeteks Olles valinud välja tunnused, valitakse välja hulk andmeid, mida kasutatakse treeningandmetena. Selle alamhulga valimisel arvestatakse asjaolu, et need oleksid võetud juhuslikult üle terve andmestiku, et vältida võimalikke mõjutusi andmestiku järjestusest. Mudeli treenimise järel on tarvis seda testandmete peal katsetada ning selleks valitakse testandmete hulk. Testandmed valitakse nii, et need ei kattu treeningandmetega, kusjuures treeningandmetega mittekattumine tähendab, et kui andmestikus on korduvate väärtustega andmeüksuseid, võis siiski üks neist esineda treeningandmetes ja teine testandmetes.

Mudeli treenimine. Seejärel treenitakse nendel andmetel Gaussi protsesse kasutades regressioonimudel. Mudeli treenimiseks kasutatakse teegi scikit-learn klassi `sklearn.gaussian_process.GaussianProcessRegressor` ning sellest on näide allpool kujutatud koodiplokis.

```
import sklearn.gaussian_process as GP

kernel = GP.kernels.RBF() + GP.kernels.WhiteKernel()
GPm = GP.GaussianProcessRegressor(n_restarts_optimizer=10, kernel=kernel)
GPm.fit(trainx, trainy)
```

Hüperparameetrite optimeerimine toimub mudeli sobitamise ajal ehk meetodi `fit()` väljakutsel (Pedregosa jt, 2011). Argument `n_restarts_optimizer` kirjeldab, mitu korda alustatakse juhuslike väärtustega optimeerimist. `trainx` ja `trainy` tähistavad vastavalt sisend- ja väljundtunnuseid, millel mudelit treenitakse.

Testandmetel kvantiilide ennustamine. Mudelit kasutatakse testandmetel ennustamiseks ning seejärel küsitakse mudelilt standardhälve igale punktile testandmestikus. Standardhälbest konstrueeritakse usaldusvahemik tõenäosusel 0,95, kasutades selleks jaotuse kvantiile $q_{0,025}$ ja $q_{0,975}$ - just nende vahele jääb 95% usaldusvahemik (Traat ja Lepik, 2013). Nimetatud kvantiilide leidmiseks saab kasutada standardhälvet ning tõenäosusele 0,95 vastavat standardskoori, mille väärtus on ligikaudu 1,96; kvantiilide leidmise näide on olemas ka scikit-learni Gaussi protsesside kasutamise näidises (Pedregosa jt, 2011). Normaaljaotuse keskväärtusest ja standardhälbest usaldusintervalli piiride leidmise valem näeb välja järgmine:

$$(q_{0,025}; q_{0,975}) \approx (\mu - 1,96 \cdot \sigma; \mu + 1,96 \cdot \sigma), \tag{1}$$

kus μ tähistab jaotuse keskväärtust ja σ standardhälvet. Valemi (1) implementatsioon on kuvatud alljärgnevas koodiplokis ning seda on rakendatud iga testväljundiks ennustatud keskväärtusele ja standardhälbele, st alljärgnevas implementatsioonis on tehted vektoritega.

```
mean, std = GPm.predict(testx, return_std=True)
low, up = mean - 1.96 * std, mean + 1.96 * std
```

Õigesti ennustatud osahulga leidmine. Usaldusintervalle võrreldakse tegelike testandmete märgenditega ning leitakse osakaal sellistest märgenditest, mis tõepoolest kuuluvad mudeli poolt esitatud usaldusvahemikku. Defineerime indikaatorfunktsiooni I_{CI}

järgnevalt:

$$I_{CI}\colon X \to \{0;1\} \text{ nii, et}$$
$$\begin{cases} I_{CI}(x_i) = 1 & \text{kui } x_i \in CI \\ I_{CI}(x_i) = 0 & \text{kui } x_i \notin CI, \end{cases} \tag{2}$$

kus CI tähistab usaldusvahemikku. Siinkohal kasutatava funktsiooni väärtuseks on 1, kui päris märgend kuulus usaldusvahemikku ning 0, kui ei kuulunud.

Korduv testimine. Kuna üks mõõtmine ei näitaks usaldusintervallide täpsust piisavalt ülevaatlikult, on tarvis eelnevaid samme korduvalt teha, alustades uute treeningandmete valimisest.

3.2.6 Mõõtmiste tulemuste mõistmine

Kui tegelike märgendite, mis kuuluvad usaldusvahemikku, osakaal on sama suur, kui usaldusvahemikku kirjeldav tõenäosus, võib väita, et mudel on parajalt enesekindel. Kui märgendite osakaal on väiksem, kui usaldusvahemiku tõenäosuse järgi peaks olema, on tegemist liigselt enesekindla mudeliga ning vastupidiselt — kui märgendite osakaal on suurem, kui usaldusvahemiku tõenäosuse järgi oleks täpne — on tegemist liigselt madala enesekindlusega mudeliga.

Arvestades indikaatorfunktsiooni valemit (2) ja teades, et positiivse tulemuse saamise tõenäosus on 0,95, saab ühest katsest mõelda kui Bernoulli jaotusega juhuslikust suurusest

$$Y \sim \mathcal{B}(1; 0,95).$$

Et aga mõõtmisel on kogu testandmestik, toimub neid katseid sama palju, kui on testandmeid. Sellisel juhul on tegemist korduvate Bernoulli katsetega ning rakendades indikaatorfunktsiooni igale mudeli poolt ennustatud usaldusvahemikule ja vastavale päris märgendile, on tulemuseks binaarne vektor

$$\mathbf{v} = \big(I_{CI_1}(y_{*_1}); I_{CI_2}(y_{*_2}); \ldots; I_{CI_n}(y_{*_n})\big),$$

kus n tähistab testandmestiku suurust ning y_{*_i} tähistab selle i-ndat andmerida. Sellest vektorist on võimalik leida positiivsete tulemuste koguarv ning võrrelda nende osakaalu tõenäosusega, et tuleb positiivne arv. Seejärel on võimalik uurida, kas positiivsete ehk usaldusvahemikus olevate päris märgendite arv erineb oluliselt nende esinemise

tõenäosusest. Tegemist on Bernoulli jaotuse laiendamisega mitmele katsele ehk binoomjaotusega. Sellest lähtuvalt on uuritud mõõdetud tulemuste olulisust binoomtestiga. Binoomtest võimaldab uurida, kas lubatud tõenäosusest erinev märgendite osakaal on juhuslik või statistiliselt oluline. Binoomtestidel on seatud olulisusnivooks 0,01.

Siinkohal on oluline mainida, et tegemist on binoomtesti naiivse kasutamisega, sest on täitmata eeldus, et tegemist on sõltumatute katsetega. Kui mündi viskamisel on tegemist sõltumatute katsetega, sest üks vise ei sõltu eelmisest, siis regressiooniülesandes see ei kehti. Kui ühe sisendväärtuse puhul on tegelik väärtus regressioonimudeli poolt ennustatud usaldusvahemikust väljas, siis see suurendab tõenäosust, et ka teise sisendväärtuse puhul on tegelik väärtus usaldusvahemikust väljas.

Binoomtesti on kasutatud selleks, et anda statistilise olulisuse hinnang saadud tulemustele. See, et on täitmata eeldus, et tegemist on sõltumatute mõõtmistega, ei muuda antud eksperimendi tulemusi vähem relevantseks, sest nende sõltumatus ei olegi eesmärk. Eesmärgiks on vaadelda tulemuste erinevust väärtusest 0,95. Binoomtesti on kasutatud kui vahendit antud tulemuste kirjeldamiseks.

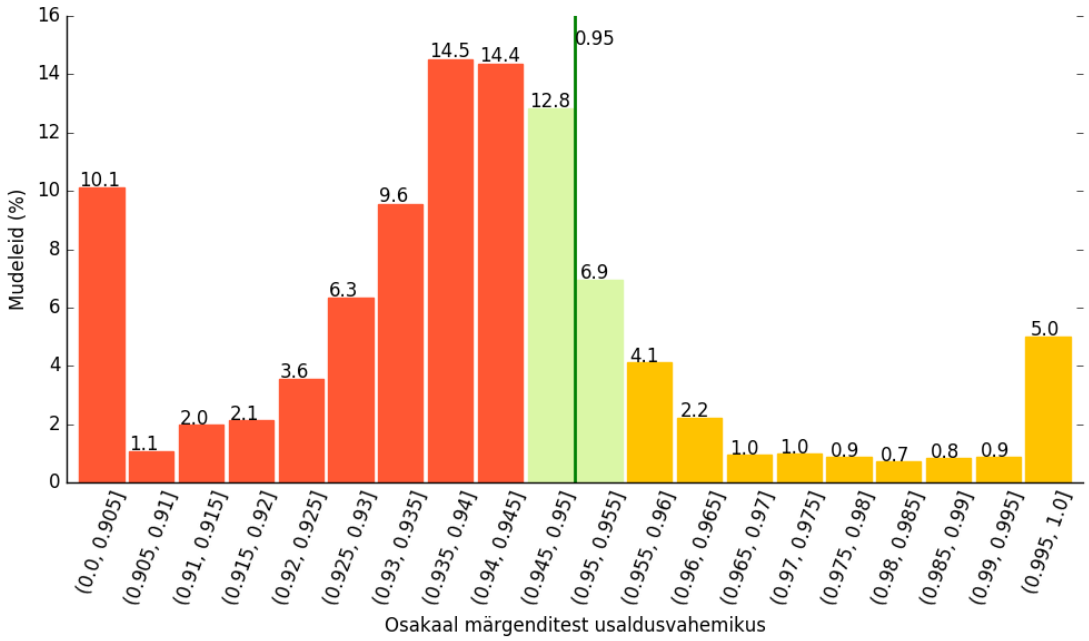
3.3 Tulemused

Antud peatükk annab ülevaate mõõtmiste tulemustest erinevatel stsenaariumitel ja analüüsib neid.

Treeniti 6651 mudelit. Neist 2166 rattaandmestikul, 2330 laevaandmestikul ning 2155 veiniandmestikul. Nende mudelite puhul usaldusvahemikku kuulunud päris märgendite osakaalude jaotus on toodud joonisel 4. Jooniselt on näha, kuidas suur osa tulemustest on väiksemad kui 0,95, kuigi ideaalis oleks 0,95 lähiümbrus kõige sagedamini esinev. Kui iga mudeli puhul oli usaldusvahemikku kuuluvuse tõenäosus 0,95, siis võiks eeldada, et kõige kõrgemad tulbad oleksid need, mis on joonisel rohelist värvi ehk väärtuse 0,95 ümbruses. Antud jooniselt on märkimisväärne tähele panna, et kuvatud osakaalude jagunemine läheneb normaaljaotusele, mille keskväärtus on madalam kui 0,95. Ligikaudu 76,5% kõikidest tulemustest on väiksemad kui 0,95 ning aritmeetiline keskmine kõikide treenitud mudelite päris märgendite osakaaludest usaldusvahemikus oli 0,9336.

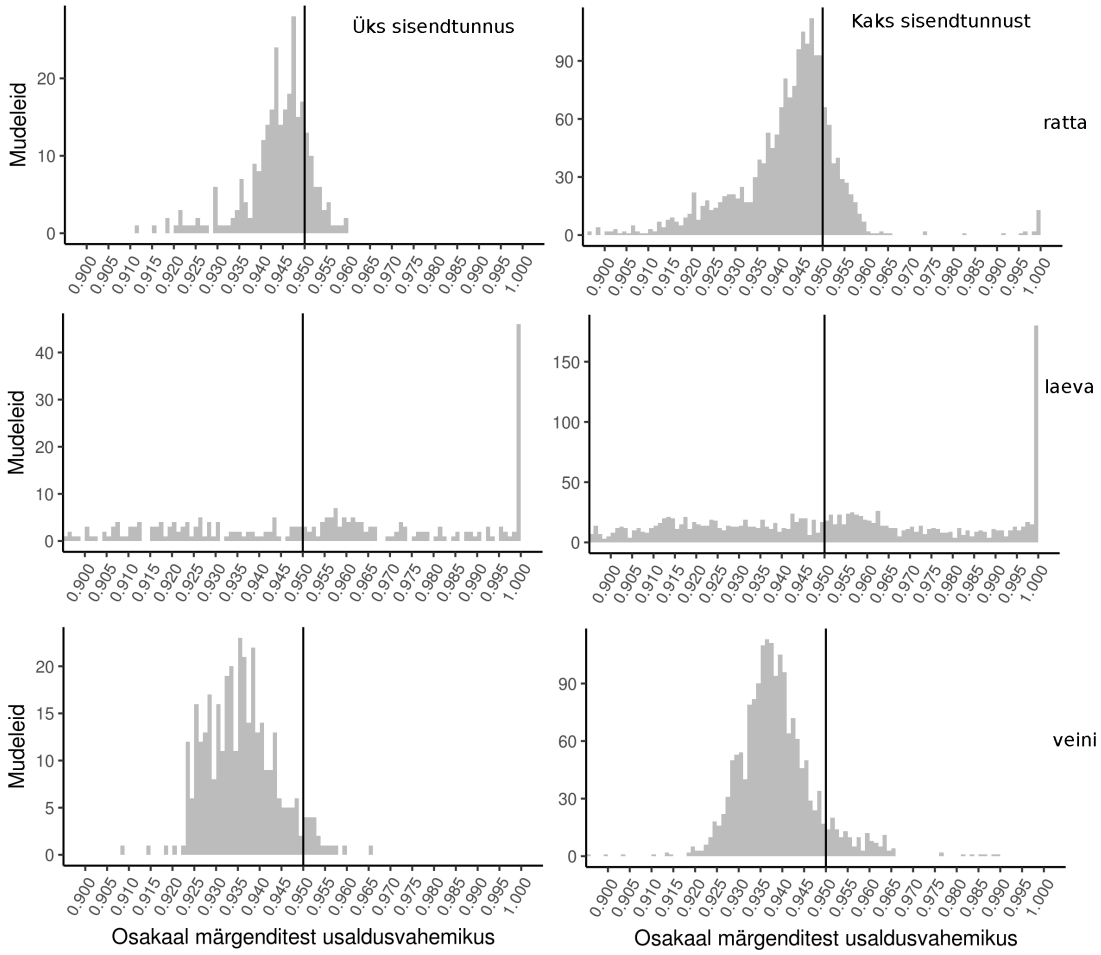
3.3.1 Erinevate tunnuste arvuga treenimine

Antud alapeatükk kirjeldab tulemusi erinevate treenimiseks kasutatud sisendtunnuste arvu kaupa.



Joonis 4. Kõikide treenitud mudelite usaldusvahemikus asuvate märgendite osakaalude jaotus. Roheline vertikaalne joon tähistab 0,95 tõenäosust. Kõige vasakpoolsem tulp kirjeldab kogusummat märgendite osakaaludest, mis on väiksemad kui 0,905 või sellega võrdsed, kusjuures need väärtused on üsna ühtlaselt jagunenud vahemikku (0;0,905].

Ühe tunnuse põhjal ennustamine Kõikidest mudelitest 931 puhul kasutati treenimiseks ühte sisendtunnust. Nendest 351 mudelit treeniti veiniandmestikul, 300 laevaandmestikul ja 280 rattaandmestikul. Ühe tunnuse põhjal treenimise tulemustest on näha, et nii ratta- kui ka veiniandmestikul treenides on 98% mudelitest tulemused vahemikus 0,92 kuni 0,96. Joonise 5 vasakpoolne veerg illustreerib ühe tunnuse põhjal treenimise tulemusi.



Joonis 5. Ühe ja kahe tunnuse põhjal ennustamise tulemused andmestiku kaupa. Vasakul on ühe tunnuse mudelid ja paremal kahe tunnuse mudelid. Ülemises reas on kujutatud rattaandmestiku tulemusi, keskmises laevaandmestiku tulemusi ning alumises veiniandmestiku tulemusi. Joonisel on kujutatud vaid tulemused vahemikus 0,9 kuni 1.

Erinevalt ratta- ja veiniandmestikust on laevaandmestiku tulemused jaotunud üpris ühtlaselt üle kõikide väärtuste. Huvitav tendents laevaandmestiku puhul on ka suur hulk selliseid mudeleid, mille puhul kõik päris märgendid kuuluvad usaldusvahemikku. Selliseid mudeleid ei esinenud teiste andmestike peal ühe tunnusega treenides.

Kahe tunnuse põhjal ennustamine Kahte tunnust kasutati kõige suurema arvu mudelite treenimiseks - selliseid mudeleid loodi 5635. Kahe tunnuse põhjal treenitud mudelite tulemused on sarnased ühe tunnuse kasutamisel treenitud mudelitele. Kahe tunnuse põhjal ennustamise tulemused on toodud joonisel 5 paremas veerus ning sellelt joonilselt on näha, kuidas erinevate andmestike peal treenitud mudelite üldised trendid on sarnased vasakpoolse veeruga, mis kirjeldab ühel tunnusel treenitud mudelite tulemusi. Sarnaselt ühe tunnuse mudelitele, on ka seekord suuremal osal ratta- ja veiniandmestikul treenitud mudelitel usaldusvahemikku kuuluvate päris märgendite osakaal vahemikus 0,92 kuni 0,96 — mõlemal andmestikul enam kui 90% tulemustest on selles vahemikus. Samuti sarnaselt ühe tunnuse mudelitele, on laevaandmestikul treenitud mudelite hulgas proportsionaalselt suurim hulk selliseid, mille puhul kõik päris märgendid kuulusid usaldusvahemikku. Teisisõnu, suur hulk laevaandmestikul treenitud mudelitest on olnud liiga vähe enesekindlad ning see on joonisel nähtav kõrge tulbana horisontaaltelje väärtuse 1 juures. Samuti on sarnaselt ühe tunnuse mudelitele laevaandmestikul treenitud mudelite õigete märgendite usaldusvahemikku kuulumise osakaalud jaotunud võrreldes teiste andmestikega ühtlasemalt.

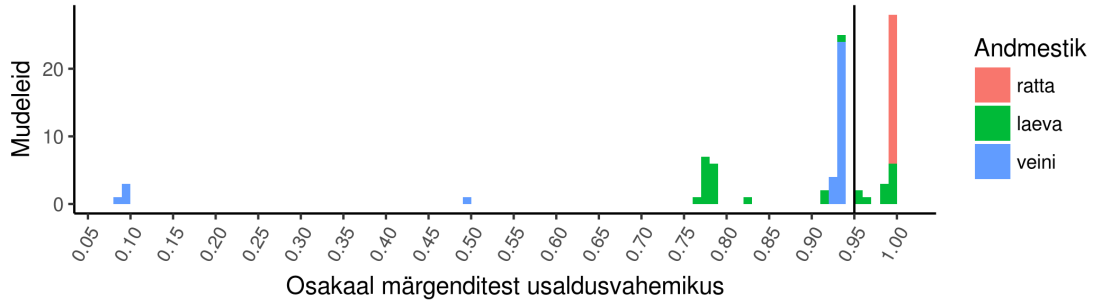
Kõikide tunnuste põhjal ennustamine Kõikide tunnuste põhjal ennustamiseks treeniti 85 mudelit, nendest rattaandmestikul 22, laevaandmestikul 30 ja veiniandmestikul 33. Võrreldes neid arve ühe ja kahe sisendtunnusega treenitud mudelite kogusega, on kõiki sisendtunnuseid kasutatud märgatavalt vähemate mudelite treenimisel.

Antud töös on erinevad uurimisalused grupid tasakaalustamata. Kõikide tunnuste põhjal on treenitud 85 mudelit, kuigi kahe tunnuse põhjal on mudeleid treenitud 5635. Samas on kõikide tunnuste kasutamisel võimalik ühel väljundtunnusel ainult üks sisendtunnuste kombinatsioon ning kahe tunnuse peal ennustades on see kogus kordades suurem. Seega on loomulik, et kahe tunnuse peal treenitud mudeleid on rohkem. Samuti on loomulik, et ühte tunnust sisendina kasutades on vähem kombinatsioone kui kahte tunnust sisendina kasutades, ent rohkem kombinatsioone kui kõiki tunnuseid korraga sisendina kasutades.

Sellegipoolest võib väike mudelite arv olla põhjuseks, miks joonisel 6 kuvatud tulemuste jaotus tundub olevat korrapäratu ega näi kuskile koonduvat.

Paistab, et kõik rattaandmestikul treenitud mudelid on liiga vähe enesekindlad ja et nende puhul kõik päris märgendid on kuulunud usaldusvahemikku. Olukordasid, kus kõik päris märgendid kuuluvad usaldusvahemikku, lahkab lähemalt järgmine peatükk.

Veiniandmestiku tulemus on suuremas jaos vahemikus 0,92 kuni 0,94, mõningate



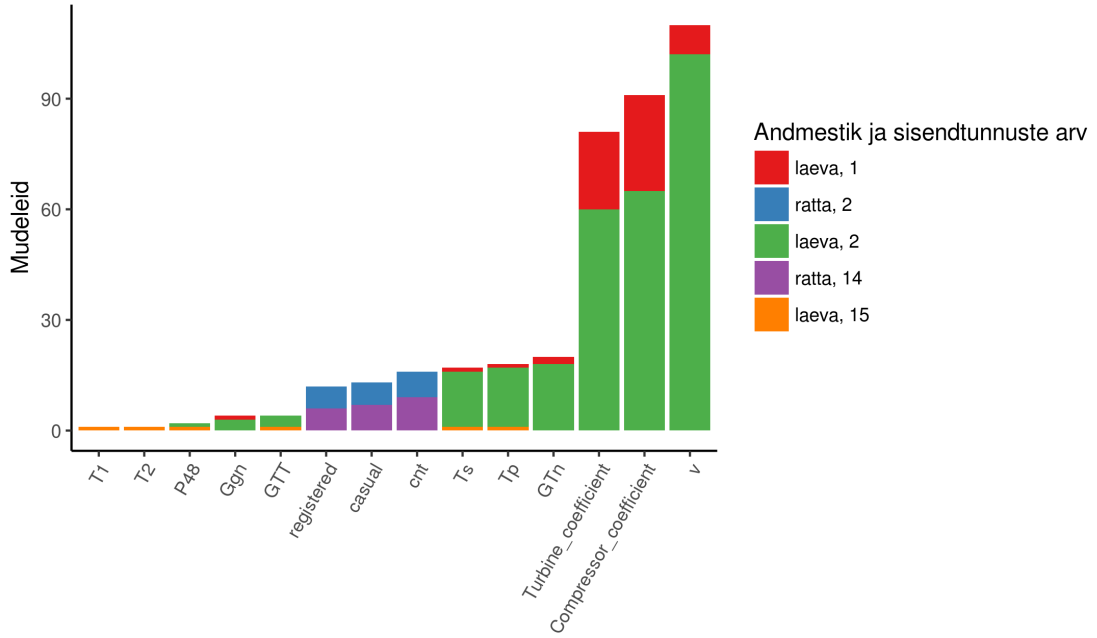
Joonis 6. Kõikide tunnuste peal treenitud mudelid

väga kehvade tulemustega alla 0,5 osakaalu, mille üle on pikemalt arutletud peatükis 3.3.4. Laevaandmestiku tulemused on jaotunud kõige korrapäratumalt.

100% usaldusvahemikku sattunud märgenditega regressiooniülesannete lahkamine Torkab silma huvitav asjaolu, et selliseid mudeleid, kus 100% testandmete märgenditest kuulub usaldusvahemikku, esineb ohtrasti laevaandmestiku puhul ning mõnel juhul ka rattaandmestiku puhul. Järgnevalt on üritatud välja selgitada, millistel tingimustel selline nähtus esineb.

Joonis 7 kujutab selliste mudelite jaotust, mis ennustasid päris märgendi usaldusvahemikku enam kui 99% juhtudest. On näha, et seda esines kõige enam laevaandmestikul kiiruse ning turbiini- ja kompressori kuluvuskoefitsiendi (`v`, `Turbine_coefficient`, `Compressor_coefficient`) ennustamisel. Võimalik, et selle põhjuseks on asjaolu, et kõik teised tunnused laevaandmestikus on nende põhjal simuleeritud ehk saadud selliste funktsioonide tulemusel, mille puhul on sõltumatuteks muutujateks just nimelt need kolm väärtust (Coraddu jt, 2016). Andmestiku kirjelduses on mainitud ka seda, et tunnus `lp` on lineaarses seoses tunnusega `v`. Kuigi andmestiku loojad on kirjutanud, et tegemist on andmetega, mis on kooskõlas päris laeva jõuseadet kirjeldavate näitudega, on tegemist siiski sünteetiliste andmetega ning võib järeldada, et nendevahelised sõltuvused on piisavalt lihtsad, et need regressioonimudelil Gaussi protsesside abil täielikult selgeks õppida. Ning tundub, et kui õpitav seos on piisavalt lihtne, siis kuuluvad peaaegu kõik päris märgendid usaldusvahemikku.

Samuti on lihtne õppida mudeleid, kus sisendväärtused kirjeldavad väljundtunnust täiuslikult. Ka rattaandmestikul on selliseid mudeleid treenitud - nimelt on kolme ennustatava atribuudi suhe järgmine: `cnt = registered+casual`. Seega alati, kui ennustatakse



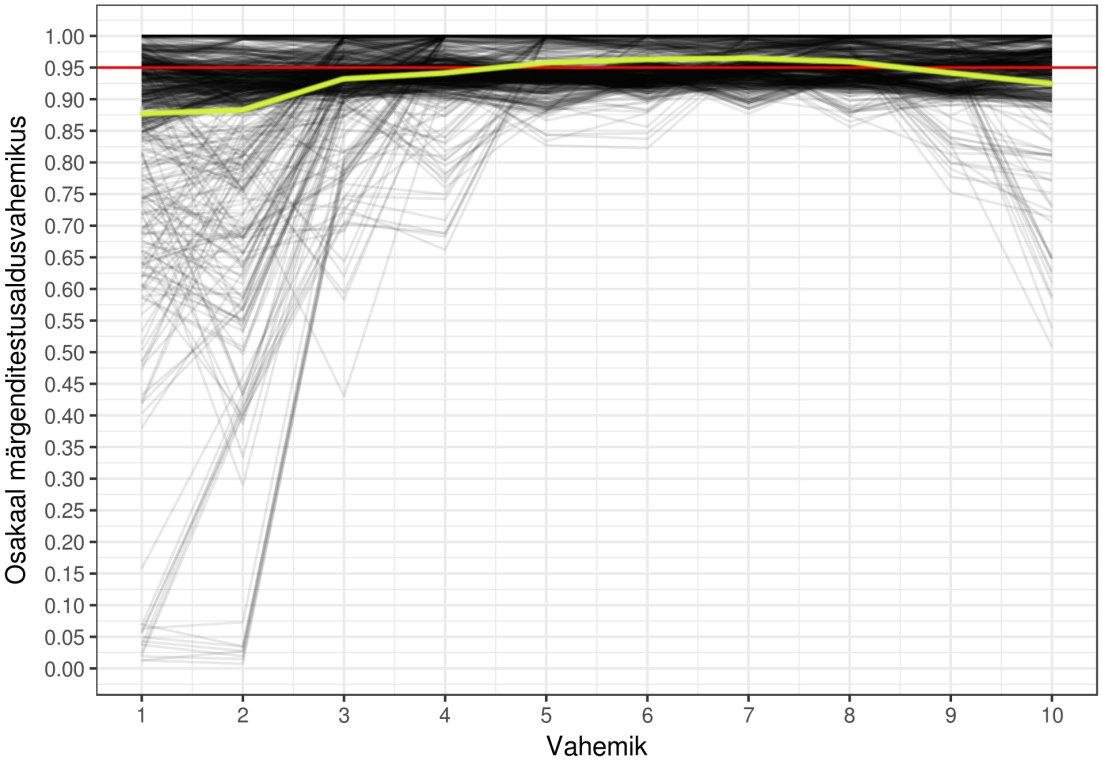
Joonis 7. Märgendid, mida ennustades enam kui 99% tulemustest kuulus usaldusvahemikku

neist ühte väärtust ja sisendis on olemas teised kaks, millest see otseselt tuletada, siis on mudeli õppimine lihtne. Selliste regressiooniülesannete lisamine katsetesse ei olnud tegelikult täiesti tahtlik, vaid pigem tundus eksperimenti disainides pädev kasutada nimetatud kolmest ühte ennustatavaks ning kuni ühte sisendina, ent algoritm kirjutati niiviisi, et see kombineeris sisenditena ka teisi ennustamiseks mõeldud väärtusi. Sellegipoolest on tegemist huvitavate tulemustega, sest saab nentida, et usaldusvahemikku sattumise tõenäosusest suuremat osakaalu päris märgenditest, mis sinna kuulusid, põhjustab asjaolu, kui mõni sisendtunnustest või kombinatsioon neist kirjeldab väljundtunnust täiuslikult.

3.3.2 Tulemused vahemike kaupa

Mudeleid, mis sobivad vahemike kaupa testimiseks, treeniti 700. Need on mudelid, millel on üks sisendtunnus ning see sisendtunnus on reaalarvuliste väärtustega. Joonisel 8 on kujutatud kõikide selliste mudelite kumulatiivsed tulemused.

Jooniselt on näha, et vahemikes 1, 2, 3 ja 10 on päris märgendite osakaal kõige suure-



Joonis 8. Vahemiku kaupa mõõtmiste tulemused. Testandmestik on jaotatud kümneks vahemikuks ning igas vahemikus on välja arvutatud päris märgendite osakaal, mis kuuluvad usaldusvahemikku. Iga joon tähistab ühe testimise tulemusi ning iga vahemiku tulemus on ühendatud omavahel. Mida tumedam on mingi ala, seda rohkem jooni on seda kohta läbinud. Kollane joon tähistab aritmeetilist keskmist.

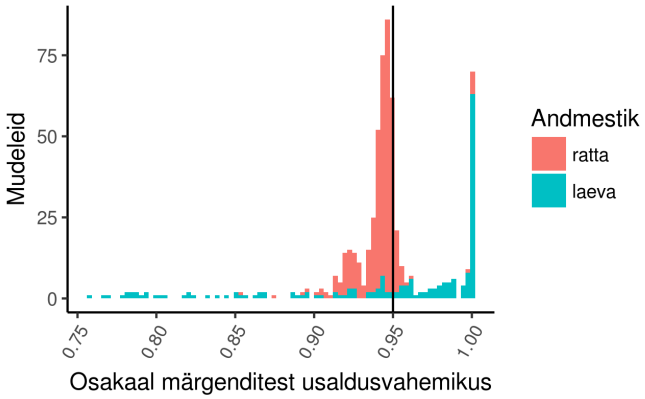
ma hälbega tõenäosusest 0,95. Tegemist on huvitava avastusega, sest enne katsetamist oli hüpotees pigem selline, et kõige väiksemate ja kõige suuremate sisendväärtuste vahemike puhul on mudel liiga vähe enesekindel. Seevastu näitavad katsed teistsuguseid tulemusi.

3.3.3 Erinevate treeningsuurustega treenimine

Treeniti kolme suurusega andmete hulgaga, 1000 treeningandmega 6059 mudelit, 2500 treeningandmega 509 mudelit ning 5000 treeningandmega 83 mudelit, kusjuures viimases stsenaariumis on kõik mudelid rattaandmestikul.

Vähemalt 2500 andmepunktiga treenitud mudelitel ei esinenud päris märgendite

osakaalu usaldusvahemikus, mis oleks madalam kui 0,75, kusjuures madalamaid tulemusi kui 0,85 esines vaid laevaandmestikul treenides. Selliste treeningsuurustega mudelite tulemused on kujutatud joonisel 9.



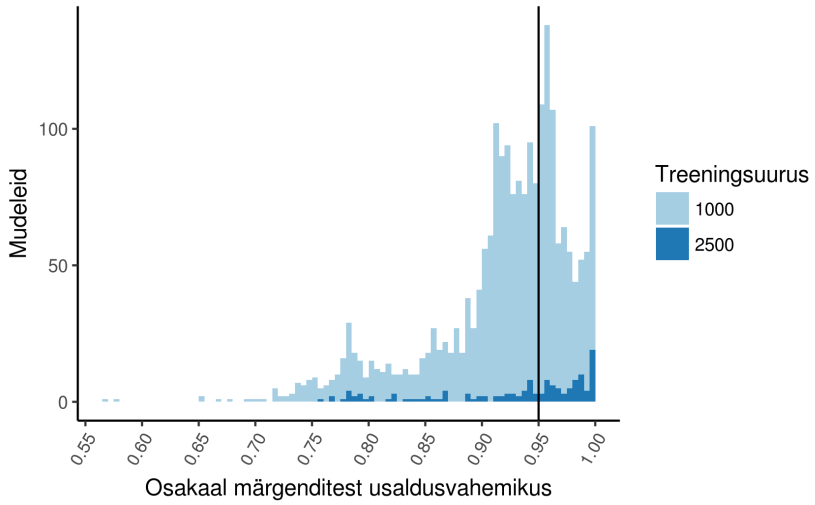
Joonis 9. Vähemalt 2500 treeningandmega treenitud mudelite tulemused

Muidugi tasub siinkohal ka arvestada, et vähemalt 2500 treeningandmega mudeleid treeniti enam kui kümme korda vähem kui 1000 treeningandmega mudeleid ning see kindlasti mõjutab asjaolu, et kõige madalam tulemus esines väiksema treeningkogusega mudelite puhul. Erinevalt rattaandmestikul lahendatud regressiooniülesannetest, on laevaandmestiku ülesannete testimiste tulemuste puhul jooniselt 9 märgata, kuidas osakaalud ei koondu ühegi punkti lähedale, vaid jaotuvad vahemikku 0,75 kuni 1. Väärtus 1,0 on siinkohal erand, mille tagamaid lahati peatükis 3.3.1.

Laevaandmestikul treenimise tulemused Analüüsides eelnevaid stsenaariume, on ilmnenud, et laevaandmestikul treenitud mudelitel on kõige suurem varieeruvus tulemustes ning erinevus väärtusest 0,95. Jooniselt 10 on näha, kuidas laevaandmestikul treenitud mudelite tulemused siiski koonduvad 0,95 ümbrusesse, aga suurema hälbega kui teistel andmestikel läbi viidud mõõtmised. Alla 0,95 jäävad tulemused moodustavad 63,36% kogu tulemustest.

3.3.4 Binoomtesti tulemused

Antud peatükis uuritakse binoomtestiga tulemuste olulisust ning vastavalt selle testi tulemustele viiakse läbi juhtumianalüüs kõige märkimisväärsematele tulemustele.

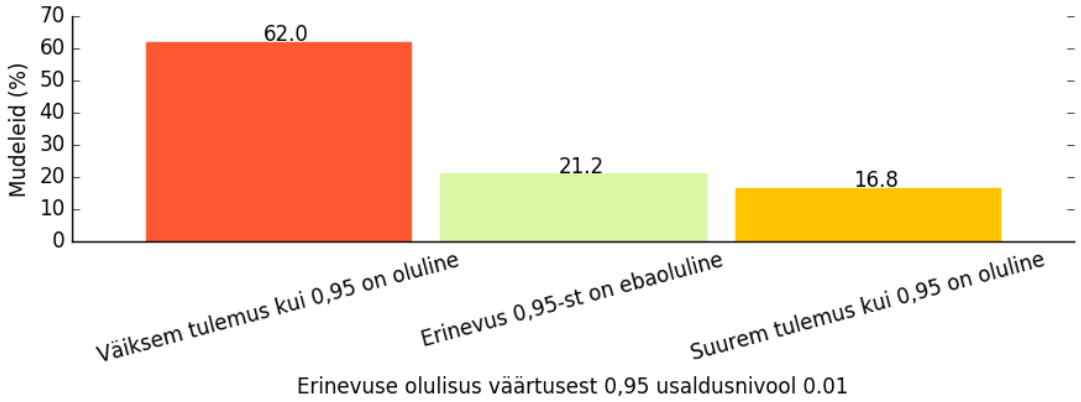


Joonis 10. Laevaandmestikul treenitud mudelite tulemuste varieeruvus

Valitud on tulemused, mis vastavad olulisusnivoole 0,01. Hinnangut kõikide mudelite usaldusvahemikku kuulunud märgendite osakaalu erinevusele tõenäosusest 0,95 kuvab joonis 11.

Jooniselt 11 on näha, et ligikaudu 62% mudelitest hindasid oma usaldusvahemikku liiga enesekindlalt. Nendel mudelitel kuulus usaldusvahemikku vähem kui 95% päris märgenditest ning seda sellisel määral, et olulisusnivool 0,01 on tulemus statistiliselt oluline, eeldades ennustuste sõltumatust. 21,2% tulemustest olid 0,95-st ebaoluliselt erinevad. Sellised on siis tulemused, mis on veidi väiksemad või suuremad kui 0,95, ent mille erinevused on piisavalt väikesed, et mitte olla olulisusnivool 0,01 olulised. Seda võib mõista niiviisi, et 21,2% tulemustest olid parajalt enesekindla mudeli omad ehk piisavalt lähedal tõenäosusele 0,95. 21,2% on kõikidest treenitud mudelitest pigem väike kogus, eriti kui siinkohal ei ole oluline, et tulemus oleks täpselt võrdne 0,95-ga, vaid kuuluks selle lähiümbrusesse. Veel vähem mudeleid on liiga vähe enesekindlad ehk oluliselt suuremate tulemustega kui 0,95. Valdav enamus on siiski liigselt enesekindlad mudelid.

Kuna korraga testitakse suurt hulka erinevaid hüpoteese, on võimalik, et esineb mitmese testimise probleem ehk mõni statistiliselt oluline seos on leitud ka siis, kui seda tegelikult ei eksisteeri. Et mitmese testimise vältimise eesmärk on vähendada I liiki vigade ehk valepositiivsete seoste leidmist, siis pole mitmese testimise korrektsiooni



Joonis 11. Kõikide treenitud mudelite tulemuste erinevuste olulisuse jagunemine

siinkohal rakendatud. Antud juhul on eesmärgiks saada üldpilt tulemustest ning mitte välistada valepositiivseid tulemusi.

Statistilises analüüsis on olulisel kohal lisaks erinevuse olulisusele ka erinevuse efekti suurus. Kümme suurima efektiga tulemust on toodud tabelis 1. Oluliselt erinevad tulemused on järjestatud efekti järgi, et uurida suurimaid erinevusi, mis on statistiliselt olulised. Efektiks on siin loetud 0,95-st erinevuse absoluutväärtust. Peamiseks eesmärgiks on leida liigselt enesekindlaid mudeleid. Siinkohal on keskendutud just nendele tulemustele, mis on olulised nivool 0,01.

Kõige suuremad erinevused väärtusest 0,95 ilmnesid veiniandmestikul testides ning kõik need erinevused oli negatiivsed ehk osakaalud oli 0,95-st väiksemad. Tulemused on huvitavad. On võimalik, et tabelis 1 märgitud sisendtunnuste põhjal treenides on keeruline veini kvaliteeti hindama õppida. Kõige suurema erinevusega tulemuste puhul on treenimiseks kasutatud kõiki andmestikus olevaid tunnuseid ning võrdlus teiste sellis-te regressiooniülesannetega (kokku 33) näitab, et veiniandmestikul on kõiki tunnuseid kasutades alati olnud tulemus alla 0,95 ning seda olulisusnivool 0,01. Võimalik, et suurte sisenddimensioonide kasutamine põhjustab Gaussi protsesside mudelis liigset enesekindlust. Erinevalt ratta- ja laevaandmestikust, puudusid veiniandmestikus atribuutide hulgas sellised tunnused, mis täiuslikult kirjeldaksid väljundtunnust, sest sisendandmed on füüsiliselt mõõdetud ning väljundandmed on hinnatud inimese maitsemeele järgi, omamata üksikasjalikke teadmisi veini keemiliste omaduste kohta (Cortez jt, 2009).

Seevastu, kui on kasutatud sisendtunnustena paare `total_sulfur_dioxide, density`

Tabel 1. Suurima efektiga erinevused

märgend	sisendtunnused	treeningkogus	osakaal usaldusvahemikus
quality	kõik andmestikus	1000	0,08902001
quality	kõik andmestikus	1000	0,09030272
quality	kõik andmestikus	1000	0,09261160
quality	kõik andmestikus	1000	0,09748589
quality	kõik andmestikus	1000	0,49050795
quality	total_sulfur_dioxide, density	1000	0,49743458
quality	free_sulfur_dioxide, density	1000	0,50564392
quality	total_sulfur_dioxide, density	1000	0,50590046
quality	free_sulfur_dioxide, density	1000	0,51513597
quality	free_sulfur_dioxide, density	1000	0,51924064

või `free_sulfur_dioxide, density`, on mõnikord treenitud ka mudeleid, mis ennustavad usaldusvahemikku rohkem kui 95% testmärgenditest (kokku seitsmel juhul 63-st, kus on sarnaseid tingimusi kasutatud).

Tulemuste varieeruvust suurima efektiga erinevustega mudelite seas võib põhjustada ka stohhastilisus eksperimendis. Juhuslikkust kasutatakse kahes kohas - treeningandmete valimisel ning hüperparameetrite optimeerimisel. On võimalik, et mõningad tulemused on nii märkimisväärselt oodatud 0,95-st erinevad just seetõttu, et juhuslikkusest ei piisanud hüperparameetrite optimeerimisel lokaalsest miinimumist pääsemiseks. Sellegipoolest, kui proovida sama regressiooniülesannet korduvalt, siis võiks loota, et vähemalt ühe korra saavad hüperparameetrid piisavalt optimeeritud, et mudel ei oleks liigselt enesekindel. Seepärast on see tulemus eriline, et veiniandmestiku puhul kõikidelt tunnustelt ennustades ei saadud kordagi usaldusvahemikku vähemalt 95% tulemustest.

Kümne suurima erinevusega tulemuste edetabelis ei ole ühtegi ühe sisendtunnusega treenitud mudelit ja seetõttu on otsustatud selliseid mudeleid eraldi uurida ning vastavad tulemused on toodud tabelis 2.

Tabelis 2 on näha, et kõik väärtusest 0,95 kõige rohkem erinevad tulemused on treenitud laevaandmestikul. Jääb silma, et kümnest viiel on sisendtunnuseks `GTT` ja neljal `GTn`. Kui vaadata teisi mudeleid, kus `GTn` on olnud sisendiks, siis 18 puhul 23-st on tulemus madalam kui 0,95. Kusjuures, ennustatud on erinevaid tunnuseid ning neil kõigil (välja arvatud tunnust `GTT` ennustades) on olnud tunnust `GTn` sisendina kasutades

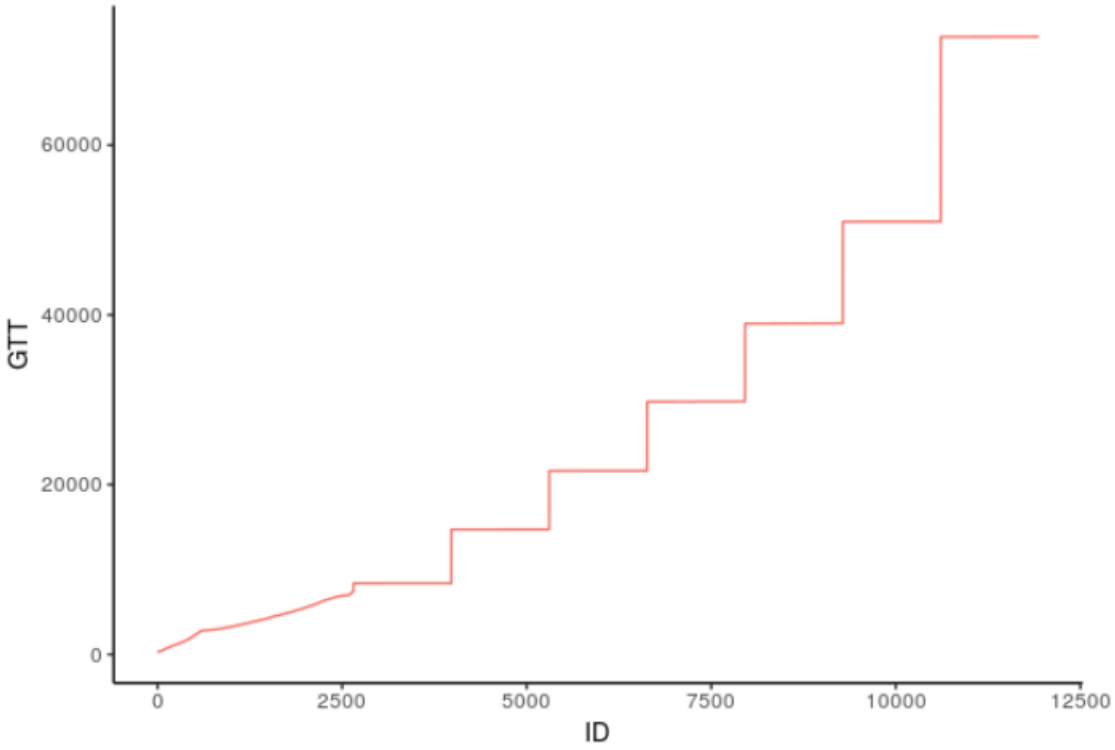
Tabel 2. Ühe sisendtunnusega mudelite suurima efektiga erinevused

märgend	sisendtunnus	treeningsuurus	osakaal usaldusvahemikus
Ggn	GTT	1000	0,7194074
P48	GTn	1000	0,7204134
v	GTT	2500	0,7779309
v	GTn	1000	0,7783977
GTT	Ts	1000	0,7802268
Tp	GTn	1000	0,7817816
Tp	GTT	1000	0,7826962
GTT	GTn	1000	0,7838851
v	GTT	1000	0,7848912
GTn	GTT	1000	0,7867203

nii madalamaid kui ka kõrgemaid tulemusi kui 0,95. Samas kui sisendtunnusena on kasutatud tunnust GTT, siis on alla 0,95 saadud tulemusi 21 23-st. Eelnevast jääb mulje, et GTT on kehv sisendtunnus, mida kasutada, ning tasub lähemalt uurida, miks.

Joonisel 12 on kujutatud tunnuse GTT väärtused kasvavas järjestuses. Esialgu andmed kasvavad pidevalt ligikaudu väärtuseni 8000 ning seejärel need kasvavad diskreetselt. Võimalik, et selliste astmeliselt kasvavate sisendväärtuse korral on mudel oma ennustustes liiga enesekindel. Seda toetab ka asjaolu, et ka tunnus GTn kasvab esialgu pidevalt ning seejärel diskreetselt astme kaupa sarnaselt GTT tunnusele ning suurem osa mudeleid, kus on seda kasutatud sisendina, on ennustanud usaldusvahemikku vähem kui 95% päris märgenditest.

Samas on lisaks GTn ja GTT tunnustele veel tunnuseid laevaandmestikus, mille väärtused kasvavad esialgu pidevalt ning seejärel diskreetselt, näiteks tunnused Tp ja Ts. Küll aga neid tunnuseid sisendina kasutades ei esinenud tulemustes mustreid, mis kinnitaksid, et just sellise sisendväärtuse jaotuse korral on mudel tihedamini liigselt enesekindel. Seega on põhjust uurida, kas tunnused GTT ja GTn on veel millegi pärast erilised. Selgub, et GTT teeb kõikidest teistest tunnustest eriliseks see, et selle väärtused varieeruvad suurimas vahemikus: 253,5 kuni 72784,8. Ning GTn on sarnase jaotusega tunnuste hulgas varieeruvuse suuruselt teine, varieerudes vahemikus 1307,7 kuni 3560,7. Seega võib olla, et kasutades Gaussi protsesside mudelil sisendtunnustena mittepidevaid tunnuseid, mis varieeruvad suurtes vahemikes, siis on võimalik, et mudel on liigselt enesekindel.



Joonis 12. Tunnuse GTT väärtused sorteeritud kasvavas järjekorras. ID kirjeldab indeksit.

3.4 Tulemuste kokkuvõte ja järeldused

Võib väita, et suurel osal juhtudest on Gaussi protsessid liigselt enesekindlad. Asjaolu, et 62% tulemustest on olulisusnivool 0,01 oluliselt madalamad kui usaldusvahemikku kuulumise tõenäosus, on selle üks suurimaid argumente. Seda väidet toetavad ka tulemuste jaotuse visualisatsioonid, millelt võib hinnata, et kõige enam esinev tulemus on alla 0,95 ning et ülejäänud tulemused ümbritsevad seda kui normaaljaotuse keskväärtust. Keskmine päris märgendite osakaal on samuti alla 0,95 ning selle väärtus on ligikaudu 0,93. Kusjuures, keskmine väärtus oleks veelgi madalam, kui treenitud mudelite hulgas ei oleks selliseid, kus sisendtunnus kirjeldab väljundit täiuslikult.

Küll aga on keeruline põhjendada, miks sellised tulemused esinevad ja millistel tingimustel on Gaussi protsesside kasutamisel suurem võimalus, et tulemused jäävad usaldusvahemikust tihedamini välja, kui seda kirjeldav tõenäosus lubab. Sellegipoolest on võimalik täheldada juhtumit, kus veiniandmestikul kõikide tunnustega treeniti kokku

33 mudelit erinevate treeningandmete kombinatsioonidega ning kõikide nende mudelite tulemused olid väiksemad kui 0,95.

Üks huvitav muster on see, et kui sisendtunnuseks on üks astmeliselt kasvavate väärtustega tunnus, mille väärtuste suurused varieeruvad suures mahus, siis esineb tihti liigselt enesekindlaid mudeleid.

Gaussi protsessidel regressioonimudeli puhul on võimalik, et väikeste ning suurte sisendväärtuste puhul on mudel liigselt enesekindel, nagu seda näitab vahemike kaupa testimine. Ennustades väärtusi selliste sisendväärtuste põhjal, mille puhul võib olla tarvis ekstrapoleerida, on mudeli liigne enesekindlus tihedamini esinev ning suurema keskmise hälbega usaldusvahemikku sattumise tõenäosusest.

4 Kokkuvõte

Töö kirjeldus Antud töö annab põhjaliku ülevaate Gaussi protsesside olemusest ning nende kasutamisest masinõppes. Töös on selgitatud ka mudeli liigse ja liiga vähese enesekindluse tähendust ning on läbi viidud eksperiment selle uurimiseks. Eksperimendi raames treeniti 6651 masinõppemudelit kolmel andmestikul, varieerides treenimisel kasutatavaid tingimusi.

Tulemused Eksperimendi tulemusena on põhjust arvata, et mõningatel tingimustel võib esineda Gaussi protsesside põhjal treenitud mudelitel liigselt enesekindlaid tulemusi. Sellisteks olukordadeks on suure hulga (antud töös 11) sisendtunnuste kasutamine mudeli treenimisel ning sellistelt sisendväärtustelt ennustades, mis kuuluvad kogu sisendväärtuste vahemiku väiksemate ning suurimate hulka nagu ilmnes vahemike kaupa testides. Liigset enesekindlust esineb ka ühe tunnuse kasutamisel, mille väärtused kasvavad astmeliselt ning mille väärtuste suurused varieeruvad suures mahus.

Gaussi protsesside puhul esineb ka mudeleid, mis on liiga vähe enesekindlad, ning selle peamiseks põhjuseks on olukord, kus mõni sisendtunnus kirjeldab väljundtunnust täiuslikult või kui neid seostav funktsioon on triviaalne (näiteks lineaarfunktsioon). Sellisel juhul õpib mudel selle väga hästi selgeks ning suurem osa, kui mitte kõik, päris märgenditest satuvad usaldusvahemikku.

Üldiselt on kõikide tulemuste keskmine madalam, kui võiks oodata. Aritmeetiline keskmine tulemus on 0,9336, mis on madalam kui 0,95. Samuti on oluline märkida, et 73% mõõtetulemustest on madalamad, kui 0,95. Tõenäoliselt praktikas ei ole 93% ja 95% erinevus väga oluline. Kuid nagu töö tulemustest näha, esineb ka suuremaid erinevusi ning on oluline välja uurida, millest taolised suured erinevused on tingitud ning kuidas ära tunda, et on oht suurematekserinevusteks. Teoreetiliselt poolelt on ka erinevus 93% ja 95% vahel siiski oluline, sest võib-olla saab Gaussi protsesside algoritmi modifitseerida kuidagi nii, et sellist süstemaatilist erinevust vältida.

Tulemuste rakendatavus Üheks võimaluseks antud bakalaureusetöö tulemuste rakendamiseks on valdkondades, mis kasutavad masinõpet ning kus vastuvõetavatel otsustel on suur kahjukulu. See tähendab olukordades, kus väära ennustuse puhul on tõsised tagajärjed. Sellistel juhtudel on oluline teada, kui suure riskiga mingeid otsuseid vastu võetakse. Tänu käesolevale uurimusele on võimalik pöörata erilist tähelepanu Gaussi

protsesside abil paljudelt tunnustelt ennustades või siis väga väikeste või suurte väärtustega sisendtunnuseid kasutades, kuna võib juhtuda, et ekstrapoleerimine mõjutab mudeli enesekindlust.

Ent pigem on saadud tulemuste rakendatavus teoreetiline ning tõmbab tähelepanu äärejuhtumitele, millele keskendudes võib Gaussi protsesside enesekindlust parandada — näiteks Gaussi protsesside sellisel implementeerimisel, mis võtab arvesse liigset enesekindlust väikeste ja suurte väärtuste vahemikus.

Uurimise jätkamise võimalused Antud teema uurimist saab kindlasti jätkata. Üheks võimalikuks viisiks on lisada tingimusi ning uurida, kuidas mõjutab enesekindlust protsesside keskväärtusfunktsiooni ning kovariatsioonifunktsiooni varieerumine ning võimalik on uurida ka seda, kuidas mõjutab mudeli enesekindlust ligikaudsete meetodite (hõredate mudelite) kasutamine treenimise keerukuse vähendamiseks. Antud uurimuse tulemustest saab jätkata tööd teemal, kuidas ja miks mõjutab astmeline funktsioon mudeli enesekindlust, sest siin töös on avastatud ainult sellise sisendi kasutamise potentsiaalne mõju mudeli enesekindlusele.

Viidatud kirjandus

Bailey, K. (2016). From both sides now: the math of linear regression. Vaadatud 17.04.2018 allikast http://katbailey.github.io/post/from-both-sides-now-the-math-of-linear-regression/. Veebipäevik. AI, Machine Learning, Data Science, Language. (Viidatud lehel 8).

Bhinge, R., Biswas, N., Dornfeld, D., Park, J., Law, K. H., Helu, M. ja Rachuri, S. (2014), Teoses *2014 IEEE International Conference on Big Data (Big Data)* (l. 978–986). doi:10.1109/BigData.2014.7004331. (Viidatud lehel 5)

Cipollini, F., Oneto, L., Coraddu, A., Murphy, A. J. ja Anguita, D. (2016). Condition Based Maintenance of Naval Propulsion Systems. Vaadatud 04.05.2018 allikast https://sites.google.com/view/cbm/home. (Viidatud lehel 17).

Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D. ja Figari, M. (2016). Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, *230*(1), 136–153. doi:10.1177/1475090214540874. (Viidatud lehekülgedel 16, 25)

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. ja Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Elsevier*, *47*(4), 547–553. (Viidatud lehekülgedel 17, 30).

Dheeru, D. ja Karra Taniskidou, E. (2017). UCI Machine Learning Repository. Vaadatud 24. 03. 2018 allikast http://archive.ics.uci.edu/ml. (Viidatud lehel 16)

Duvenaud, D. (2014). *Automatic model construction with Gaussian processes* (doktoritöö, University of Cambridge). (Viidatud lehel 15).

Fanaee-T, H. ja Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 1–15. doi:10.1007/s13748-013-0040-3. (Viidatud lehel 16)

Gal, Y., van der Wilk, M. ja Rasmussen, C. E. (2014). Distributed Variational Inference in Sparse Gaussian Process Regression and Latent Variable Models. Teoses Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence ja K. Q. Weinberger (Toim.), *Advances in Neural Information Processing Systems 27* (l. 3257–3265). Curran Associates, Inc. (Viidatud lehel 15).

GPy. (2012). GPy: A Gaussian process framework in Python. Vaadatud 24. 03. 2018 allikast http://github.com/SheffieldML/GPy. (Viidatud lehel 14)

Hensman, J., Fusi, N. ja Lawrence, N. D. (2013). Gaussian Processes for Big Data. *CoRR*, *abs/1309.6835*, 1–7. arXiv: 1309.6835. (Viidatud lehel 15)

Martino, L., Laparra, V. ja Camps-Valls, G. (2017). Probabilistic cross-validation estimators for Gaussian process regression. Teoses *2017 25th European Signal Processing Conference (EUSIPCO)* (l. 823–827). doi:10.23919/EUSIPCO.2017.8081322. (Viidatud lehel 8)

Murray, I. (2016). Lecture notes in Machine Learning and Pattern Recognition. University of Edinburgh. (Viidatud lehekülgedel 14, 18).

Nguyen-Tuong, D. ja Peters, J. (2008). Local Gaussian process regression for real-time model-based robot control. Teoses *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems* (l. 380–385). doi:10.1109/IROS.2008.4650850. (Viidatud lehekülgedel 5, 6)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... ja Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. (Viidatud lehekülgedel 14, 16, 19).

Platt, J. C., Burges, C. J., Swenson, S., Weare, C. ja Zheng, A. (2002). Learning a Gaussian process prior for automatically generating music playlists. Teoses *Advances in neural information processing systems* (l. 1425–1432). (Viidatud lehel 5).

Raschka, S. (2016). Mlxtend. doi:10.5281/zenodo.49235. (Viidatud lehel 18)

Rasmussen, C. E. (1999). *Evaluation of Gaussian processes and other methods for non-linear regression*. Citeseer. (Viidatud lehel 5).

Rasmussen, C. E. ja Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press. (Viidatud lehekülgedel 7, 14).

Schreiter, J., Englert, P., Nguyen-Tuong, D. ja Toussaint, M. (2015). Sparse Gaussian process regression for compliant, real-time robot control. Teoses *Robotics and Automation (ICRA), 2015 IEEE International Conference on* (l. 2586–2591). IEEE. (Viidatud lehel 6).

Schreiter, J., Nguyen-Tuong, D. ja Toussaint, M. (2016). Efficient sparsification for Gaussian process regression. *Neurocomputing*, *192*, 29–37. (Viidatud lehel 6).

Traat, I. ja Lepik, N. (2013). E-kursuse "Bayesi statistika Markovi ahelatega" materjalid. Tartu Ülikool. (Viidatud lehekülgedel 8, 10, 19).

Lisad

I. Repositoorium

Töös kasutatud skriptid ning tulemusi sisaldavad .csv failid leiab repositooriumist, mis on avalikult saadaval Bitbucketis aadressil:

- https://bitbucket.org/lauraruusmann/confidence_of_gaussian_processes/src/master/.

Repositoorium sisaldab ka eestikeelselt kommenteeritud koodifaile ning kirjeldust .csv failide sisu kohta.

II. Andmestikud

Tabel 3. Laevaandmestiku atribuutide tähendused ja mõõtühikud

Atribuudi nimi	Eestikeelne selgitus	Mõõtühik
lp	kangi asend	
v	laeva kiirus	sõlm
GTT	gaasiturbiini võlli pöördemoment	kN·m
GTn	gaasiturbiini pöörete arv	p/min
Ggn	gaasigeneraatori pöörete arv	p/min
Ts	tüürpoordi propelleri pöördemoment	kN
Tp	pakpoordi propelleri pöördemoment	kN
T48	kõrgrõhu turbiini väljumistemperatuur	°C
T1	gaasiturbiini kompressori sisselaskeõhu temperatuur	°C
T2	gaasiturbiini kompressori väljalaskeõhu temperatuur	°C
P48	kõrgrõhu turbiini väljundrõhk	baar
P1	gaasiturbiini kompressori õhu sisselaskerõhk	baar
P2	gaasiturbiini kompressori õhu väljalaskerõhk	baar
Pexh	gaasiturbiini heitgaasi rõhk	baar
TIC	turbiini sissepritse kontroll	%
mf	kütusevoog	kg/s

Tabel 4. Inglisekeelsed laevaandmestiku atribuutide kirjeldused

Atribuudi lühend	Inglisekeelne selgitus
lp	Lever position
v	Ship speed
GTT	Gas Turbine shaft torque
GTn	Gas Turbine rate of revolutions
Ggn	Gas Generator rate of revolutions
Ts	Starboard Propeller Torque
Tp	Port Propeller Torque
T48	HP Turbine exit temperature
T1	GT Compressor inlet air temperature
T2	GT Compressor outlet air temperature
P48	HP Turbine exit pressure
P1	GT Compressor inlet air pressure
P2	GT Compressor outlet air pressure
Pexh	Gas Turbine exhaust gas pressure
TIC	Turbine Injecton Control
mf	Fuel flow

Tabel 5. Jalgrattalaenutuste andmestik

Tunnuse nimetus	Eestikeelne selgitus	Võimalikud väärtused
dteday	kuupäev	päev kujul aaaa-kk-pp
season	aastaaeg	1 - kevad, 2 - suvi, 3 - sügis, 4 - talv
yr	aasta	0 - 2011, 1 - 2012
mnth	kuu	1 kuni 12
hr	tund	0 kuni 23
holiday	töövaba päev ehk püha	0 - ei ole püha, 1 - on püha
weekday	nädalapäev	0 kuni 6
workingday	tööpäev	1 - ei ole tööpäev ega nädalavahetus, 0 - muul juhul
weathersit	ilmastikuolud	1 - selge; mõned pilved; osaliselt pilves, 2 - udu ja pilves; udu ja poolpilves; udu ja mõned pilved; udu, 3 - kerge lumesadu; kerge vihm, äikesetorm ja hajuspilvisus; kerge vihm ja hajuspilvisus, 4 - tugev vihm, jäätkruubid, äikesetorm ja udu; lumi ja udu
temp	normaliseeritud temperatuur, °C vahemikus -8 kuni +39	0.0 kuni 1.0
atemp	normaliseeritud tajutav temperatuur, °C vahemikus -16 kuni +50	0.0 kuni 1.0
hum*	normaliseeritud õhuniiskus, väärtused on jagatud 100-ga (maksimaalne võimalik)	0.0 kuni 1.0
windspeed**	normaliseeritud tuule kiirus väärtused on jagatud 67-ga (maksimaalne võimalik)	0.0 kuni 1.0
casual	juhuslike laenutajate arv	
registered	registreeritud laenutajate arv	
cnt	laenutuste koguarv	

* - esialgne mõõtühik allikas kirjeldamata

Tabel 6. Veiniandmestiku tunnused

Tunnuse nimetus	Eestikeelne selgitus
fixed_acidity	fikseeritud happesus
volatile_acidity	volatiilne happesus
citric_acid	tsitrushappesus
residual_sugar	jääksuhkur
free_sulfur_dioxide	vaba vääveldioksiid
total_sulfur_dioxide	kogu vääveldioksiid
density	tihedus
pH	happesus
sulphates	sulfaadid
alcohol	alkohol
quality	kvaliteet

III. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Laura Ruusmann**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   Gaussi protsesside usaldusvahemik,

   mille juhendaja on Meelis Kull,

   1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

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

Tartus, 14.05.2018