Helsingin yliopisto - Helsingfors universitet - University of Helsinki ID 2007-277

Tiedekunta-Fakultet-Faculty
Valtiotieteellinen tiedekunta

Laitos-Institution-Department
Matematiikan ja tilastotieteen laitos

Tekijä-Författare-Author
Blomstedt, Paul

Työn nimi-Arbetets titel-Title
Bayesläiset menetelmät diskriminatiivisessa ja generatiivisessa luokittelussa

Oppiaine-Läroämne-Subject
Tilastotiede

Työn laji-Arbetets art-Level
Pro gradu

Aika-Datum-Month and year
2007-05-14

Sivumäärä-Sidantal-Number of pages
47 s.

Tiivistelmä-Referat-Abstract
Tilastollisessa luokittelussa kiinnostuksen kohteena oleva havaintoyksikkö sijoitetaan tätä kuvaavien havaittujen ominaisuuksien perusteella johonkin luokkaan. Esim. sähköpostiohjelmien roskapostisuodattimet hyödyntävät luokittelumenetelmiä luokitellessaan viestit näiden sisällön perusteella joko roskapostiksi tai ”oikeaksi” sähköpostiviestiksi. Tässä työssä taas tarkastellaan lääketieteellistä sovellusta, jossa potilaan terveydentilaa koskevien tietojen perusteella pyritään päättelemään onko potilaalla jokin määrätty sairaus vai ei. Luokitelussa käytettävä luokittelumalli estimoidaan luokiteltavan havaintoyksikön kanssa samasta perusjoukosta olevasta, valmiiksi luokitellusta aineistosta, jota kutsutaan opetusaineistoksi.

Luokittelumalleja voidaan muodostaa monin eri tavoin. Tässä työssä käsiteltävät mallit perustuvat havaintoyksikön ominaisuuksille ehdollistetun, luokkamuuttujan ehdollisen jakauman mallintamiseen. Luokittelija sijoittaa tällöin havaintoyksikön luokkaan, jonka ehdollinen todennäköisyys on suurin. Ehdollisiin todennäköisyyksiin perustuvat luokittelijat voidaan muodostaa joko diskriminatiivisesti tai generatiivisesti. Edellisessä estimoidaan suoraan luokkamuuttujan ehdollista jakaumaa vastaava malli kun taas jälkimmäisessä estimoidaan ensin havaintoyksikön ominaisuuksia kuvaavien muuttujien sekä luokkamuuttujan yhteisjakaumaa vastaava malli, josta etsitty ehdollinen jakauma saadaan käyttämällä Bayesin kaavaa. Tutkimuksessa tarkastellaan binääriseen luokitteluun soveltuvaa, diskriminatiivisesti muodostettavaa logistista regressiota sekä naiivia Bayes-luokittelijaa, joka tiettyjen oletusten vallitessa on tämän generatiivinen vastine.

Modernissa tilastotieteessä on viime vuosina huomattavasti lisääntynyt ns. bayesläisten menetelmien käyttö. Ominaista näille menetelmille on kaiken tilastollisen epävarmuuden ilmaiseminen todennäköisyysjakaumien avulla. Tässä työssä tutkitaan kokeellisesti bayesläisen lähestymistavan vaikutusta naiivin Bayes-luokittelijan ja logistisen regressiomallin luokitustarkkuuteen. Tämän lisäksi tarkastellan diskriminatiivisten ja generatiivisten luokittelumallien välisiä eroja ja arvioidaan opetusaineiston koon vaikutusta näiden luokituskykyyn. Luokittelumallien vertailussa käytetään Tampereen yliopistollisesta sairaalasta peräisin olevaa aineistoa, joka koostuu sepelvaltimovarjoainekuvattujen potilaiden terveydentilaa koskevista tiedoista.

Luokitustarkkuudeltaan generatiivinen luokittelija oli diskriminatiivista luokittelijaa parempi, joskin erot pienenivät mitä suuremmaksi opetusaineiston kokoa kasvatettiin. Tämä on sopusoinnussa kirjallisuudessa esitetyn tuloksen kanssa, jonka mukaan generatiiviset luokittelijat ovat diskriminatiivisia luokittelijoita tarkempia juuri pienillä opetusaineistoilla kun taas jälkimmäiset ovat tarkempia suurilla opetusaineistoilla. Bayesläisen lähestymistavan soveltaminen paransi jossain määrin kummankin mallin luokituskykyä etenkin pienimmillä opetusaineistoilla.

Avainsanat-Nyckelord-Keywords
logistinen regressio
regressioanalyysi
naiivit Bayes-luokittelijat
bayesilainen mallikeskiarvoistaminen
bayesilaiset menetelmät
binääriset luokittelumenetelmät
tilastomenetelmät
priorijakaumat

Säilytyspaikka-Förvaringsställe-Where deposited

Muita tietoja-Övriga uppgifter-Additional information