Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Edwart Ždanovitš

Müügikvaliteedi parandamine tugivektormasinate abil

Magistritöö

finants- ja kindlustusmatemaatika erialal (30 EAP)

Juhendaja:

Raul Kangro (PhD)

TARTU
2017

Müügikvaliteedi parandamine

tugivektormasinate abil

Magistritöö

Edwart Ždanovitš

Lühikokkuvõte. Käesoleva töö eesmärgiks on leida statistilise õppe meetod parandamaks laenutoote müügikvaliteeti. Probleemipüstitus taandub kahe klassiga klassifitseerimisülesandele. Töö keskseks statistilise õppe meetodiks on tugivektormasinad (TVM). Ühe osa tööst moodustavad TVM mittesümmeetrilised kaofunktsioonid. Nimetatud meetodeid rakendatakse testandmestikule – tulemused on toodud töö viimases osas. Võrdleva meetodina kasutatakse klassifitseerimispuud.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine , finants- ja kindlustusmatemaatika.

Märksõnad: Tugivektormasinad, mittesümmeetriline kaofunktsioon, statistilised katsed

Sales quality improving by support vector machines

Master’s thesis

Edwart Ždanovitš

Abstract. The objective of these master’s theses is to find the machine learning method to improve the credit product sales quality. The aim is to solve a two group classification problem - divide observation to risky and not risky groups. Central classification method in these papers is support vector machine classifier (SVM). One part from the papers is involving SVM non-symmetric loss functions. Classification tree method is used as reference method. Methods are applied to the data set – results are presented in the final part of the papers.

CERCS research specialisation: P160 Statistics, operational research, programming, actuarial mathematics.

Keywords: support vector machine classifier, non-symmetric loss function, testing

Sisukord

Sissejuhatus

Käesoleva töö eesmärgiks on leida statistilise õppe meetod parandamaks laenutoote müügikvaliteeti. Probleemipüstitus taandub kahe klassiga klassifitseerimisülesandele – eristamist vajavateks klassideks on hea maksekäitumisega laenutoote kliendid ning riskantsemad ehk tõenäoliselt kehvemini laenu teenindavad kliendid. Kuna ajas järjest rohkem kogutakse kliendi kohta erinevaid andmeid, siis võib tekkida huvi, kas nende põhjal on võimalik ennustada klientide maksekäitumist.

Magistritöö on algab ülevaatega meetodist ning kaofunktsioonidest. Seejärel rakendatakse kirjeldatud meetodid ja referentsmeetodit avalikul Saksa krediidi näidisandmestikul [1] ning võrreldakse tulemusi. Töö kekseks statistilise õppe meetodiks on tugivektormasinad. Võrdleva meetodina kasutatakse klassifitseerimispuud . Selle meetodi kirjeldust ei esitata.

Main coal of current papers is to introduce and use the statistical learning methods for predicting the credit default rate for each client. The aim is to solve a two group classification problem - divide observation to risky and not risky groups. According to nowadays more and more wider data collecting on personal level motives to improve or invent methods how to interpret or predict their behavior. Project starts with the method description and related loss functions and ends with the results obtained by the two methods are presented and compared. Central classification method in these papers is support vector machine classifier. Classification tree method is used as reference method. Methods are tested and compared on German Credit data [1].

1 Probleemi detailne kirjeldus

1.1 Probleemi kirjeldus

Käesoleva töö on fokuseeritud kahe klassiga klassifitseerimisülesande lahendamisele. Üldise klassifitseerimisülesande eesmärgiks on leida klassifitseerija g, mis jaotab objektid K erinava grupi vahel vastavalt objektile kuuluva p kirjeldava tunnuse x_{ij} alusel, tehes seejuures võimalikult vähe kulukaid valesti liigitamisi. Objekti uuritav tunnus y_i määrab, millisesse gruppi ta kuulub. Valesti klassifitseerimiseks loetakse sündmust, kus objekt liigitatakse gruppi mis ei ole objekti tegelik grupp. Võttes aluseks näiteks laenutoote klientide andmestik, on gruppi 1 kuuluvad objektid need finantsasutuse kliendid, kes krediiditoodet kasutades on jäänud r päevasesse võlgnevusse. Gruppi -1 kuuluvad need kliendid, kelle tooted ei ole üle r päevases võlas.

Üle r päevases võlas olevad klientide lepingud loetakse ebatõenäoliselt laekuvateks ning need müüakse maha. Lõplik eesmärk on vähendada kulu ja tõsta tulu - suurendada kasumit. Iga õigesti gruppi -1 liigitatud leping toob ettevõttele tulu intressi näol, iga gruppi 1 õigesti liigitatud leping hoiab ära kulu, mis tekkib krediiditoote mitte teenindamisest. Üldiselt on iga mitteteenindava nõude turuväärtus madalam kui tema jääkväärtus, seega mitteteenindava nõude müügiga kaasneb kahju. Nendest eraldi osa moodustavad mitteteenindavad lepingud, kus teenindatud osalt saadud tulu ületab lepingu jäägi, kuid neid eraldi ei käsitleta. Käesoleva probleemipüstituse juures tuleb otsus alati langetada, seega ei uurita võimalikku kahju või kasu otsuse tegemata jätmise korral.

Klassifitseerimismeetod, mis suudab kõik halvaks minevad ehk tulevikus mitteteenindavad laenud ennustada halva krediidikäitumisega klientide gruppi, ei pruugi olla kasu toov meetod, kui sealjuures ka enamus häid laene klassifitseeritakse halvaks minevateks. Töös lähtutakse olukorrast, kus erinevatesse gruppidesse valesti liigitamine toob erineval määral kahju. Jättes tehingu sõlmimata põhjusel, et hea maksekäitumisega leping liigitati valesse grupp, kaotatakse tulu, mis üldjuhul on oluliselt väiksem kui kulu, mis kaasneb tulevikus mitteteenindava lepingu sõlmimistest. Selleks, et kirjeldatud situatsioonis erinevad valesti otsustamised omaks võrreldavat mõju, defineeritakse kaofunktsioon.

Probleemi lahendamisel ei eeldata, et kirjeldava ja uuritavate tunnuste vahel on ainult lineaarsed sõltuvused. Seega kasutatud meetodid võimaldavad ka mittelineaarsete seoste kirjeldamist tunnuste vahel.

1.2 Kaofunktsioon ja risk

Olgu Y uuritava tunnuse võimalike väärtuste hulk ehk käesolevas töös hulk $\{1, -1\}$. Kaofunktsioon

$$L: Y \times Y \rightarrow \mathbb{R} \quad (1.1)$$

näitab kahju, mis tekib objekti klassist *i* klassi *j* liigitamisel. Kahju, mis tekib õigesti hindamisel on null ehk $L(i, i) = 0$ iga klassi *i* korral. Lähtudes eelnevalt püstitatud probleemikirjeldusest, on eesmärgiks leida klassifitseerija, mis minimiseerib kahju. Kuna objekt ja tema klass on juhuslikud, on juhuslik ka klassifitseerija kadu. Parimaks loetakse klassifitseerija, mille keskmine kahju ehk risk on minimaalne.

Klassifitseerija g risk on keskmine kahju üle tunnusvektori ja ühisjaotuse $F(x, y)$:

$$R(g) = \int L(y, g(x)) dF(x, y). \quad (1.2)$$

Valesti klassifitseerimisel tekkiv kahju on seotud paljude väliste teguritega. Kui majandusel läheb hästi, on valesti hindamise kahju reeglina väiksem, kuna klientide maksevõime on suurem, müüdud tooted võimaldavad teenida suuremat intressi ning võlas nõudeid saab edasi müüa suurema hinnaga. Teisalt on konkureerivaid tooteid ning neid teenuseid pakkuvaid ettevõtteid rohkem. Järelikult ei ole otstarbekas fikseerida kindlat valesti hindamise kahju. Käesolevas töös kasutatakse kolme erinevat kahjumäära – autori poolt valitud vähimat kahju saamist, võimalikku maksimaalset kahju saamist ja nimetatute vahepealset. Maksimaalseks kahju juhtumina käsitletakse olukorda, kus võlga sattunud nõudeid ei õnnestu maha müüa. Minimaalseks kahju saamisena vaadatakse olukorda, kus mitteteenindavat tagatiseta tarbimislaenu nõuete portfelli on võimalik maha müüa 40% selle väärtusest. Kolmas, ehk eelnevate keskmine kahju, kirjeldab olukorda, kus 20% võlaportfellist õnnestub maha müüa. Eeldades lihtsustatult, et korrektselt teenindavate laenude portfellis on intress 20%, on maksimaalsele kahjule, keskmisele kahjule ja minimaalsele kahjule vastavad kaofunktsioonid järgmised:

$$L_1(i,j) = \begin{cases} 0, kui\ i = j \\ 1, kui\ i = -1\ , j = 1 \\ 5, kui\ i = 1, j = -1 \end{cases} \quad (1.3)$$

$$L_2(i,j) = \begin{cases} 0, kui\ i = j \\ 1, kui\ i = -1\ , j = 1 \\ 4, kui\ i = 1, j = -1 \end{cases} \quad (1.4)$$

$$L_3(i,j) = \begin{cases} 0, kui\ i = j \\ 1, kui\ i = -1\ , j = 1 \\ 3, kui\ i = 1, j = -1. \end{cases} \quad (1.5)$$

2 Tugivektorklassifitseerija

Käesolev peatükk tugineb teatmikul [6, pt. Support Vector Machines, lk. 337-375] ning loengukonspektil [9, lk.31,64-85]. Tugivektorklassifitseerijad on maksimaalse marginaaliga klassifitseerija üldistus. Kogu meetodi kirjeldus on toodus kaheklassilise klassifitseerimisülesande näitel.

2.1 Klassifitseerimine kahte klassi hüpertasandi abil

Olgu X Hilberti ruum, $f(x) = \langle \omega, x \rangle + \omega_0$ sellel ruumil antud lineaarne funktsionaal ning H olgu funktsionaali f abil defineeritud afiinne hulk ehk hüpertasand:

$$H = \{x: \langle \omega, x \rangle + \omega_0 = 0\} \quad (2.1)$$

[9, lk. 31].Vaadates näitena p-mõõtmelist ruumi $\mathbb{R}^p$, on hüpertasand defineeritud järgmiselt:

$$\langle \omega, x \rangle + \omega_0 = 0, \quad (2.2)$$

kus $\omega = (\omega_1, \dots, \omega_p)$ ning $\langle \omega, x \rangle = \sum_{j=1}^{p} \omega_j x_j$. Juhul, kui viimase võrduse asemel on võrratus, siis punkt x ei asu hüpertasandil vaid ühel või teisel pool tasandit selles ruumis. Esindagu uuritav tunnus kahte gruppi väärtustega -1 ja 1, ehk $K = 2$ ning $y = 1$ või $y = -1$. Olgu $x^* = (x_1^*, \dots, x_p^*)$ testobjekt, mida soovitakse klassifitseerida ning y^* hinnang, kuhu klassi uuritav objekt kuulub. Eeldades, et eksisteerib hüpertasand, mis suudab täielikult eraldada kõik teadaoleva klassikuuluvusega objektid nii, et kõik klassi 1 objektid x rahuldavad kõik võrratust $\langle \omega, x \rangle + \omega_0 > 0$ ning kõik klassi -1 objektid rahuldavad võrratust $\langle \omega, x \rangle + \omega_0 < 0$, siis on võimalik seada testobjekti klassifitseerimiseeskiri:

$$\text{kui } g(x^*) = sign(\langle \omega, x^* \rangle + \omega_0) = 1, \text{ siis loetakse objekt } x^* \text{ klassi 1 kuuluvaks ehk } y^* = 1 \quad (2.3)$$

ja

$$\text{kui } g(x^*) = sign(\langle \omega, x^* \rangle + \omega_0) = -1, \text{ siis loetakse objekt } x^* \text{ klassi -1 kuuluvaks ehk } y^* = -1. \quad (2.4)$$

Teisisõnu, kui leidub täielikult treeningandmeid eraldav hüpertasand, siis on võimalik objektid jaotada esimesse või teise klassi vastavalt reeglile, kas need vaatlused asetsevad ühel või teisel pool tasandit. Seega klassifitseerija $g(x)$ antud klassifitseerimisülesande korral võib defineerida kujul $g(x) = sign\big(f(x)\big)$.

2.2 Maksimaalse marginaaliga eraldaja

Hüpertasandi poolt täielikult eraldatavate vaatluste korral on neid eraldavaid hüpertasandeid võimalik leida lõpmatul hulgal. Eesmärgiks on nende seast valida selline hüpertasand, mille korral kõik treeningpunktid ehk vaatlused on temast kõige kaugemal. Sellist tasandit nimetatakse maksimaalse marginaaliga eraldajaks. Selleks on vaja leida iga vaatluse kaugus d ehk marginaal otsitavast tasandist. Suvalise punkti x kaugus tasandist *H* avaldub kujul:

$$d(x, H) = \frac{|\langle \omega, x \rangle + \ \omega_0|}{||\omega||} = \frac{|f(x)|}{||\omega||} \tag{2.5}$$

[9, lk. 31]. Võttes $||\omega||$ võrdseks ühega on $|\langle \omega, x \rangle + \ \omega_0|$ punkti x kaugus tasandist. Seega kui $f(x)$ erineb nullist palju, siis asub ta hüpertasandist kaugel. Teisalt kui $f(x)$ on nullilähedane arv, siis asub ta hüppertasandi lähedal. Üks võimalus maksimaalsete marginaaliga eraldaja leidmiseks on lahendada optimiseerimisülesanne:

$$\max_{\omega, \omega_0} d, \tag{2.6}$$

$$tingimusel, et\ ||\omega|| = 1, \tag{2.7}$$

$$y_i(\langle \omega, x_i \rangle + \ \omega_0) \geq d\ iga\ i = 1, \dots, n\ korral. \tag{2.8}$$

Pärast toodud ülesande lahendamist on osad punktid leitud hüpertasandist täpselt kaugusel d ning ülejäänud punktid kaugemal kui d. Saab näidata, et leitav hüpertasandi määravad ainult need punktid, mis asuvad kaugusel d. Seega hüpertasand jääks samaks, kui eemaldada andmestikust kõik punktid x_i, mis asuvad täieliku andmestiku jaoks leitud hüpertasandist kaugemal kui d [9, lk. 70].

Neid punkte, mis asuvad leitud hüpertasandist kaugusel d, nimetatakse tugivektoriteks. Marginaaltasanditeks nimetatakse maksimaalse marginaaliga eraldajaga paralleelseid tasandeid, mis läbivad kummagi grupi tugivektoreid ning seega asuvad maksimaalse

marginaaliga eraldajast kaugusel d. Kahemõõtmelises ruumis on tasanditeks sirged. Kahe kirjeldava tunnusega maksimaalse marginaaliga eraldaja näide on toodud joonisel 1.

Eelnevalt toodud optimiseerimisülesande ekvivalentne kuju on:

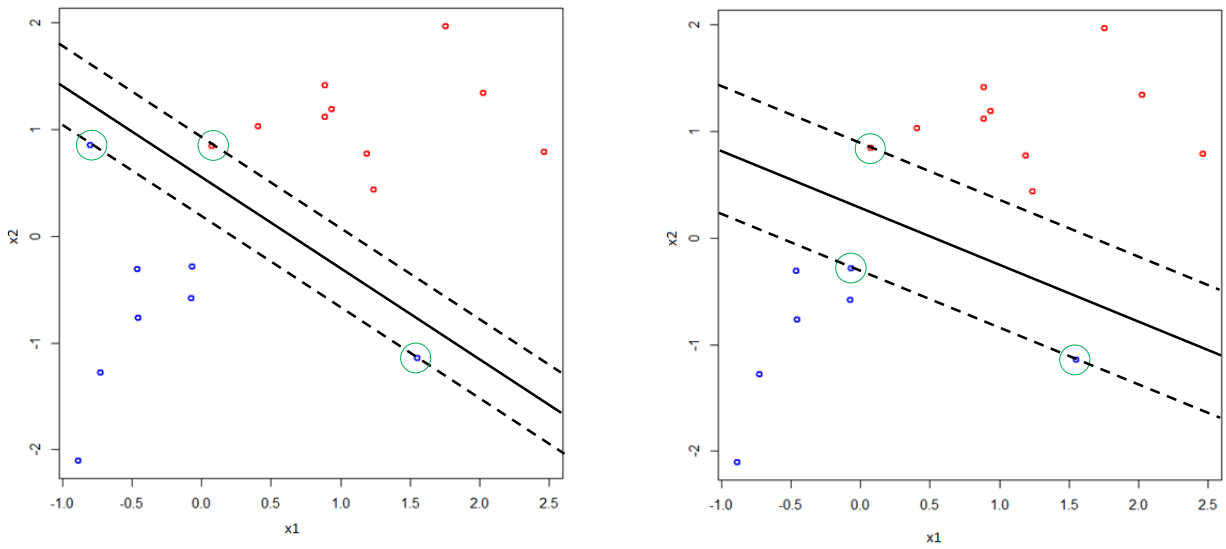
$$\min_{\omega,\omega_0} \frac{1}{2}||\omega||^2 \tag{2.9}$$

$$tingimusel, et$$

$$y_i(\langle \omega, x_i \rangle + \omega_0) \geq 1, iga\ i = 1, \dots, n\ korral, \tag{2.10}$$

mida tihti kasutatakse optimiseerimisülesande lahendamiseks.

Võib juhtuda ning üldiselt ka nii on, et tugivektorite hulk valimis võrreldes valimi suurusega on väike. Kuna hüperatasandi paiknemine on määratud tugivektorite põhjal, siis igasugune tugivektori vahetumine mõne muu tasandist kaugel või sootuks teisel pool tasandit asuva vaatlusega võib avaldab tugevat mõju hüpertasandi paiknemisele. Järelikult iga järgneva valimi tugivektorite poolt määratud hüpertasandi paiknemine võib oluliselt erineda eelneva põhjal määratud hüpertasandist.



Joonis 1. Sinised punktid on gruppi -1 ja punased punktid gruppi 1 kuuluvad objektid. Pidev sirge on maksimaalse marginaali eraldaja ning katkendsirged on marginaalsirged. Tugivektorid on ümbritsetud roheliste ringidega. Vasakul on sama valim peale ühe gruppi -1 kuuluva objekti eemaldamist.

2.3 Tugivektorklassifitseerija

Eelnevalt kirjeldatud maksimaalse marginaaliga eraldaja leidmine on võimalik ainult selliste vaatluste korral, mida on võimalik tasandi abil täielikult eraldada. Tugivektorklassifitseerija on sarnane oma ülesehituselt maksimaalse marginaaliga eraldajale, kuid ei eelda, et vaatlused peavad olema täielikult eraldatud, ehk teisisõnu on lubatud vaatluste paiknemine ka valel pool otsitavat klassifitseerivat tasandit. Järelikult, selline lähenemine annab võimaluse klassifitseerida ka selliseid andmestikke, mille korral ei ole võimalik klasse hüpertasandiga täielikult eraldada. Tugivektoreraldaja leidmiseks tuleb lahendada eelnevaga võrreldes mõnevõrra täiendatud optimeerimise ülesanne:

$$\min_{\omega,\omega_0} \frac{1}{2}\left|\left|\omega\right|\right|^2 + C\sum_{i=1}^{n} e_i \tag{2.11}$$

$$tingimusel, et$$

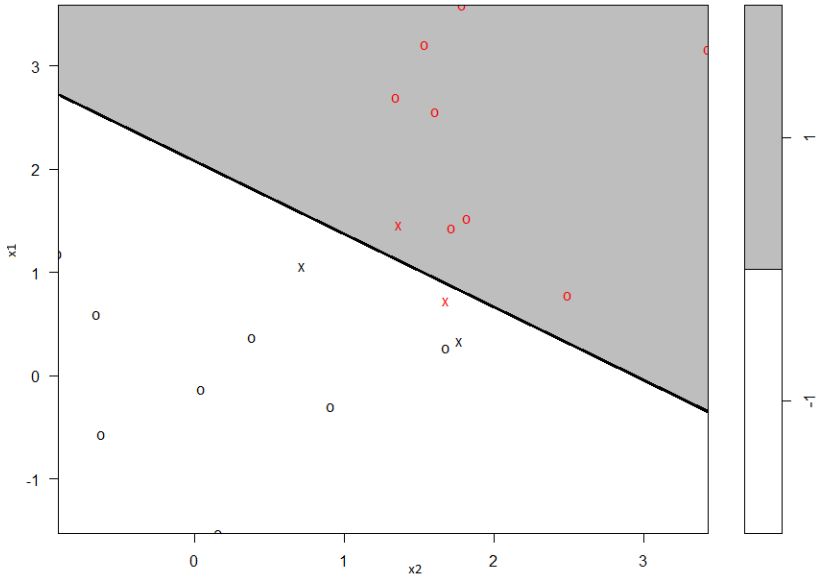
$$y_i(\langle \omega, x_i \rangle + \ \omega_0) \geq 1 - e_i, e_i \geq 0 \ iga \ i = 1, \dots, n. \tag{2.12}$$

Lahendades optimeerimisülesanne, leitakse klassifitseerimiseks kasutatav hüpertasand. Taaskord vastavalt funktsiooni $f(x) = \langle \omega, x \rangle + \omega_0$ märgile sõltub, kas testobjekt x loetakse kuuluvaks ühte või teise klassi. Lisandunud muutuja e_i võimaldab vaatlusel x_i paikneda valel pool marginaaltasandit või koguni valel pool hüpertasandit. Väljendi „valel pool“ all peetakse silmas olukorda, kus peale ülesande lahendamist näiteks gruppi -1 kuuluv objekt asub gruppi 1 klassifitseeritaval tasandi poolel.

Ülesande lahendi korral e_i on võrdne nulliga, kui *i*-s vaatluse asub õigel pool marginaaltasandit. Kui e_i väärtus on vahemikus (0,1), asub *i*-s vaatlus õigel pool klassifitseerimistasandit, selle ja marginaaltasandi vahel tekitatud piirkonnas. Muutuja e_i väärtus üks ja suurem leiab aset olukorras, kus *i*-s vaatlus on valel pool hüpertasandit. Tugivektormasinad, erinevalt maksimaalse marginaali eraldajast, ei määrata hüpertasandit ainult lähimate erinevatesse klassidesse kuuluvate vaatluste abil vaid lubatud hulgal lähimate ja valel pool otsitavat tasandit asuvate vaaluste põhjal. Sarnaselt maksimaalse marginaaliga eraldajale, ei kasuta tugivektorklassifitseerija õigel pool marginaaltasandit asuvaid, kuid väljas pool marginaaltasandit asuvaid vaatlusi ehk need ei oma mõju tugivektorklassifitseerija leidmisel. Kahe kirjeldava tunnusega tugivektoreraldaja on toodud joonisel 2.

Mittenegatiivne parameeter C on seadistuse parameeter. Parameeter C võimaldab optimiseerimisülesande lahendamisel anda kaalu valel poole marginaaltasandit asetsevate vaatluste kauguste e_i summale. Võttes C piisavalt suur, on tugivektorklassifitseerija optimiseerimise ülesanne samaväärne maksimaalse marginaaliga eraldaja leidmisega.

Praktiliste ülesannete lahendamisel on üheks võimaluseks määrata parameeter C rist-valideerimine abil.



Joonis 2. Tugivektorklassifitseerija lahutab halli ja valget ala. Ristiga tähistatud objektid on tugivektorid. On näha, et üks gruppi 1 kuuluv tugivektor asub valel pool tasandit.

2.4 Parameetri *C* leidmine ristvalideerimise abil

Masinõppemeetodi treenimiseks ning testimiseks on vaja andmestikku. Andmestik võib olla juhuslik n objektist koosned valim uuritavast populatsioonist. Kasutades kogu valimit meetodi õpetamiseks, puudub meil teadmine meetodi headusest rakendatuna populatsioonile. Selle tulemusena võib paljude mudelite hulgast osutuda valituks mudel, mis omab väga häid näitajaid treenimiseks kasutatud andmestikul, kuid ei pruugi omada sama häid näitajaid mõnel muul sama populatsiooni valimil või populatsioonil endal. Kirjeldatud nähtust nimetatakse mudeli ülesobitamiseks. Üks võimalik lähenemine on jaotada valim kaheks osaks – treeningandmestik n_1 et $n_1 < n$ ning testandmestik n_2 et $n_2 = n - n_1$. Objektid jaotatakse kahe andmestiku vahel

juhuslikult. Treeningandmesiku abil leitakse sobivad mudelid. Mudeleid testitakse testandmestikul ning valitakse nende hulgast parim. Statistiliste meetodite võrdlemine jaguneb kaheks etappiks. Esmalt leitakse testimise alusel valitud meetodi õiged parameetrid, ehk sobitatakse sama meetodi erinevaid mudeleid ning valitakse neist parim. Seda tehakse iga statistilise meetodi korral, mida plaanitud kasutada. Teise etapina võrreldakse erinevate meetodite parimaid mudeleid omavahel.

On ilmselge, et juhulikkuse alusel saadud treeningandmestik võib mõjutada nii meetodit ennast, kui ka meetodi valikut. Järelikult võib olla ekslik leida sobiv meetod ühe treeningandmestiku põhjal. Erinevate treeningandmestike jaoks on võimalik kasutada k-alagrupilist ristvalideerimise meetodi. Saadud valim jaotatakse k-alamgruppi. Järgemööda valitakse üks alamgrupp testandmestikuks ning ülejäänuid $k-1$ alamgruppi kasutatakse treeningandmetena. Seega valitud statistilise õppe meetodit rakendatakse k korda nii moodustatud $k-1$ alamgruppist moodusutatud treeningandmetel ning testitakse treenimisel mittekasutatud alamgrupil. Meetodi headust mõõdetakse test alagruppide mõõtmistulemuste keskmisena. Kuna töö eesmärk on seatud 2 klassiga klassifitseerimise probleemi lahendamisele, siis mõõdetavaks meetodi headuse näitajaks võib võtta näiteks vähima kahju. Otsitav k-alamgruppiga ristvalideerimise statistik avaldub kuju:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} Err_i, \tag{2.13}$$

kus $Err_i = L(y_i \neq \hat{y}_i)$ [6, pt. k-Fold Cross-Validation]. On ilmselge, et mida suurem on k, seda rohkem on vaja meetodit treeningandmetele sobitada. Seega meetodid, mis nõuavad palju ajalist ressurssi, muutuvad ristvalideerimise korral veelgi aeganõudvamaks protsessiks.

Kokkuvõttes toimub valitud statistilise õppe meetodite testimine kahel erineval andmestikul. Esmalt testitakse meetodi käitumist erinevate häälestusparameetrite korral ristvalideerimise käigus treeningandmestikust eraldatud testalamandmestikel ning seejärel mõõdetakse parimate parameetritega meetodi headust ning võrreldakse saadud tulemusi testandmestikku kasutades. Viimast ei ole kaasatud enam ristvalideerimise protsessi.

2.5 Mittelineaarsed klassifitseerijad

Kui uuritava ja kirjeldavate tunnuste vaheline seos on mittelineaarne, siis lineaarse hüpertasandi kasutamine ei pruugi anda head klassifitseerimise tulemust. Selleks, et saada mittelineaarne klassifitseerija, võib kasutada mittelineaarset eraldajat. Üks võimalus mittelineaarse tugivektorklassifitseerija defineerimiseks on kasutatavate argumenttunnuste hulga laiendamine arvutatavate tunnustega. Näiteks sobitades tugivektorklassifitseerijat *p* kirjeldava tunnusega

$$X_1, X_2, \dots, X_p \tag{2.14}$$

asemel hoopis 2*p* tunnusega:

$$X_1, X_1^2, X_2, X_2^2, \dots, X_p, X_p^2 \tag{2.15}$$

saab tugivektorklassifitseerija leida optimeerimisülesande:

$$\min_{\omega} \frac{1}{2} \sqrt{\omega_1^2 + \cdots + \omega_p^2 + (\omega_1')^2 + \cdots + (\omega_p')^2} + C \sum_{i=1}^{n} e_i \tag{2.16}$$

$$tingimusel, et$$

$$y_i(\langle \omega, x_i \rangle + \langle \omega', x_i^{*2} \rangle + \ \omega_0) \geq 1 - e_i, e_i \geq 0 \ \ iga\ i = 1, \dots, n\ korral, \tag{2.17}$$

lahendamisel teel. See aga tähendab uuritava tunnuse lähendamist ruutfunktsiooniga $f(x) = \langle \omega, x_i \rangle + \langle \omega', x_i^2 \rangle + \ \omega_0$. Sarnaselt võib kasutada ka kõrgema astme polünoome või veelgi keerulisemaid funktsioone.

3 Tugivektormasinad

Käesolev peatükk tugineb loengukonspektile [9, Tugivektorklassifitseerijad ja teised tuumameetodid]. Mittelineaarse klassifitseerija leidmiseks kasutatav optimeerimisülesanne on keerulise polünoomi ja suure andmemahu juures suurt arvutusressurssi nõudev protsess. Tugivektormasinad on tugivektorklassifitseerijatel põhinev meetod, mis kasutab klassifitseerimisülesande lahendamiseks tuumasid.

3.1 Tuumad

Eeldades, et leidub piisavalt punkte x_i, et nende kohavektorite abil saab esitada kõiki vektoreid ω vaadeldavas ruumis, siis saab lineaarse eraldaja esitada skalaarkorrutise abil kujul:

$$f(x) = \omega_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x \rangle, \tag{3.1}$$

kus α_i on igale treeningvaatlusele i vastav parameeter valimis suurusega n. Leidmaks $f(x)$, tuleb hinnata parameetreid ω_0 ja $\alpha_1, \dots, \alpha_n$. Viimaste hindamiseks tuleb leida skalaarkorrutis kõigi treeningvaatluste vahel. Järelikult tuleb arvutada $n(n-1)/2$ skalaarkorrutist. On ilmselge, et treeningvaatluste arvu kasvades suureneb skalaarkorrutiste arv nagu vaatluste arvu ruut ning klassifitseerimismeetodi treenimine muutub järjest rohkem ressurssi nõudvamaks protsessiks. Peale meetodi (3.11) ja (3.12) duaalsele kujule viimist ning optimaalse lahendi leidmist selgub, et kui treeningpunkt x_i ei osutunud tugivektoriks, siis α_i on võrdne nulliga [9, lk. 70]. Need summeritavad punktid, mis ei ole tugivektorid ehk mille α_i kordaja on 0, ei avalda uue vaatluse x^* korral klassifitseerija väärtuse $f(x^*)$ leidmisel mõju ning nende punktide ja x^* vahelisi skalaarkorrutisi ei ole vaja arvutada. Teisisõnu uuritava objekti klassifitseerimiseks ei ole vaja leida skalaarkorrutist objekti ning kõigi treeningandmete vahel, vaid objekti ja tugivektorite vahel. Olgu tugivektorite hulk tähistatud *S*-iga, siis tugivektoreraldaja saab esitada kujul:

$$f(x) = \omega_0 + \sum_{i \in S} \alpha_i \langle x_i, x \rangle. \tag{3.2}$$

Mittelineaarse klassifitseerija saamiseks on vaja leida teatav kujutis φ, mis teisendab ruumis $\mathbb{R}^p$ olevad tunnusvektorid Hilberti ruumi W. Kujutise kasutamise idee on püüda teisendada

uuritav tunnus ja kirjeldavad tunnused mingist ruumist ruumi W, kus uuritava tunnuse ja argumenttunnuste vaheline seos oleks lineaarselt paremini kirjeldatav kui eelnevas ruumis. Eeldades, et nüüd ruumis W on uuritava ja kirjeldavate tunnuste vaheline seos küllaltki lineaarne, võib seal kasutada uuritava tunnuse klassi määramiseks mõnda lineaarset klassifitseerijat. Ilmselgelt kasutatakse käesolevas töös äsja defineeritud lineaarset tugivektorklassifitseerijat. Kasutades sobivat mittelineaarset teisendust φ avaldub mittelineaarne tugivektorklassifitseerija kujul:

$$f(x) = \omega_0 + \sum_{i \in S}^{n} \alpha_i \langle \varphi(x_i), \varphi(x) \rangle. \tag{3.3}$$

Tugivektormasinate idee seisneb selles, et tunnusvektorid teisendatakse alati kõrgema dimensiooniga ruumi. Seega ruumi W dimensioon on suurem kui ruumi $\mathbb{R}^p$ dimensioon ning W dimensioon võib olla ka lõpmatu. Hinnatavad konstandid $\alpha_1, \dots, \alpha_n$ sõltuvad teisendusest φ läbi skalaarkorrutiste $\langle \varphi(x_i), \varphi(x) \rangle$. Selgub, et $\langle \varphi(x_i), \varphi(x) \rangle$ on võimalik leida ilma teisendust φ kasutamata kui on teada funktsioon:

$$K: \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}, \qquad K(y, x) = \langle \varphi(y), \varphi(x) \rangle. \tag{3.4}$$

Tihti on selle funktsiooni analüütiline kuju leitav ning see teeb tugivektorklassifitseerimise võimalikuks. Funktsiooni K nimetatakse tuumaks. Eelnevalt toodud tugivektoreraldaja saab esitada kujul:

$$f(x) = \omega_0 + \sum_{i \in S}^{n} \alpha_i K(x_i, x) \tag{3.5}$$

[9, lk. 89]. Saadud mittelineaarse tuumaga klassifitseerijat nimetatakse tugivektormasinaks.

3.2 Radiaaltuum

Radiaaltuum on mittelineaarne tuum, mis avaldub kujul:

$$K(y,x) = \exp(-\gamma \sum_{j=1}^{p} (y_j - x_j)^2), \tag{3.6}$$

kus γ on positiivne konstant. Mida kaugemal on treeningpunkt testvaatlusest, seda suurem on tuumas toodud summa väärtus. Teisalt, mida suurem on nimetatud summa, seda väiksemat väärtust omab tuuma funktsioon. Järelikult testvaatlusest kaugemal asetsevate treeningpunktide puhul on tuuma funktsiooni väärtus väiksem ning lähemal asuvate korral suurem [6, lk. 352-353].

Praktikas on võimalik väärtust γ leida sobitamise teel. Meetodi treenimisel antakse ette γ väärtuste vahemik ning valituks osutub γ mille korral ristvalideerimise statistik on minimaalne. On ilmselge, et iga treenimiseks etteantud γ väärtuse korral tuleb leida ka eelpool nimetatud seadistuse parameeter C. Kasutades t erinevat γ, u erinevat C väärtust ning k-alamgrupilist ristvalideerimist, tuleb meetodit treenida $t \times u \times k$ korda. Kirjeldatud treenimise protsess võib sellest tulenevalt osutuda küllaltki aeganõuvaks protsessiks.

4 Tugivektormasin mittesümmeetrilise kaofunktsiooni korral

4.1 Sümmeetriline kaofunktsioon

Olgu klassifitseerimisprobleem endiselt kaheklassiline, kus uuritava tunnuse y tinglik jaotus on kujul:

$$Y|(X=x) = \begin{cases} 1, tn. \mu(x) \\ -1, \quad tn. 1-\mu(x), \end{cases} \tag{4.1}$$

kus $\mu(x) = P(Y=1|X=x)$ on tinglik tõenäosus. Tunnused Y ja X on juhuslikud suurused ning nende ühisjaotuse jaotusfunktsiooniks on $F(x,y)$. Töö esimeses pooles kirjeldatud risk on esitatav kujul:

$$\begin{aligned} R(g) &= \int L\big(y, g(x)\big) dF(x,y) \\ &= \int \big[\mu(x) L\big(1, g(x)\big) + \big(1-\mu(x)\big) L\big(-1, g(x)\big)\big] dF(x). \end{aligned} \tag{4.2}$$

Klassifitseerija g on parim, kui $R(g)$ on minimaalne. Seega eesmärgiks on leida kõigi sobilike klassifitseerijate hulgast klassifitseerija g, mis minimiseerib riski piisavalt hästi ehk keskmine kahju oleks võimalikult väike.

Tavapärase sümmeetrilise 0-1 kaofunktsiooni:

$$L_{0,1}(y,g) = \frac{1-yg(x)}{2} = \begin{cases} 0, kui\ y = g(x) \\ 1, kui\ y \neq g(x) \end{cases} \tag{4.3}$$

korral avaldub valemi (5.2) viimase integraali all olev tinglik risk kujul:

$$\begin{aligned} R_{0,1}(g|X=x) &= \mu(x)\frac{1-g(x)}{2} + \big(1-\mu(x)\big)\frac{1+g(x)}{2} \\ &= \begin{cases} \mu(x), kui\ g(x) = -1 \\ 1-\mu(x), kui\ g(x) = 1. \end{cases} \end{aligned} \tag{4.4}$$

Seega parim klassifitseerija ehk Bayesi klassifitseerija on:

$$g^*(x) = \begin{cases} 1, & kui\ \mu(x) > \frac{1}{2} \\ -1, & kui\ \mu(x) \leq \frac{1}{2}. \end{cases} \tag{4.5}$$

Olgu nüüd $g(x) = sign(f(x))$. Saadud riski minimiseerib iga funktsioon f, mille korral:

$$\begin{cases} f(x) > 0, kui\ \mu(x) > \frac{1}{2} \\ f(x) = 0, kui\ \mu(x) = \frac{1}{2} \\ f(x) < 0, kui\ \mu(x) < \frac{1}{2} \end{cases} \tag{4.6}$$

Üks nõutud tingimust rahuldavatest funktsioonidest on $f(x) = 2\mu(x) - 1$, mis on ühtlasi ka Bayesi klassifitseerija. Selle keskmine kahju on Bayesi risk:

$$\begin{aligned} R_{0,1}^*(g) = \int \mu(x) \left(\frac{1}{2} - \frac{1}{2} sign(2\mu(x) - 1)\right) \\ + (1 - \mu(x)) \left(\frac{1}{2} + \frac{1}{2} sign(2\mu(x) - 1)\right) dF(x), \end{aligned} \tag{4.7}$$

mis väikseim võimalik.

4.2 Mittesümmeetriline kaofunktsioon 1

Olgu C_1 kahju, mis tekib tegeliku klassi 1 hindamisel klassiks -1 ning C_{-1} kahju, mis tekib klassi -1 hindamisel klassiks 1. Mittesümmeetrilise kaofunktsiooni korral $C_1 \neq C_{-1}$. Seega on soov kasutada kaofunktsiooni kujul:

$$L(y, g(x)) = \begin{cases} 0, y = g(x) \\ C_1, y = 1\ ja\ g(x) = -1 \\ C_{-1}, y = -1\ ja\ g(x) = 1. \end{cases} \tag{4.8}$$

Analoogselt eelnevaga leitakse tinglik risk:

$$\begin{aligned} R(g|X = x) &= C_1 \mu(x) \frac{1 - g(x)}{2} + C_{-1}\big(1 - \mu(x)\big) \frac{1 + g(x)}{2} \\ &= \begin{cases} C_1 \mu(x), kui\ g(x) = -1 \\ C_{-1}(1 - \mu(x)), kui\ g(x) = 1. \end{cases} \end{aligned} \tag{4.9}$$

Parim sellele kaofunktsioonile vastav klassifitseerija, Bayesi klassifitseerija, on seega defineeritud kujul:

$$g^*(x) = \begin{cases} 1, & (1-\mu(x))C_{-1} \leq \mu(x)C_1 \\ -1, & (1-\mu(x))C_{-1} > \mu(x)C_1 \end{cases} \tag{4.10}$$

ehk

$$g^*(x) = \begin{cases} 1, & \mu(x) \geq \dfrac{C_{-1}}{C_{-1}+C_1} \\ -1, & \mu(x) < \dfrac{C_{-1}}{C_{-1}+C_1}. \end{cases} \tag{4.11}$$

Kui suurus $\mu(x)$ oleks teada, siis võiks kasutada Bayesi klassifitseerijat. Kuna $\mu(x)$ ei ole teada, siis tuleb sobiv klassifitseerija leida riski minimiseerimise teel. Paraku on üle kõikide -1,1 väärtustega funktsioonide hulga minimiseerimine keeruline.

Kui nüüd õnnestub leida suvaline reaalarvuliste väärtustega „klassifitseerijat“ $f(x)$ kasutav kaofunktsioon, mis käitub riski minimiseerimisel samuti nagu ainult väärtusi -1 ja 1 kasutava klassifitseerija $g(x)$ jaoks eelnevalt defineeritud kaofunktsioon, siis võib lahendada riski minimiseerimise ülesannet üle kõikide reaalarvliste väärtustega mõõtuvate funktsioonide hulga. Hiljem saab klassifitseerimisreegli aluseks võtta näiteks reegli $g(x) = sign\big(f(x)\big)$, mis loob seose:

$$g(x) = \begin{cases} 1, & kui\ f(x) > 0 \\ -1, & vastasel\ juhul. \end{cases} \tag{4.12}$$

Olgu kaofunktsioon kujul:

$$\hat{L}(1,z) = C_1 \frac{\lfloor 1-z \rfloor_+}{2} \tag{4.13}$$

ning

$$\hat{L}(-1,z) = C_{-1} \frac{\lfloor 1+z \rfloor_+}{2}, \tag{4.14}$$

kus z on reaalarvuline väärtus ning $\lfloor z \rfloor_+ = \max\{0, z\}$. Defineeritud kaofunktsiooni korral avaldub tinglik risk kujul:

$$R_{C_1,C_{-1}}(f(x)|X = x) = \mu(x)C_1\frac{\lfloor 1 - f(x)\rfloor_+}{2} + (1 - \mu(x))C_{-1}\frac{\lfloor 1 + f(x)\rfloor_+}{2}. \quad (4.15)$$

Minimaalse riski saavutamiseks peab $f(x)$ olema selline reaalarv z, mis on ülesande:

$$\min_z \mu(x)C_1\frac{\lfloor 1 - z\rfloor_+}{2} + \big(1 - \mu(x)\big)C_{-1}\frac{\lfloor 1 + z\rfloor_+}{2} \quad (4.16)$$

lahendiks. Eeldades, et $C_1 > 0$ ja $C_{-1} > 0$, on minimiseeritav funktsioon (5.16) kahe murdekohaga tükiti lineaarne ning kumer funktsioon (vt. joonis 3). Sellise funktsiooni ekstremaalsed väärtused on kindlasti saavutatud murdepunktides. Seega vaadeldava optimaalne lahend saadakse, kui valitakse $f(x)$ väärtuseks 1 või -1 vastavalt selle, kumb annab väiksema tingliku riski väärtuse. Seega optimaalne lahend ülesandele:

$$\min_f \int \hat{L}\big(y, f(x)\big)dF(x, y) \quad (4.17)$$

avaldub kujul:

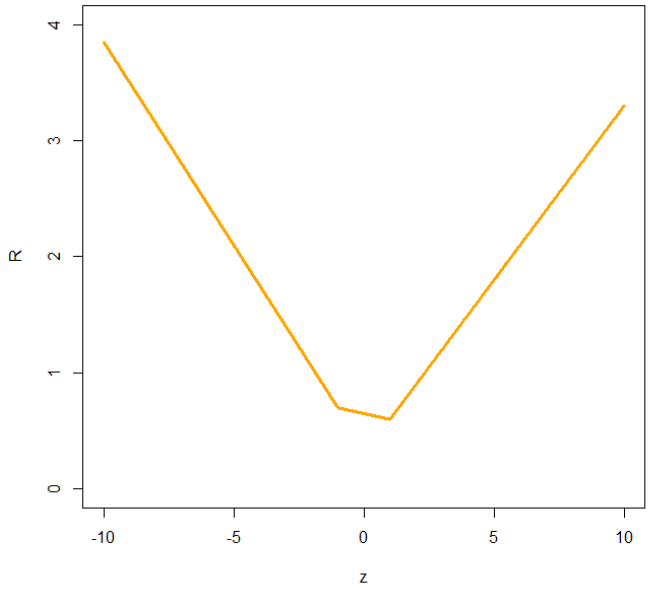
$$f(x) = \begin{cases} 1, & (1 - \mu(x))C_{-1} \le \mu(x)C_1 \\ -1, & (1 - \mu(x))C_{-1} > \mu(x)C_1. \end{cases} \quad (4.18)$$

ehk

$$f(x) = \begin{cases} 1, & \mu(x) \ge \dfrac{C_{-1}}{C_{-1} + C_1} \\ -1, & \mu(x) < \dfrac{C_{-1}}{C_{-1} + C_1}. \end{cases} \quad (4.19)$$

Praktilise ülesande lahendamiseks minimiseeritakse empiirilist riski, milleks on kaofunktsiooni väärtuste summa üle valimi. Võttes nüüd $f(x) = \langle \omega, x_i \rangle + \omega_0$, saadakse ülesanne:

$$\min_{\omega,\omega_0} \sum_{\{i|y_i=1\}}^{n} C_1\lfloor 1-(\langle\omega,x_i\rangle+\omega_0)\rfloor_+ + \sum_{\{i|y_i=-1\}}^{n} C_{-1}\lfloor 1+(\langle\omega,x_i\rangle+\omega_0)\rfloor_+ . \tag{4.20}$$



Joonis 3. Funktsioon $\mu(x)C_1\lfloor 1-z\rfloor_+ + (1-\mu(x))C_{-1}\lfloor 1+z\rfloor_+$, kus $C_1 = 2$, $C_{-1} = 1$ ning $\mu(x)$ on valitud 0,7.

Kuna aga vaadeldud ülesande (5.20) lahend ei pruugi olla ühene ning suure arvu tunnuste korral võib tekkida ülesobitamise oht, siis on vaja lisada ülesandesse liige, mille abil saab need probleemid lahendada. Seetõttu esitatakse ülesanne kujul, kus minimiseeritakse kombinatsiooni tinglikust riskist ja liigset keerukust karistavast liikmest $\frac{1}{2}||\omega||^2$. Seega ülesanne on kujul:

$$\min_{\omega,\omega_0} \sum_{\{i|y_i=1\}}^{n} C_1\lfloor 1-(\langle\omega,x_i\rangle+\omega_0)\rfloor_+ + \sum_{\{i|y_i=-1\}}^{n} C_{-1}\lfloor 1+(\langle\omega,x_i\rangle+\omega_0)\rfloor_+ + \frac{1}{2}||\omega||^2. \tag{4.21}$$

Defineeritud ülesandele lisatakse ka töös eelpool kirjeldatud seadistusparameeter C. Seega lõplik ülesanne primaarsel kuju on:

$$\min_{\omega,\omega_0,e} \frac{1}{2}\left|\left|\omega\right|\right|^2 + C\left[C_1 \sum_{\{i|y_i=1\}}^{n} e_i + C_{-1} \sum_{\{i|y_i=-1\}}^{n} e_i\right] \tag{4.22}$$

$$tingimusel, et$$

$$1 - e_i - y_i(\langle\omega, x_i\rangle + \omega_0) \leq 0, e_i \geq 0 \; iga \; i = 1 \ldots n \; korral. \tag{4.23}$$

Eelnevalt defineeritud tuumade rakendamiseks viiakse optimeerimise ülesanne duaalsele kujule Lagrange'i määramata kordajate meetodil. Eeldades esmalt, et $\langle\omega, x_i\rangle = \omega' x_i$, avaldub Lagrange'i funktsionaal kujul:

$$\begin{aligned} L(\omega, \omega_0, e, \alpha, \beta) &= \frac{\left|\left|\omega\right|\right|^2}{2} + C\left[C_1 \sum_{\{i|y_i=1\}}^{n} e_i + C_{-1} \sum_{\{i|y_i=-1\}}^{n} e_i\right] \\ &+ \sum_{i=1}^{n} \alpha_i\left(1 - e_i - y_i(\omega' x_i + \omega_0)\right) - \sum_{i=1}^{n} \beta_i e_i \end{aligned} \tag{4.24}$$

$$\begin{aligned} &= \frac{\left|\left|\omega\right|\right|^2}{2} + C\left[C_1 \sum_{\{i|y_i=1\}}^{n} e_i + \sum_{\{i|y_i=-1\}}^{n} e_i\right] + \sum_{i=1}^{n} \alpha_i - \sum_{i=1}^{n} \alpha_i e_i \\ &- \sum_{i=1}^{n} \alpha_i y_i \omega' x_i - \omega_0 \sum_{i=1}^{n} \alpha_i y_i - \sum_{i=1}^{n} \beta_i e_i, \\ &kus \; \alpha_i \geq 0, \beta_i \geq 0 \text{ iga } i = 1 \ldots n \; korral. \end{aligned} \tag{4.25}$$

Karush-Kuhn-Tuckeri tingimus: esialgse ülesande miinimumkoha ω^*, ω_0^* ja e^* korral leiduvad sellised mittenegatiivsed α_i ja β_i, et Lagrange'i funktsionaali kõik osatuletised on võrdsed nulliga [7]. Seega, leidmaks ekstreemumit, võetakse tuletis $L(\omega, \omega_0, e, \alpha, \beta)$ iga vektori ω ja vektori e komponendi kohta ning seejärel võrdsustatakse nulliga. Tuletist ω_0 järgi:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial \omega_0} = \sum_{i=1}^{n} \alpha_i y_i \; |= 0. \tag{4.26}$$

seega

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{4.27}$$

Tuletis vektori ω komponentide kaupa:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial \omega_j} = \ \omega_j - \sum_{i=1}^{n} \alpha_i y_i x_{ij} \ | = 0, \tag{4.28}$$

iga $j \in \{1, \dots, p\}$. Järelikult

$$\omega_j = \sum_{i=1}^{n} \alpha_i y_i x_{ij}, \tag{4.29}$$

iga $j \in \{1, \dots, p\}$ ning tuletis vektori e komponentide kaupa:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial e_i} = \ C(C_1 I_{\{i|y_i=1\}} + C_{-1} I_{\{i|y_i=-1\}}) - \alpha_i - \beta_i | = 0 \tag{4.30}$$

iga $i \in \{1, \dots, n\}$. Sellest α_i on esitatav kujul:

$$\alpha_i = \ C\left(C_1 I_{\{i|y_i=1\}} + C_{-1} I_{\{i|y_i=-1\}}\right) - \beta_i, iga\ i = \{1, \dots, n\}\ korral\ ning \tag{4.31}$$

$$\beta_i \geq 0 \text{ iga } i = \{1, \dots, n\}\ korral, \tag{4.32}$$

kus I on indikaatorfunktsioon. Kuna β_i ja α_i on mittenegatiivsed, siis tulemuse (5.31) põhjal $0 \leq \alpha_i \leq CC_1$, kui $y_i = 1$ ning $0 \leq \alpha_i \leq CC_{-1}$, kui $y_i = -1$, seda iga $i = \{1, \dots, n\}$ korral. Asendades saadud tulemused esialgsesse Lagrange'i funktsionaali, on tulemus järgmine:

$$L(\omega, \omega_0, e, \alpha, \beta) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j + \sum_{i=1}^{n} \alpha_i \tag{4.33}$$

$$0 \leq \alpha_i \leq CC_1, \text{kui } y_i = 1,$$
$$0 \leq \alpha_i \leq CC_{-1}, \text{kui } y_i = -1\ ja \tag{4.34}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0. \tag{4.35}$$

Seega duaalne ülesanne on kujul:

$$\max_{\alpha_i \in \mathbb{R}^n} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j \tag{4.36}$$

$$0 \leq \alpha_i \leq CC_1, \text{kui } y_i = 1,$$
$$0 \leq \alpha_i \leq CC_{-1}, \text{kui } y_i = -1 \; ja \tag{4.37}$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0. \tag{4.38}$$

Olgu optimiseerimisülesanne kujul:

$$\min_{\alpha_i \in \mathbb{R}^n} f(x) \tag{4.39}$$

$$tingimusel, et$$

$$g_i(x) \leq 0, iga \; i = 1, \dots, n \; korral, \tag{4.40}$$

siis Slateri tingimuste:

1. funktsioon f ja g_i on kumerad;
2. leidub $x_0 \in \mathbb{R}^n$ nii, et $g_i(x) < 0$ iga $i = 1, \dots, n$ korral

täidetuse korral saab näidata, et primaarne ja duaalne ülesanne on võrdsed sadulpunktis [9, lk. 64-66]. On lihtne näha, et primaarse ülesande korral on Slateri tingimused täidetud, seega duaalse ülesande lahendi α^* abil saame leida esialgse ülesande lahendivektori kujul:

$$\omega^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i. \tag{4.41}$$

Kui ω^* on teada, siis saab parameetri ω_0^* määrata kasutades mõnda tugivektorit x_i:

$$\omega_0^* = y_i - \omega^{*'} x_i. \tag{4.42}$$

Karush-Kuhn-Tuckerti tingimused sellisel juhul on:

$$\alpha_i^* \geq 0, iga\ i = 1 \ldots n\ korral, \tag{4.43}$$

$$\alpha_i^*(1 - e_i - y_i(\omega_i^{*'} x_i + \omega_0^*)) = 0, iga\ i = 1 \ldots n\ korral. \tag{4.44}$$

Tingimuse (5.41) põhjal on selge, et $\alpha_i^* > 0$ ainult siis, kui $1 - e_i = y_i(\omega' x_i + \omega_0)$ ehk punkti x_i kaugus on võrdne marginaaliga, kui $e_i = 0$ või marginaaliga, millest on maha lahutatud e_i kui $e_i > 0$. Teisisõnu on α_i^* suurem nullist ainult tugivektorite korral.

Mittesümmeetrilise kaofunktsiooniga tugivektormasina saamiseks tuleb $x_i' x_j$ asendada tuumafunktsiooniga ning lahendada saadud duaalne ülesanne. Testobjekti x^* klassifitseerimiseeskirja võib nüüd esitada kujul:

kui

$$g(x^*) = sign\left(\omega_0 + \sum_{i \in S}^{n} y_i \alpha_i K(x_i, x^*)\right) = 1, \tag{4.45}$$

siis loetakse objekt x^* klassi 1 kuuluvaks ning ülejäänud juhtudel klassi -1 kuuluvaks.

4.3 Mittesümmeetriline kaofunktsioon 2

Töös tuuakse ka teine kaofunktsiooni definitsioon, mis on esitatud artiklis [10]. Eelnevalt vaadeldud mittesümmeetrilisele kaole vastava riski minimiseerimise ülesanne taandub lineaarselt eralduva valimi ja piisavalt suure C korral maksimaalse marginaaliga eraldaja leidmisele, st klassifitseeriv hüpertasand paigutatakse sama kaugele mõlema klassi tugivektoritest. Võib argumenteerida, et mõistlikum võiks olla paigutada see tasand lähemale selle klassi vaatlustele, mille valesti klassifitseerimine toob kaasa väiksema kahju. Artiklis [10] on toodud minimiseerimisülesanne, mille lahendil on selline omadus. Olgu C_1 endiselt kahju, mis tekib tegeliku klassi 1 hindamisel klassiks -1 ning C_{-1} kahju, mis tekib klassi -1 hindamisel klassiks 1. Olgu kaofunktsioon kujul:

$$L_{C_1, C_{-1}}(y, g(x)) = \frac{1 - yg(x)}{2}\left(C_1 \frac{1 - g(x)}{2} + C_{-1} \frac{1 + g(x)}{2}\right)$$

$$= \begin{cases} 0, kui\ y = g(x) \\ C_1, kui\ y = 1\ ja\ g(x) = -1 \\ C_{-1}, kui\ y = -1\ ja\ g(x) = 1. \end{cases} \tag{4.46}$$

Vaadates defineeritud tulemuse erandit, kus $C_1 = C_{-1}$ on tegu sümmeetrilise kaofunktsiooniga. Kaofunktsioonile $L_{C_1,C_{-1}}(y, g)$ vastav tinglik risk avaldub nüüd kujul:

$$R_{C_1,C_{-1}}(g|X = x) = \mu(x)C_1\frac{1 - g(x)}{2} + (1 - \mu(x))C_{-1}\frac{1 + g(x)}{2} = \begin{cases} C_1\mu(x), kui\ g(x) = -1 \\ C_{-1}(1 - \mu(x)), kui\ g(x) = 1. \end{cases} \tag{4.47}$$

Lähtuvalt defineeritud riskist, on artiklis [10] toodud parameetrite hindamiseks vajalik minimiseerimise ülesanne kujul:

$$\min_{\omega,\omega_0} \sum_{\{i|y_i=1\}}^{n} \lfloor C_1 - C_1(\langle \omega, x_i \rangle + \omega_0) \rfloor_+ + \sum_{\{i|y_i=-1\}}^{n} \lfloor 1 - (2C_{-1} - 1)(\langle \omega, x_i \rangle + \omega_0) \rfloor_+ + \frac{1}{2}||\omega||^2, \tag{4.48}$$

mis primaarsel kujul on:

$$\min_{\omega,\omega_0} \frac{1}{2}||\omega||^2 + C\left[C_1 \sum_{\{i|y_i=1\}} e_i + (2C_{-1} - 1) \sum_{\{i|y_i=-1\}} e_i\right] \tag{4.49}$$

$$tingimusel, et$$

$$-y_i(\langle \omega, x_i \rangle + \omega_0) + \frac{y_i + 1}{2} - \frac{y_i - 1}{2(2C_{-1} - 1)} - e_i \leq 0\ ja$$

$$e_i \geq 0\ iga\ i = 1 \ldots n\ korral. \tag{4.50}$$

Nii nagu ka eelmise kaofunktsiooni kirjelduse korral, viiakse ülesanne (5.50) tuumade kasutamise eesmärgil duaalsele kujule. Eeldades, et $\langle \omega, x_i \rangle = \omega' x_i$, avaldub Lagrange'i funktsionaal kujul:

$$
\begin{aligned}
&L(\omega, \omega_0, e, \alpha, \beta) \\
&= \frac{||\omega||^2}{2} + C\left[C_1 \sum_{\{i|y_i=1\}} e_i + (2C_{-1} - 1) \sum_{\{i|y_i=-1\}} e_i\right] \\
&+ \sum_{i=1}^{n} \alpha_i \left(-y_i(\omega' x_i + \omega_0) + \frac{y_i + 1}{2} - \frac{y_i - 1}{2(2C_{-1} - 1)} - e_i\right) \\
&- \sum_{i=1}^{n} \beta_i e_i \qquad (4.51) \\
&= \frac{||\omega||^2}{2} + C\left[C_1 \sum_{\{i|y_i=1\}}^{n} e_i + (2C_{-1} - 1) \sum_{\{i|y_i=-1\}}^{n} e_i\right] \\
&- \sum_{i=1}^{n} \alpha_i y_i \omega' x_i - \omega_0 \sum_{i=1}^{n} \alpha_i y_i \\
&+ \sum_{i=1}^{n} \alpha_i \left(\frac{y_i + 1}{2} - \frac{y_i - 1}{2(2C_{-1} - 1)}\right) - \sum_{i=1}^{n} \alpha_i e_i - \sum_{i=1}^{n} \beta_i e_i, \\
&kus\ \alpha_i \geq 0, \beta_i \geq 0, e_i \geq 0\ iga\ i = 1 \ldots n\ korral. \qquad (4.52)
\end{aligned}
$$

Ekstreemumite leidmiseks võetakse taaskord tuletis $L(\omega, \omega_0, e, \alpha, \beta)$ iga vektori ω ja vektori e komponendi kohta ning seejärel võrdsustatakse nulliga. Tuletist ω_0 järgi on:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial \omega_0} = \sum_{i=1}^{n} \alpha_i y_i \ | = 0. \qquad (4.53)$$

seega

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \qquad (4.54)$$

Järgnevalt leitakse tuletis vektori ω komponentide kaupa:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial \omega_j} = \omega_j - \sum_{i=1}^{n} \alpha_i y_i x_p^j \ | = 0, \qquad (4.55)$$

iga $j \in \{1, \ldots, p\}$, seega:

$$\omega_j = \sum_{i=1}^{n} \alpha_i y_i x_i, \tag{4.56}$$

iga $j \in \{1, \dots, p\}$. Seejärel võetakse tuletis vektori e komponentide kaupa:

$$\frac{\partial L(\omega, \omega_0, e, \alpha, \beta)}{\partial e_i} = CC_1 I_{\{i|y_i=1\}} + C(2C_{-1} - 1) I_{\{i|y_i=1\}} - \alpha_i - \beta_i| = 0 \tag{4.57}$$

iga $i \in \{1, \dots, n\}$. Seega:

$$\alpha_i = CC_1 I_{\{i|y_i=1\}} + C(2C_{-1} - 1) I_{\{i|y_i=1\}} - \beta_i, \tag{4.58}$$

iga $i \in \{1, \dots, n\}$. Kuna β_i ja α_i on mittenegatiivsed, siis tulemuse (5.58) põhjal $0 \leq \alpha_i \leq CC_1$, kui $y_i = 1$ ning $0 \leq \alpha_i \leq C(2C_{-1} - 1)$, kui $y_i = -1$. Asendades saadud tulemused esialgsesse Lagrange'i funktsionaali, on tulemus järgmine:

$$\begin{aligned} L(\omega, \omega_0, e, \alpha, \beta) \\ &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i' x_j \\ &+ \sum_{i=1}^{n} \alpha_i \left(\frac{y_i + 1}{2} - \frac{y_i - 1}{2(2C_{-1} - 1)} \right) \end{aligned} \tag{4.59}$$

$$0 \leq \alpha_i \leq CC_1, \text{kui } y_i = 1, \tag{4.60}$$

$$0 \leq \alpha_i \leq C(2C_{-1} - 1), \text{kui } y_i = -1 \; ja$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0. \tag{4.61}$$

Mittesümmeetrilise kaofunktsiooniga tugivektormasina saamiseks tuleb jällegi $x_i' x_j$ asendada tuumafunktsiooniga ning lahendada saadud duaalne ülesanne.

5 Andmestik ning meetodi sobitamine

5.1 Saksa krediidi andmestik

Saksa krediidi näidisandmestik [1] on internetist leitav näidisandmestik, mis koosneb 1000 objektist. Uuritavaid gruppe on kaks – võlas ja mitte võlas kliendid. Andmestikus on toodud vastav binaarne üks-null tunnus. Võlas olevate klientide arv on 300, ehk 30% kõigist vaatlustest.

Andmestik koosneb diskreetsetest ja pidevastest tunnustest. Kokku on 20 kirjeldavat tunnusest. Meetodite sobitamisel eeldatakse, et „Ülalpeetavate arv“, „Telefon“ ja „Võõrtööline“ on diskreetsed tunnused ning kõik ülejäänud 17 on pidevad tunnused. Tunnuste loetelu on toodud lisas 1.

Andmestik jaotatakse juhuslikkuse alusel kahte gruppi suhtes 80% ja 20%. Väiksem ehk 200 objektiga grupp on testandmestik ning 800 objektiga grupp on treeningandmestik. Igat meetodit treenitakse 10 juhuslikul testandmestiku järjestusel, kasutades 10 alamgrupilist ristvalideerimist. Parimaks loetakse mudel, mille korral keskmine Win (vt. valem 6.1) väärtus üle kõigi järjestuste on suurim. Kirjeldatud protsess viiakse läbi iga halvaks läinud portfelli müügi osakaalu korral.

5.2 Otsustuspuu sobitamine

Otsustuspuu meetodi kirjeldused on toodud [6, pt. Tree-Based Methods] ning [4, pt. Otsustuspuu]. Otsustupuu sobitamiseks kasutatakse tarkvara R paketti „rpart“ [2]. Kuna valesti klassifitseerimine erinevates klassides toob kaasa erineval määral kahju, siis kasutatakse mittesümmeetrilist kaofunktsiooni. Töö esimeses pooles toodud kolmele erinevale kahju saamisele vastavad mittesümmeetrilise kaofunktsiooni eeljaotused on:

Loss	$\tilde{\pi}_1$	$\tilde{\pi}_{-1}$
L_1	0,68	0,32
L_2	0,63	0,37
L_3	0,56	0,44

Otsustuspuu kasvatamiseks kasutatakse Gini indeksit ning 10-alamgruppilist riskvalideerimist. Kärbitud otsustuspuu suurust määrav keerukusparameeter leitakse igale eeljaotusele eraldi ristvalideerimise abil, kasutades testandmestikku.

Tulemuse põhjal, mis saadakse otsustuspuu meetodi rakendamisel testandmetele, arvutatakse suurus Win, mis on defineeritud kujul:

$$Win = \left(\frac{\sum_{i=1}^{n_2} I_{y_i=1,\hat{y}_i=1}}{\sum_{i=1}^{n_2} I_{y_i=1,\hat{y}_i=1} + \sum_{i=1}^{n_2} I_{y_i=1,\hat{y}_i=-1}} (1 - sr) od - \frac{\sum_{i=1}^{n_2} I_{y_i=-1,\hat{y}_i=1}}{\sum_{i=1}^{n_2} I_{y_i=-1,\hat{y}_i=1} + \sum_{i=1}^{n_2} I_{y_i=-1,\hat{y}_i=-1}} i_{rate} (1 - od) \right) P, \quad (5.1)$$

kus

- P – on portfelli suurus
- od – on halvaks minevate laenude osakaal portfellis ehk hetkel 30%
- sr – halvaks läinud portfelli müügist saadud tulu osakaal halba portfelli
- $I_{y=\ldots,\hat{y}=\ldots}$ – indikaatorfunktsioon, kus $\hat{y}_i$ on i-nda objekt hinnang
- n_2 – testobjektide arv
- i_{rate} – teenindavalt portfellilt teenitava intressimäär.

Parameeter Win näitab, kui suur hulk raha hoitakse kokku suurusega P portfelli pealt, kui rakendatakse vastava meetodi mudelit. Seega väärtust Win kasutatakse edasises meetodite headuse võrdlemiseks. Ilmselgelt loetakse parimaks mudel, mille korral nimetatud parameeter on suurim.

Teise etapina peale keerukusparameetri leidmist treenitakse otsustuspuu meetodit 10-l juhulikul treeningandmete järjestusel. Igal järjestuse korral kasutatakse taaskord 10-alamgruppilist ristvalideerimist.

5.3 Tugivektormasinate sobitamine

Tugivektor masina sobitamiseks on võimalik kasutada erinevaid tuumasid. Käesolevas töös kasutatakse eelnevalt kirjeldatud radiaaltuuma.

Mudeli sobitamiseks viiakse pidevat tüüpi tunnused normaliseeritud kujule ehk keskväärtusega 0 ning dispersiooniga 1. Diskreetsete tunnuste väärtused asendatakse väärtustega 1 ja -1.

Parim mudel, sõltuvalt tugivektormasina seadistusparameetrile C ning radiaaltuuma parameetrile γ leitakse testimise teel, andes ette erinevaid C ja γ kombinatsioone. Iga kombinatsiooni korral kasutatakse 10-alamgrupilist ristvalideerimist. Parameetreid hinnatakse iga halvast portfellist müüduks osutunud osakaalu korral. Mudeli headuse hindamiseks kasutatakse eelnevas alam-peatükkis defineeritud parameetri Win väärtust.

Kui sobivad parameetrid C ja γ on leitud, siis järgmise sammuna sobitatakse 10 erineval testandmestiku järjestusel samade parameetritega mudeleid. Iga järjestuse korral kasutatakse taaskord 10-alamgruppilist ristvalideerimist.

Tugivektormasinate rakendamiseks on olemas tarkvara R pakett „e1071“ [5]. Nimetatud pakett võimaldab sobitada ainult sümmeetrilise kaofunktsiooniga mudeleid. Kuna puudub teadmine, et eksisteerib avalik R lähtekood, mis võimaldab kasutada töö esimeses pooles kirjeldatud mittesümmeetrilisi kaofunktsioone, siis kirjutatakse vastav optimeerimise ülesanne programmi R abil. Autori poolt loodud kood on andmekandjal pandud kaasa töö kirjalikule versioonile. Ülesande lahendamiseks jaoks kasutatakse R pakette „quadprog“[3], "caret"[8] ning "matrixcalc"[11].

6 Rakendatud meetodite tulemused ja võrdlused

Käesolevas peatükis kirjeldatakse töö praktilise poole tulemused. Kasutatud otsustuspuu mudelite keerukusparameetrid CP, tugivektormasinate seadistusparameetrid C ja radiaaltuuma väärtused γ on toodud tabelites 2 kuni 4, kus on toodud vastavate mudelite treenimistulemused. Lisa 2 on toodud parameetri Win väärtus erinevate C, γ ja sr väärtuste korral.

6.1 Tulemused

Esmalt vaadatakse olukorda, kus halvaks läinud nõudeid ei õnnestu maha müüa ehk iga mitteteenindava nõudega saadakse maksimaalne kahju. Tabelis 1 on toodud meetodite treenimiste tulemused 10 juhusliku testandmestiku järjestuse korral. Vaadates tabeli viimast veergu ehk keskmisi üle järjestuste, on ilmselge, et tugivektormasinad on selle andmestiku seadistuse ja ülesande püstituse korral rohkem tulutoovamad kui otsustuspuu meetod. Erinevate kaofunktsioonidega tugivektormasinaid võrreldes on näha, et mittesümmeetrilise kaofunktsiooniga meetodid annavad keskmiselt parema tulemuse. Teisalt, jättes kõrvale keskmised tulemused üle järjestuste, on näha, et mittesümmeetrilist kaofunktsiooni kasutavad tugivektormasinad on oluliselt ebastabiilsemad – sõltuvalt järjestusest võivad nad anda vägagi erinevaid tulemusi. Ehk et mittesümmeetriliste kaofunktsioonide korral on standardhälve oluliselt kõrgem.

Tabel 1. Väärtus Win 10 juhusliku testandmesiku järjestuse korral, kui laenu mitte teenindavat portfelli ei õnnestu müüa.

	Meetod			
Järjestus	**Otsustuspuu CP: 0,017**	**S-TVM C:40, γ:0,0075**	**MS1-SVM C:0,75, γ:0,03**	**MS2-SVM C:0,75, γ:0,03**
1	88 008	123 392	22 421	32 785
2	100 171	123 945	172 397	175 033
3	67 042	130 237	139 170	136 858
4	109 507	125 516	189 516	181 328
5	107 743	123 793	97 903	86 152
6	91 468	112 559	184 823	193 201
7	92 238	119 189	94 155	104 883
8	95 327	123 346	181 909	190 209
9	79 923	122 523	159 508	168 414
10	80 990	115 403	178 454	187 392
Keskmine	**91 242**	**121 990**	**142 026**	**145 625**

	Otsustuspuu CP**: 0,017**	**S-TVM** C**:40,** γ**:0,0075**	**MS1-SVM** C**:0,75,** γ**:0,03**	**MS2-SVM** C**:0,75,** γ**:0,03**
Standardhälve	**13 050**	**5 075**	**54 553**	**54 427**

Olukorras, kus laenu mitte teenindavast portfellist õnnestub maha müüa 20% ulatuses selle väärtusest, on keskmiste Win väärtuste erinevus meetodite lõikes väiksem kui eelneva puhul. Suurim erinevus on tugivektorite ja otsustuspuu mudelite vahel. Seega vaadeldava andmestiku korral oleks igati põhjendatud eelistada tugivektormasinaid otsustuspuudele.

Tabel 2. Väärtus Win 10 juhusliku testandmesiku järjestuse korral, kui halvast portfellist õnnestub maha müüa viiendik.

	Meetod			
Järjestus	**Otsustuspuu** CP**: 0,01**	**S-TVM** C**:15,** γ**:0,01**	**MS1-SVM** C**:0,5,** γ**:0,05**	**MS2-SVM** C**:15,** γ**:0,0075**
1	73 347	101 248	99 194	98 698
2	72 002	87 953	138 800	130 304
3	72 747	94 471	37 592	67 155
4	84 411	90 154	44 518	71 337
5	75 462	96 009	140 081	110 947
6	72 188	93 282	8 524	33 942
7	78 662	91 386	144 597	133 767
8	74 661	93 511	99 468	107 188
9	72 885	88 704	136 357	129 163
10	80 031	93 009	139 919	118 205
Keskmine	**75 640**	**92 973**	**98 905**	**100 071**
Standardhälve	**4 119**	**3 861**	**50 931**	**32 792**

Viimane stsenaarium, kus halvast portfellist õnnestub maha müüa 40% selle väärtusest, on vaadeldava andmestiku korral taaskord parimat tulemust kajastavad meetodid mittesümmeetrilist kaofunktsiooni kasutavad tugivektormasinad. Endiselt tagastab halvimaid tulemusi otsustuspuu meetod. Kuigi mittesümmeetrilist kaofunktsiooni kasutavatel tugivektormasinatel on väga kõrge standardhälve, siis sümmeetrilise kaofunktsiooniga tugivektormasinal on see meeldivalt madal kõigil kolmel korral.

Tabel 3. Väärtus Win 10 juhusliku testandmestiku järjestuse korral, kui mitteteenindavast portfellist õnnestub maha müüa 40%.

	Meetod			
Järjestus	**Otsustuspuu CP: 0,016**	**S-TVM C:10, γ:0,01**	**MS1-SVM C:0,5, γ:0,01**	**MS2-SVM C:0,75, γ:0,05**
1	59 470	64 973	65 709	78 270
2	56 790	64 780	73 318	75 803
3	54 663	65 358	65 358	65 395
4	52 364	64 641	88 642	88 564
5	53 526	63 856	84 644	83 499
6	52 092	65 125	23 713	49 481
7	51 850	65 715	61 918	73 756
8	46 186	62 953	83 709	79 213
9	43 223	59 424	90 359	76 552
10	54 331	66 923	80 355	59 436
Keskmine	**52 449**	**64 375**	**71 773**	**72 997**
Standardhälve	**4 744**	**2 033**	**19 717**	**11 713**

Viimaks rakendatakse eelneva põhjal välja valitud mudeleid testandmestikule. Tulemused on toodud tabelis 4. On näha, et kohati võib otsustuspuu anda paremaid tulemusi, kui sümmeetrilise kaofunktsiooniga tugivektormasin. Mittesümmeetrilise kaofunktsiooniga tugivektormasinad suudavad käesoleva andmestiku seadistuse korral anda parimaid tulemusi. Kohati isegi paremaid, kui eelneva treenimise käigus. See võib olla põhjustatud juhuslikkuse poolt, kuid samas on seekord ka treeningandmestik mõne võrra suurem.

Tabel 4. Tulemused, rakendades saadud mudeleid testandmestikul

Meetod	**Õigesti hinnatud**		**Valesti hinnatud**		**Kaotatud kasulik müük**	**Võidetud halb müük**	**Tulu**
	Hea krediit	**Halb krediit**	**Tegelik hea krediit**	**Tegelik halb krediit**			
Otsustuspuu							
L1	136	19	16	29	10,53%	39,58%	104 013
L2	124	23	28	25	18,42%	47,92%	89 211
L3	118	28	34	20	22,37%	58,33%	73 684
S - TVM							
L1	140	20	12	28	7,89%	41,67%	113 947
L2	142	20	10	28	6,58%	41,67%	90 789
L3	141	20	11	28	7,24%	41,67%	64 868
MS1-TVM							
L1	98	41	54	7	35,53%	85,42%	206 513
L2	100	40	52	8	34,21%	83,33%	152 105
L3	100	37	52	11	34,21%	77,08%	90 855

Meetod	Õigesti hinnatud		Valesti hinnatud		Kaotatud kasulik müük	Võidetud halb müük	Tulu
	Hea krediit	**Halb krediit**	**Tegelik hea krediit**	**Tegelik halb krediit**			
MS2 - TVM							
L1	98	41	54	7	35,53%	85,42%	206 513
L2	103	38	49	10	32,24%	79,17%	144 868
L3	98	39	54	9	35,53%	81,25%	96 513

Kokkuvõte

Käesoleva töö eesmärgiks oli leida statistilise õppe meetod parandamaks laenutoote müügikvaliteeti – hinnata kliendi tunnuste alusel tema makseraskustesse jäämist. Töö keskseks meetodiks olid tugivektormasinad, mida võrreldi otsustuspuu meetodiga. Kuna erinevad valesti klassifitseerimised tõid erineval määral kahju, siis kasutati mudelite sobitamisel mittesümmeetrilisi kaofunktsioone.

Tehtud katsete käigus selgus, et defineeritud andmestiku seadistuse korral suudavad tugivektormasinad anda märkimisväärselt paremaid tulemusi, kui otsustuspuu meetod. Eriti häid tulemusi andsid just mittesümmeetrilist kaofunktsiooni kasutavad tugivektormasinad. Võttes arvesse, et makseraskustesse sattuvate klientide hulk selles andmestikus on 30%, siis rakendades mittesümmeetrilist kaofunktsiooni kasutavat tugivektormasinat, vähenes see parematel juhtudel ligikaudu 85%. Samal ajal kui otsustuspuu suutis halva krediidikvaliteediga klientidest paremal juhul ära tunda ligikaudu 58%. Hoolimata väga heast mittesümmeetrilist kaofunktsiooni kasutava tugivektormasina keskmistest treeningtulemusest, tekitas vastuolulisi tundeid tulemuste küllaltki lai varieerumine.

Main purpose of the current paper was to introduce and use the statistical learning methods for predicting the credit default rate for each client - divide observed clients into risky and not risky groups. Central classification method in this paper was support vector machine classifier. Classification tree method was used as reference method. According to the fact that different miss-classifications returns different loss, the non-symmetric loss function approach was used.

The symmetric and non-symmetric loss function implementing support vector machines returned remarkable results comparing to the decisions tree method, especially the non-symmetric loss function using ones. According to the fact that overdue rate was 30% in used data whilst using named data setup and support vector machine with non-symmetric loss function, it was possible to reduce this rate approximately 85% in average. At the same time when decision tree model on that data setup reduced the overdue rate approximately 58%. On the other hand, variation between the results on different training data sequences was very high while using non-symmetric loss function producing support vector machine.

Kasutatud kirjandus

[1] Analysis of German Credit Data. https://onlinecourses.science.psu.edu/stat857/node/215. Külastatud 29.04.2016.

[2] Atkinson, B., Ripley B., Therneau, T. (2015). Package *'rpart'*. http://cran.r-project.org/web/packages/rpart/rpart.pdf. Külastatud 04.02.2016.

[3] Berwin, A., Turlach, R., Weingessel, A., (2015). Package *'quadprog'*. https://cran.r-project.org/web/packages/quadprog/quadprog.pdf. Külastatud 01.05.2016.

[4] Binsol, P., *Maksevõimetuse hindamine* (2015). http://dspace.ut.ee/bitstream/handle/10062/47541/binsol_paavo_msc_2015.pdf. Külastatud 04.02.2016.

[5] David, M., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C. (2015). Package *'e1071'*. https://cran.r-project.org/web/packages/e1071/e1071.pdf. Külastatud 16.02.2016.

[6] James, G., Witten, D., Hastie,T., Tibshirani, R. (2006). *An Introduction to Statistical Learning 4th.* http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Fourth%20Printing.pdf. Külastatud 15.11.2015.

[7] Karush–Kuhn–Tucker conditions. https://en.wikipedia.org/wiki/Karush%E2%80%93Kuhn%E2%80%93Tucker_conditions. Külastatud 13.01.2017.

[8] Kuhn M.. Contributions from Wing J., Weston, S., Williams, A., Keefer, C., Allan Engelhardt, A., Cooper T., Mayer, Z., Kenkel B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang,Y., Candan, C., Hunt.T. Package *'caret'*. ftp://cran.r-project.org/pub/R/web/packages/caret/caret.pdf. Külatatud 12.06.2016

[9] Lember, J. (2008). *Tehisõpe 1*. Loengukonspekt. http://www-1.ms.ut.ee/ained/Tehis%f5pe/tehisope8.pdf. Külastatud 29.02.2016.

[10] Masnadi-Shirazi, H., Vasconcelos, N. (2010). *Risk minimization, probability elicitation, and cost-sensitive SVMs*. San Diego, La Jolla, California.: Statistical Visual Computing Laboratory, University of California.

[11] Frederick, N. Package *'matrixcalc'*. ftp://cran.r-project.org/pub/R/web/packages/caret/caret.pdf. Külatatud 12.06.2016.

Lisa 1. Saksa krediidi andmestik

Kirjeldavad tunnused

Kirjeldav tunnus	Nimi andmestikus	Väärtus 1	Väärtus 2	Väärtus 3	Väärtus 4	Väärtus 5
Konto jääk	Account.Balance	Konto puudub	Jääk puudub	Jääk alla 200 SM	Jääk üle 200 SM	
Osamaksete staatus	Payment.Status.of.Previous.Credit	Hilines	Muu maksed	Makstud	Probleemid puuduvad	Eelnevad on makstud
Säästud	Value.Savings.Stocks	Puuduvad	Alla 100 SM	100 kuni 500 SM	500 kuni 1000 SM	Üle 1000 SM
Kliendi vanus	Age..years.	Vanus				
Laenu suurus	Credit.Amount	Summa				
Osamaksete arv	Duration.of.Credit..month.	Arv				
Praeguse töösuhte pikkus	Length.of.current.employment	Töötu	Alla 1 aasta	1 kuni 4 aastat	4 kuni 7 aastat	Üle 7 aasta
Osamakse suurus	Instalment.per.cent	Üle 35%	25%-35%	20% - 25%	Alla 20%	
Kvalifikatsioon	Occupation	Töötu	Ajutine töötaja	Oskustööline	Täidesaatev	
Sugu ja suhe	Sex...Marital.Status	Mees, lahutatud	Mees, vallaline	Mees, abielus	Naine	
Sama elukoht	Duration.in.Current.address	Alla aasta	1 kuni 4 aastat	4 kuni 7 aastat	üle 7 aasta	
Elukoha tüüp	Type.of.apartment	Vaba	Rendi	Omanik		
Väärtuslikuim vara	Most.valuable.available.asset	Puuduvad	Auto	Elukindlustus	Kinnisvara	
Krediite pangas	No.of.Credits.at.this.Bank	1	2 kuni 3	4 kuni 5	üle 6	
Käendajaid	Guarantors	Puuduved	Kaas taotleja	Käendaja		
Muud kohustused	Concurrent.Credits	Teise pangas	Järelmaks	Puudub		
Ülalpeetavate arv	No.of.dependents	Üle 3	Alla 3			
Telefon	Telephone	Jah	Ei			
Võõrtööline	Foreign.Worker	Jah	Ei			

Kir. tun.	Nimi and.	Vä. 1	Vä. 2	Vä. 3	Vä. 4	Vä. 5	Vä. 6	Vä. 7	Vä. 8	Vä. 7	Vä. 9
Laenu eesmärk	Purpose	Uus auto	Kasutatud auto	Mööbel	Raadio/TV	Muu seade	Parandus	Puhkus	Koolitus	Töö	Muu

Lisa 2. Seadistusparameetri C ja radiaaltuuma γ leidmine

Pahaks läinud nõudeid ei õnnestu müüa

S-TVM	Win						Keskmine tugivektorite arv					
C\	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	20 475	62 824	97 261	113 235	108 549	64 328	449	451	448	446	463	524
0.75	80 426	107 818	118 300	128 311	119 129	88 343	447	443	437	434	457	523
1	103 395	116 803	115 204	122 942	124 365	105 140	442	433	428	431	452	525
2	121 890	117 666	127 181	125 269	130 545	130 183	421	415	412	423	449	537
5	120 156	132 681	125 879	131 706	134 506	117 892	402	401	400	419	453	541
10	127 760	126 095	133 913	137 362	127 068	125 763	396	396	396	419	454	539
15	134 253	127 242	128 109	125 987	125 729	123 195	393	393	393	413	450	533
20	125 317	130 770	133 885	133 673	126 919	128 318	391	394	389	416	447	536
40	133 782	137 554	121 979	124 108	121 195	119 732	388	384	385	406	436	532
80	119 011	132 727	129 045	129 147	125 215	126 297	383	379	375	398	429	532
160	133 273	128 528	132 309	131 358	121 673	133 056	374	369	373	392	426	534
320	137 101	135 880	126 588	120 701	123 000	126 936	364	365	364	386	423	530
640	131 829	128 445	124 286	124 218	124 073	117 928	358	358	358	380	421	532

MS1-TVM	Win						Keskmine tugivektorite arv					
C/γ	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	66 376	129 627	132 169	169 554	190 875	181 435	580	566	561	550	559	611
0.75	114 748	135 903	148 691	196 338	179 984	157 576	562	552	547	538	545	589
1	146 471	146 431	134 589	190 005	177 391	145 012	553	544	541	529	536	577
2	169 965	156 300	178 982	184 876	160 645	130 513	533	531	527	508	511	552
5	162 321	169 738	176 188	162 585	137 914	115 181	521	513	504	474	478	540
10	176 444	176 646	171 413	145 728	129 540	123 221	511	494	482	452	453	537
15	172 802	179 521	151 928	136 783	112 865	134 738	500	483	470	438	441	533
20	177 134	167 009	152 011	131 852	115 262	124 058	491	472	460	426	437	533
40	165 266	161 655	156 809	120 191	128 704	113 096	471	450	437	407	434	531
80	150 780	153 977	142 075	116 209	123 575	110 101	448	430	416	393	425	532
160	146 533	153 907	130 191	129 700	124 939	123 661	425	411	399	389	423	534
320	145 009	123 301	130 256	129 430	118 834	117 598	406	389	380	381	425	533
640	145 955	141 433	128 190	126 534	122 794	120 285	394	378	364	378	426	531

MS2-TVM	Win						Keskmine tugivektorite arv					
C/γ	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	110 481	99 731	101 037	179 630	192 903	180 390	579	566	560	549	560	609
0.75	99 180	154 965	150 908	199 535	180 541	167 657	564	551	547	538	547	588
1	131 686	149 089	153 806	193 479	177 016	143 689	553	545	540	528	536	578
2	151 465	154 767	162 202	178 967	156 754	121 703	535	530	527	509	509	555
5	156 525	171 115	175 994	155 422	133 177	126 346	520	514	503	475	479	539
10	172 585	191 561	163 137	143 556	130 648	114 174	511	496	482	453	450	535
15	175 815	168 772	156 909	142 778	114 240	117 433	502	482	470	441	440	533
20	176 485	162 562	152 957	137 451	127 395	127 389	493	472	461	428	434	533
40	171 199	165 660	158 674	127 450	123 193	124 219	472	451	439	404	431	531
80	158 501	152 706	145 716	118 269	124 999	118 992	449	427	419	395	425	532
160	145 467	145 092	133 334	114 019	124 642	122 056	424	409	399	388	419	532
320	139 920	134 230	130 243	119 612	117 990	114 963	407	392	380	381	426	531
640	140 641	136 206	130 179	108 938	123 574	120 714	388	376	361	379	423	530

Pahaks läinud nõudeid õnnestub müüa 20% väärtuses

S-TVM	*Win*						Keskmine tugivektorite arv					
C/γ	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	16 570	52 446	70 237	82 446	80 903	45 730	449	451	450	447	462	526
0.75	63 319	84 381	90 019	96 739	96 585	70 762	449	441	438	434	456	525
1	79 246	92 089	91 034	94 391	101 389	82 232	443	434	428	432	452	526
2	99 989	98 169	95 146	98 889	100 874	96 276	421	413	411	423	451	538
5	99 387	102 404	97 815	104 747	98 752	91 517	402	402	400	420	454	543
10	99 032	97 490	107 607	100 630	93 740	96 142	395	395	397	420	452	536
15	101 173	101 463	108 732	104 947	92 412	97 625	393	392	394	414	451	535
20	99 818	101 094	102 533	97 335	99 518	92 283	392	393	391	413	445	536
40	101 095	102 859	104 380	107 968	87 479	95 779	390	386	386	408	436	532
80	100 322	101 333	107 085	98 069	93 602	90 671	381	380	380	404	431	534
160	97 303	103 366	97 990	81 067	86 634	90 735	373	371	373	393	426	530
320	102 883	96 209	98 813	86 151	98 478	98 312	362	367	366	384	421	532
640	105 055	99 948	89 367	89 819	98 366	91 060	358	358	360	378	425	531

MS1-TVM	*Win*						Keskmine tugivektorite arv					
C/γ	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	52 387	94 305	82 745	127 564	136 013	128 942	570	552	544	534	546	602
0.75	83 409	84 671	85 589	123 392	132 626	119 232	549	535	531	523	537	584
1	74 418	87 159	110 500	131 595	132 170	104 279	538	527	521	516	527	573
2	110 140	104 273	109 344	135 556	116 789	95 329	516	510	507	499	505	558
5	103 793	119 080	128 933	119 588	107 849	84 069	501	497	492	468	478	538
10	120 321	128 874	124 917	106 101	102 290	92 021	493	480	469	448	452	536
15	122 384	126 273	114 850	103 809	97 502	97 877	486	469	459	437	442	535
20	123 357	124 180	117 778	99 901	101 714	92 419	477	460	450	425	439	531
40	123 853	117 358	120 910	90 676	89 491	93 035	458	443	430	406	433	532
80	122 970	116 507	116 278	90 428	92 711	90 172	438	422	412	391	424	531
160	119 057	106 583	101 812	86 550	96 792	93 035	419	402	399	385	424	532
320	108 094	109 264	107 956	89 483	94 888	97 328	399	388	380	381	426	531
640	96 479	99 785	96 638	89 571	93 960	91 318	384	373	360	380	424	529

MS2-TVM	*Win*						Keskmine tugivektorite arv					
C/γ	0.005	0.0075	0.01	0.03	0.05	0.1	0.005	0.0075	0.01	0.03	0.05	0.1
0.5	62 280	71 031	83 335	111 371	122 777	134 581	569	551	545	532	548	601
0.75	87 245	85 519	87 655	121 802	135 612	125 703	548	536	529	522	535	582
1	85 254	95 869	99 862	130 887	130 972	119 216	536	526	521	515	527	572
2	88 479	101 962	109 180	126 976	118 529	97 451	516	511	505	497	507	557
5	105 314	114 613	122 361	124 658	104 293	96 167	501	497	491	466	476	541
10	114 400	125 498	121 091	106 579	96 960	93 774	493	481	472	446	452	537
15	122 310	137 239	118 404	109 205	95 742	86 298	486	470	459	437	443	533
20	127 819	119 366	115 461	102 929	88 381	91 940	478	463	451	427	438	532
40	132 370	126 679	111 655	97 065	96 344	96 767	460	442	430	408	433	530
80	125 000	110 972	111 028	95 228	90 345	95 080	441	421	409	396	427	533
160	114 413	106 064	108 446	92 586	93 275	95 838	419	403	397	387	423	531
320	102 130	96 184	87 339	86 765	96 482	90 466	398	388	380	379	425	531
640	107 215	91 407	96 473	92 866	87 895	89 148	385	374	364	382	425	532

Pahaks läinud nõudeid õnnestub müüa 40% väärtuses

S-TVM	*Win*						Keskmine tugivektorite arv					
C\Mu	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**
0.5	12 906	38 833	51 297	63 271	61 382	39 323	450	450	448	445	462	525
0.75	40 578	58 682	66 850	69 698	67 906	48 558	448	443	439	437	456	523
1	52 680	66 996	68 594	69 094	69 948	60 077	443	432	428	429	454	526
2	68 605	68 416	64 802	68 726	76 013	69 037	421	414	412	425	451	537
5	71 109	70 535	69 835	71 857	70 728	60 561	403	400	401	420	455	542
10	70 898	69 028	77 894	73 937	67 441	61 194	395	396	396	418	452	539
15	75 601	70 356	74 641	71 883	67 768	62 474	394	391	395	415	450	535
20	76 353	76 082	71 391	62 664	67 975	65 451	392	391	393	411	447	536
40	71 765	73 190	72 807	58 413	59 416	63 577	388	387	386	406	438	531
80	72 151	69 927	73 940	57 287	58 995	64 657	383	375	377	401	430	532
160	72 354	70 733	66 040	64 368	56 890	66 029	374	368	371	394	422	535
320	67 944	69 914	67 920	63 546	57 692	60 479	363	362	365	385	424	531
640	66 965	64 168	65 209	62 242	60 976	65 651	356	360	359	380	423	530

MS1-TVM	*Win*						Keskmine tugivektorite arv					
C/γ	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**
0.5	27 099	26 585	31 459	62 043	76 729	91 208	556	537	525	518	530	590
0.75	31 329	40 024	50 045	68 707	89 542	80 533	532	516	509	503	520	574
1	32 730	35 640	50 817	73 942	89 358	76 526	518	505	501	499	514	567
2	38 172	42 899	48 674	89 505	79 965	65 942	493	488	485	481	495	555
5	46 434	55 381	69 646	82 019	73 411	61 821	477	472	469	458	472	543
10	57 086	82 803	74 928	73 311	66 401	70 950	471	462	456	441	453	535
15	61 418	76 276	72 867	69 878	60 410	62 748	464	455	444	433	448	533
20	72 195	84 164	78 328	63 960	63 278	63 441	459	447	440	426	442	533
40	85 604	84 283	78 311	64 808	63 653	61 890	446	431	420	405	432	532
80	76 225	74 082	74 142	63 801	61 544	61 945	428	413	404	393	425	530
160	80 778	68 614	71 136	58 186	62 069	67 409	411	397	392	388	426	532
320	83 742	74 440	64 775	60 774	63 306	68 175	392	383	378	381	424	532
640	71 404	66 650	64 041	54 994	58 561	62 218	378	371	360	379	423	531

MS2-TVM	*Win*						Keskmine tugivektorite arv					
C/γ	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**	**0.005**	**0.0075**	**0.01**	**0.03**	**0.05**	**0.1**
0.5	17 487	26 532	30 595	62 574	79 334	91 304	555	537	526	517	529	588
0.75	28 902	35 067	38 473	70 732	92 454	72 783	532	518	510	506	518	575
1	23 671	33 484	56 300	78 427	85 087	74 547	519	506	501	498	512	569
2	35 154	41 549	50 480	87 813	77 124	64 649	494	489	486	481	497	556
5	54 013	60 696	68 660	82 571	68 585	64 390	477	475	471	458	472	541
10	54 931	76 552	87 949	71 599	61 666	62 768	471	464	457	440	456	538
15	64 463	76 338	90 372	66 514	63 456	62 949	464	455	446	434	444	532
20	73 962	81 624	80 743	65 395	67 981	64 868	460	450	439	426	440	530
40	91 077	81 275	78 595	61 560	61 561	60 356	447	430	422	405	431	533
80	82 179	75 178	75 588	67 034	64 064	60 105	427	414	407	397	423	531
160	75 886	77 207	73 592	63 870	63 631	63 400	410	398	389	390	423	532
320	75 836	67 014	63 587	60 123	64 294	66 594	391	385	379	379	421	531
640	72 283	71 468	60 842	61 462	64 596	62 197	379	376	362	380	423	533

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Edwart Ždanovitš,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose Müügikvaliteedi parandamine tugivektormasinate abil, mille juhendaja on Raul Kangro,
   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.

Tartu, **19.01.2017.**