TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Magnus Paal

Andmepunktide panuse visualiseerimine binaarklassifitseerija kaos

Bakalaureusetöö (9 EAP)

Juhendaja: Meelis Kull

Tartu 2021

Andmepunktide panuse visualiseerimine binaarklassifitseerija kaos

Lühikokkuvõte: Masinõppe klassifikaatorite hindamiseks kasutatud kaofunktsioonide väljundi kahanemine on üks viis kuidas hinnata klassifikaatori paranemist. Kaofunktsiooni väljundi ehk kao korral on tegemist ainult ühe arvulise väärtusega ja see ei anna alati head ülevaadet nii klassifikaatorist kui ka andmetest. Töö eesmärk oli uurida, kuidas oleks kadu võimalik tõlgendada andmepunktide kaudu ja leida viis selle visualiseerimiseks. Visualisatsioonid, mis töö käigus leiti ja tutvustati võimaldavad eelkõige uurida, kuidas iga eraldi seisev andmepunkt tervesse kaosse panustab. Selle abil saab uurida milliste andmepunktide panus kaosse on üldiselt suurem, kas nende, mida on rohkem ja millele ennustatud väärtus on lähemal tegelikule väärtusele või nende, mida on vähem ja mille ennustatud väärtus on kaugemal tegelikust väärtusest. Teiseks saab võrrelda klassifikaatoreid ja võrrelda kadude erinevuse korral, millised on andmepunktid, mis on sellesse erinevusse peamised panustajad. Kolmandaks sai eri andmepunkte visualiseeringus eraldada tunnuste põhjal ja uurida, millised neist tunnustest panustavad kaosse kõige rohkem.

Võtmesõnad: masinõpe, binaarne klassifitseerimine, kaofunktsioon, Brier'i skoor, ristentroopia

CERCS: P176 Tehisintellekt

Visualizing the Contribution of Datapoints in a Binary Classifier's Loss

Abstract: The decrease of a classifier's loss function's output is one way to know if a classifier is improving. The output of a loss function which is also known as loss is just one value and doesn't give a complete overview of the classifier and dataset as a whole. The aim of this thesis was to find a way how to interpret loss through datapoints and visualize it. The visualizations found can help to grasp how each datapoint contributes in the whole loss. These visualizations could be used to find out which sets of datapoints contribute the most in loss, the ones whose predicted value is farther from their actual value and which make up a smaller number of points, or those whose predicted value is closer to the actual value and which make up a bigger number of points. Secondly these visualizations could

be used to compare the different losses of two classifiers and find out which datapoints are the ones that contribute most in that difference. Lastly the visualizations could be used to find out which datapoints with which features contribute the most in a loss.

Keywords: machine learning, binary classification, loss function, Brier score, cross entropy

CERCS: P176 Artificial Intelligence

Sisukord

Sissejuhatus

Masinõppes kasutatakse klassifikaatoreid, mis määravad andmepunktidele klasse. Näiteks klassifikaator võib inimese tunnuste puhul pakkuda tema soo. Klassifikaatorite headuse hindamiseks kasutatakse erinevaid meetodeid. Üks selline hindamismeetod on kaofunktsioon (*loss function*). Üks klassifikaator hinnatakse teisest paremaks, kui selle kaofunktsiooni väljund, ehk kadu (*loss*) on väiksem, kui teise klassifikaatori kadu. See tähendab, et kaofunktsiooni väljundi minimeerimine on üks viis, kuidas mudeli headuse paranemist saab hinnata.

Kaofunktsiooni väljund on ainult üks arvuline väärtus ja see ei anna alati head ülevaadet klassifikaatorist ja andmetest. Töö eesmärgiks on uurida, kuidas kaofunktsiooni väljundit saaks paremini tõlgendada andmepunktide kaudu ja välja mõelda visualiseeringud, mille abil saaks kergesti leida vastuse järgmistele küsimustele:

1. Kas suurema osa kaost panustavad andmepunktid, mida on rohkem, aga millele ennustatud väärtus on lähemal tegelikule väärtusele, või pigem andmepunktid, mida on vähem ja millele ennustatud väärtus on kaugemal tegelikust väärtusest?

2. Kui mõne klassifikaatori kadu on suurem kui teise oma, siis millised andmepunktid on sellesse peamised panustajad?

3. Kas mingite konkreetse tunnustega andmepunktid panustavad kaosse ebaproportsionaalselt palju võrreldes teisi tunnuseid omavate andmepunktidega?

Eelnevalt on erinevaid kadusid kujutatud pindaladena kõverate all [3, 6, 8]. Nendeks viisideks on hinnakõverad, mille esialgu pakkusid välja Drummond ja Holte (2006) [3], ning Brier'i kõverad, mis olid täiendatud hinna kõverad, mille pakkusid välja Hernández-Orallo jt. (2011) [6]. Viimast täiendas Kull (2020) [8] võimalusega kujutada ristentroopia kadu kõvera all.

Töö käigus tutvustatakse valminud meetodit, kuidas kujutada iga andmepunkti panust pindala all. Seda tehakse kahe erineva visualisatsiooniga Brier'i kõverate ja tulpdiagrammide abil. Töö vältel kasutatakse krediidiriski andmestikku [7]. Selle andmestiku

klassifitseerimise eesmärk on ennustada, kas inimesele laenu andmisel nad maksavad selle tagasi või mitte.

Esimeses osas seletatakse tööga seonduvaid mõisteid, definitsioone ja tutvustatakse varem ilmunud uurimusi. Töö teises osas tutvustatakse andmepunktide panuse visualiseerimiseks kahte erinevat viisi Brier'i kõverate ja tulpdiagrammi abil. Töö kolmandas uuritakse visualiseeringuid sissejuhatuses püstitatud küsimuste abil. Töö käigus valmis ka Pythoni teek Brier'i ja hinnakõverate joonistamiseks ja andmepunkti panuse visualiseerimiseks. Lisast I on kätte saadav Githubi koodi repositoorium, kus on nii teek, kui ka visualisatsioonide joonistamiseks loodud kood.

1 Mõisted ja definitsioonid

Selles töö osas selgitatakse peamiseid tööga seonduvaid mõisteid ja definitsioone. Samuti antakse ülevaade varem ilmunud uurimuste kohta. See aitab paremini mõista töö järgmiseid osi.

1.1 Binaarne klassifikaator

Klassifikatsioon on masinõppe ülesanne, mille eesmärgiks on kasutada klassifikaatoreid, mis määravad andmepunktidele klasse lõplikust hulgast. Tähistame kõikide andmepunktide hulka kui $\mathcal{X} \times \mathcal{Y}$, mille elemendid on paarid $z_i = (x_i, y_i)$, kus x_i on andmepunktile z_i kuuluvad tunnused ja y_i on andmepunkti tegelik klass.

Kui võtta, et andmepunkt z_i on $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ ja klassifikaator on funktsioon $f(x)$, siis andmepunktile ennustatud klass on $\hat{y}_i = f(x_i)$. Selles töös kasutatakse aga peamiselt binaarset klassifitseerijat, mis tähendab, et kõikide võimalike klasside hulk koosneb ainult kahest võimalikust klassist [5]. Nendeks klassideks on negatiivne (tähistatakse 0) ja positiivne (tähistatakse 1).

Mõne klassifikaatori rakenduse juures on vaja ennustust, mis annab tõenäosuse, et ennustus kuulub mingisse klassi. Tõenäosusliku klassifikatsiooni eesmärk ongi ennustada elementide kuuluvust populatsioonis nii, et binaarse klassifikatsiooni tulemuseks on vektor (p_i^0, p_i^1) andmepunkti z_i igasse klassi kuulumise tõenäosustest [5]. Siin on p_i^0 andmepunktile z_i määratud negatiivsesse klassi kuulumise tõenäosus ja p_i^1 andmepunktile z_i positiivsesse klassi kuulumise tõenäosus. Kuna $p_i^0 + p_i^1 = 1$ ja teades, et $p_i^0 = 1 - p_i^1$, siis võib selle vektori asendada ainult positiivse klassi tõenäosusega $p_i^1 = \hat{p}_i$, mida edaspidi nimetatakse andmepunktile ennustatud tõenäosuseks.

1.2 Segadusmaatriks

Binaarse klassifitseerimise korral on kahte tüüpi ennustusi, mida saab jagada neljaks. Kui ennustatakse positiivsele andmepunktile positiivne klass, siis on tegemist õige-positiivsega (*True Positive* (TP)), kui negatiivne klass siis vale-negatiivsega (*False Negative* (FN)).

Samuti kui ennustatakse negatiivsele andmepunktile negatiivne klass, siis on tegemist õige-negatiivsega (*True Negative* (TN)) ja kui ennustatakse positiivne klass, siis saadakse vale-positiivne (*False Positive* (FP)).

Need ennustuste tüübid on üks viis kuidas hinnata klassifikaatori headust. Seda visualiseeritakse ka tabeli kujul, mida kutsutakse segadusmaatriksiks (*confusion matrix*) (Tabel 1) [5]. Tabeli ülemises päises on andmepunktidele ennustatud väärtused, mida binaarse klassifitseerija korral on kaks: negatiivne ja positiivne. Tabeli vasakpoolses päises on andmepunktide tegelikud väärtused. Tabelis on 2x2 maatriks, mille iga väärtus vastab igat tüüpi ennustusele.

Tabel 1. Segadusmaatriks

	Ennustatud positiivne	Ennustatud negatiivne
Tegelik positiivne	TP	FN
Tegelik negatiivne	FP	TN

Esimeses veerus on õige-negatiivsete ja vale-negatiivsete arv. Teises veerus on vale-positiivsete ja õige-positiivsete arv.

1.3 Hinnatundlik klassifitseerimine

Enamjaolt on klassifitseerimisel eesmärgiks minimeerida valesid ennustusi, ehk vale-negatiivsete ja vale-positiivsete arvu. Mõnel juhul omavad vale-negatiivsed ja vale-positiivsed erinevat kaalu, ehk klassifikaatori jaoks on tähtsam, et neist ühte tüüpi ennustusi oleks vähem kui teisi.

Näiteks kui ennustatakse suure andmestiku puhul inimestel mõne haiguse olemasolu. Selles andmestikus on kaks klassi: terve ja haige (vastavalt negatiivne ja positiivne). Vale-positiivne ennustus tähendab, et inimene on ekslikult ennustatud haigeks. Tagajärjeks on see, et ravitakse inimest, kes pole haige ja kulutatakse aega ja ressursse. Vale-negatiivse puhul, on haige inimene ennustatud terveks ja tagajärjeks on inimese mitte ravimine ja võimalik surm. Seega on klassifikaatori jaoks tähtsam, et mudel ennustaks võimalikult

vähe vale-negatiivseid.

Hinna saab määrata kõikidele ennustuste liikidele. Selles töös on hind ainult vale-negatiivsetel ja vale-positiivsetel ennustustel, mida tähistatakse vastavalt C_{FN} ja C_{FP}. Samamoodi võib tähistada ka õige-negatiivsete ja õige-positiivsete hinnad C_{TN} ja C_{TP}, aga nende väärtus on siin alati 0. Nüüd saab leida iga andmepunkti jaoks võimaliku ehk oodatava hinna. Oodatav hind näitab, milline on võimalik hind andmepunkti negatiivseks või positiivseks ennustamisel.

Kui võtta suvaline andmepunkt z_i, millele ennustatud tõenäosus on $\hat{p}_i$ ja andmepunkti tegelik väärtus pole teada, siis selle andmepunkti negatiivseks ennustamisel on oodatav hind $(1-\hat{p}_i)\cdot C_{TN} + \hat{p}_i \cdot C_{FN} = \hat{p}_i \cdot C_{FN}$ [10] ja andmepunkti positiivseks ennustamisel on oodatav hind $(1-\hat{p}_i)\cdot C_{FP} + \hat{p}_i \cdot C_{TP} = (1-\hat{p}_i)\cdot C_{FP}$ [10]. See klass, mille oodatav hind on kõige väiksem, on see, mille hinnatundliku ennustamise korral peaks valima [10]. Ehk kui $\hat{p}_i \cdot C_{FN} > (1-\hat{p}_i)\cdot C_{FP}$, siis peaks valima positiivse klassi, vastupidisel juhul negatiivse. Eelnevast oodatud hindade võrdlusest saab tuletada lävendi:

$$
\begin{aligned}
\hat{p}_i C_{FN} &> (1-\hat{p}_i)C_{FP} && \mid \div C_{FN} \\
\hat{p}_i &> \frac{C_{FP}}{C_{FN}}(1-\hat{p}_i) \\
\hat{p}_i &> \frac{C_{FP}}{C_{FN}} - \frac{C_{FP}}{C_{FN}}\hat{p}_i \\
\hat{p}_i + \frac{C_{FP}}{C_{FN}}\hat{p}_i &> \frac{C_{FP}}{C_{FN}} \\
(1+\frac{C_{FP}}{C_{FN}})\hat{p}_i &> \frac{C_{FP}}{C_{FN}} \\
\frac{C_{FN}+C_{FP}}{C_{FN}}\hat{p}_i &> \frac{C_{FP}}{C_{FN}} && \mid \div \tfrac{C_{FN}+C_{FP}}{C_{FN}} \\
\hat{p}_i &> \frac{C_{FP}}{C_{FP}+C_{FN}} && (1)
\end{aligned}
$$

Enamjaolt ennustatakse andmepunktile kindel klass tõenäosusliku klassifikaatori kasutamisel mingi lävendi t seadmisel. Ehk kui ennustatud tõenäosus $\hat{p}_i > t$, siis ennustatak-

se sellele punktile positiivne klass, vastupidisel juhul negatiivne [6]. Tuletusest (1) saadud lävendit $\frac{C_{FP}}{C_{FP}+C_{FN}}$ nimetatakse edaspidi hindade suhteks C. Ka Hernández-Orallo jt (2011) kasutavad Brier'i kõverates lävendina hindade suhet $c = \frac{C_{FP}}{B}$, kus $B = C_{FP} + C_{FN}$ [6]. Iga andmepunkti z_i puhul ennustab klassifikaator nüüd selle klassiks

$$\hat{y}_i = \begin{cases} 1 & \text{kui } \hat{p}_i > c \\ 0 & \text{vastasel juhul} \end{cases} \tag{2}$$

[4].

Kui võtta nüüd hinnad $C_{FP} = 1$ ja $C_{FN} = 4$, siis iga vale-negatiivse ennustamine on mudeli kogu hinna jaoks neli korda halvem kui vale-positiivse ennustamine. Seega on mudeli huvides ennustada võimalikult vähe vale-negatiivseid. Selle näite juures määratakse andmepunkt positiivseks, kui klassifikaator ennustab tõenäosuseks $\hat{p}_i > \frac{1}{5} = 0.2$.

1.4 Kaofunktsioonide definitsioon

Vigade hulk ja hindade summa võivad mõlemad olla erinevad kaofunktsiooni väljundid ja nendeks sooritatud tehted kaofunktsioon ise. Kaofunktsioonid on viis, kuidas hinnata kui hästi klassifikaator ennustas. Kaofunktsioonid võrdlevad kui lähedal oli antud andmepunktidele ennustatud klassid punkti tegelikule väärtusele. Mida väiksemad ja lähedasemad on tulemused 0-ile, seda parem on klassifikaator [2]. Seega klassifikaator hinnatakse järjest paremaks kaofunktsiooni väärtuse minimeerimisel.

Kõikide andmepunktide tegelike klasside järjend on $\mathbf{y} = (y_1, y_2, ..., y_n)$ ja kõikide ennustatud tõenäosuste järjend on $\hat{\mathbf{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_n)$. Ühe andmepunkti z_i panustatud kadu on $l(y_i|\hat{p}_i)$, kus andmepunkti tegelik väärtus on y_i ja ennustatud tõenäosus on $\hat{p}_i$. Kui ennustus kaugeneb tegelikust väärtusest, siis $l(y_i|\hat{p}_i)$ kasvab [2]. Vastupidisel juhul, kui ennustus läheneb tegelikule väärtusele, siis $l(y_i|\hat{p}_i)$ kahaneb [2]. Iga andmepunkti jaoks hulgast $\{z_1, z_2, ..., z_n\}$ saab iga kaofunktsiooni kirjutada kui [2]:

$$L(\mathbf{y}|\hat{\mathbf{p}}) = \frac{1}{n}\sum_{i=1}^{n} l(y_i|\hat{p}_i). \tag{3}$$

Kuna kasutatakse peamiselt binaarset klassifitseerijat, siis on kadu ühe andmepunkti jaoks võimalik ümber kirjutada järgmiselt [2]:

$$l(y_i|\hat{p}_i) = (1 - y_i)l(0|\hat{p}_i) + y_i l(1|\hat{p}_i). \tag{4}$$

Eelneva (4) paremaks mõistmiseks saab tuua kaks näidet. Klassifikaator ennustab andmepunktile z_i klassi $\hat{y}_i$ nagu valemis (2). Hindade suhteks, ehk lävendiks on $c = 0.5$. Seega, kui $\hat{p}_i > 0.5$, siis saab ennustatud klassiks $\hat{y}_i = 1$ ja vastupidisel juhul $\hat{y}_i = 0$. Kaofunktsioon tagastab nende näidete korral, mitu viga klassifikaator tegi. Ehk ühe andmepunkti kao $l(y_i|\hat{p}_i)$ võib asendada ühe andmepunkti kaoga $l(y_i|\hat{y}_i)$, kus $l(y_i|\hat{y}_i) = 0$, kui $y_i = \hat{y}_i$ ja $l(y_i|\hat{y}_i) = 1$, kui $y_i \neq \hat{y}_i$.

Kõigepealt on andmepunkti tegelikuks väärtuseks negatiivne $y_1 = 0$. Klassifikaator ennustab andmepunktile tõenäosuse $\hat{p}_1 = 0.2$. Kuna $\hat{p}_1 \leq 0.5$, siis määratakse andmepunktile negatiivne klass $\hat{y}_1 = 0$. Nüüd saab ühe andmepunkti kao summa (4) vasakult poolelt $(1 - 0) \cdot l(0|0) = 0 \cdot 0 = 0$ ja paremalt poolt $0 \cdot l(1|0) = 0 \cdot 1 = 0$. Ennustatud klass ja tegelik väärtus olid samaväärsed ja ka kadu oli $0 + 0 = 0$.

Teiseks näiteks võetakse andmepunkt, mille tegelik väärtus on negatiivne $y_2 = 0$. Andmepunktile ennustatud tõenäosuse $\hat{p}_2 = 0.8$ järgi määratakse sellele hoopis positiivne klass $\hat{y}_2 = 1$. Nüüd saab andmepunkti kao vasaku liidetava väärtuseks $(1-0)l(0|1) = 1 \cdot 1 = 1$ ja parema liidetava väärtuseks $0 \cdot l(1|1) = 0 \cdot 0 = 0$. Ennustatud klass ja tegelik väärtus erinesid ja kadu oli $1 + 0 = 1$.

Viimase näite korral oli tegemist vale-positiivse ennustusega, sest mudel ennustas positiivse, aga tegelik klass oli negatiivne. Võib öelda, et kui parempoolne liidetav oli võrdne 1ga, siis oli ennustus vale-positiivne. Samamoodi vasakpoolse liidetava võrdumisel 1ga on tegemist vale-negatiivse ennustusega. Kui kasutatakse teisi kaofunktsioone ja hindade suhteid, siis vasaku poole korral tekib nullist suurem kadu vale-positiivse korral, kui $\hat{p}_i > c$ ja $y_i = 0$. Seda saab ümber kirjutada kui $I[\hat{p}_i > c, y_i = 0]$, mis kehtivuse korral võrdub 1ga ja muidu 0ga [8]. Parema poole korral tekib nullist suurem kadu vale-negatiivse korral, kui $\hat{p}_i \leq c$ ja $y_i = 1$, ehk $I[\hat{p}_i \leq c, y_i = 1] = 1$ [8].

1.5 Erinevad kaofunktsioonid

Eelneva näite korral oli tegemist kaofunktsiooniga, mille lähenemisel nullile väheneb vigade hulk. Iga vale ennustuse korral lisati kaole 1 ja õigete ennustuste korral 0. Olemas on ka teisi kaofunktsioone, nagu Brier'i skoor, ristentroopia ja keskmine hind.

Brier'i skoor, mille pakkus välja Brier (1950) [1], on üks üldlevinumaid kaofunktsioone. Teisiti nimetatakse seda ka ruutkeskmiseks veaks. Kuna aga tõenäosusliku ennustamise korral kasutatakse peamiselt nimetust Brier'i skoor, siis on ka selles töös kasutusel see termin. Brier'i skoori korral on iga andmepunkti panustatud kadu $l_{BS}(y_i|\hat{p}_i) = (y_i - \hat{p}_i)^2$ ja terve skoori saab kirjutada, kasutades kaofunktsiooni üldkuju (3):

$$BS = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2. \tag{5}$$

Teine väga laialdasti kasutatav kaofunktsioon on ristentroopia ehk log-kadu. Selle iga andmepunkti panustatud kadu on $l_{CE}(y_i|\hat{p}_i) = -y_i \log(\hat{p}_i) - (1-y_i)log(1-\hat{p}_i)$ [2]. Jällegi on võimalik kasutada kaofunktsiooni üldkuju (3), et kirjutada funktsioon üle kõikide andmepunktide:

$$CE = \frac{1}{n}\sum_{i=1}^{n}\big[-y_i \log(\hat{p}_i) - (1-y_i)log(1-\hat{p}_i)\big]. \tag{6}$$

Hinnatundlikku klassifitseerimise korral kasutatakse iga vea kohta erinevat hinda. Seega tuleks kasutada kaofunktsiooni, mis minimeerib hindu. Iga andmepunkti kao leidmiseks tuleks vale-positiivse või vale-negatiivse ennustuste korral lisada vastava vea hind:

$$l(y_i|\hat{p}_i) = I[\hat{p}_i > c, y_i = 0]C_{FP} + I[\hat{p}_i \leq c, y_i = 1]C_{FN}. \tag{7}$$

Kõikide andmepunktide hinna summeerimisel ja keskmise võtmisel saab kaofunktsiooni, mida nimetatakse keskmiseks hinnaks:

$$L(C_{FP}, C_{FN}) = \frac{1}{n}\sum_{i=1}^{n} I[\hat{p}_i > c, y_i = 0]C_{FP} + I[\hat{p}_i \leq C, y_i = 1]C_{FN}. \tag{8}$$

Olukorras, kus hinnad pole teada, ei ole mõistlik minimeerida kaofunktsiooni fikseeritud hindade juures, sest hinnad võivad päris maailmas olla teistsugused [8]. Seetõttu ei an-

na lõplik klassifikaator nii häid tulemusi kui kadu on minimeeritud teistsuguste hindade põhjal. Tuleks leida viis, kuidas minimeerida kaofunktsiooni üle kõikide võimalike hindade suhete.

Seda saab teha võttes kaofunktsioonist keskväärtuse üle hindade suhte ühtlase jaotuse lõigul (0,1) [2, 8]. Selleks tuleks keskmise hinna kaofunktsiooni kujutada ka ühe ainsa muutujaga hindade suhtega C. Hindade suhte saab kirjutada kahte erinevat viisi $C = \frac{C_{FP}}{B}$ ja $1 - c = \frac{C_{FN}}{B}$, kus $B = C_{FP} + C_{FN}$. Siis saab hinnad kujutada läbi hindade suhete järgmiselt: $C_{FP} = B \cdot C$ ja $C_{FN} = B \cdot (1 - C)$. Nende kujutiste abil saab keskmise hinna korral asendada hinnad hindade suhtega. Saab kaofunktsiooni, mida kutsutakse edaspidi keskmise hinna keskväärtuseks [8]:

$$E\big[L(C_{FP}, C_{FN})\big] = E\big[L(B \cdot C, B \cdot (1 - C)\big] = E\big[B \cdot L(C, 1 - C)\big], \tag{9}$$

kus keskmise hinna kaofunktsioon on lahti kirjutatult [8]:

$$B \cdot L(C, 1 - C) = B \cdot \frac{1}{n} \sum_{i=1}^{n} I[\hat{p}_i > C, y_i = 0]C + I[\hat{p}_i \leq C, y_i = 1](1 - C). \tag{10}$$

1.6 Range korralikkus

Kui kaofunktsiooni keskväärtuse minimeerimisel minimaalseim väärtus vastab olukorrale $\hat{p}_i = p_i$, ehk ennustatud tõenäosus vastab samade tunnustega andmepunktide tegelikule tõenäosusele, siis kutsutakse seda korralikuks kaofunktsiooniks (*proper scoring rule*) [9]. Rangelt korralikuks kaofunktsiooniks (*strictly proper scoring rule*) kutsutakse juhtu kui minimaalseim väärtus saadakse siis ja ainult siis kui andmepunkti ennustatud tõenäosus on võrdne samade tunnustega andmepunktide tegeliku tõenäosusega $\hat{p}_i = p_i$ [9]. See tähendab, et mitte range korralikkuse puhul võib minimaalseima väärtuseni jõuda ka mõnel muul juhul, range korralikkuse korral ainult eelnevalt nimetatud olukorras. Buja jt. (2005) tõestatu põhjal saab näidata, et keskmise hinna keskväärtus (9) on rangelt korralik kaofunktsioon, sest on võimalik jõuda järgmise integreerimiseni [8]:

$$E\big[B \cdot L(C, 1 - C)\big] =$$
$$E\Big[E\big[B \cdot L(C, 1 - C)\big|C\big]\Big] =$$

$$E\Big[E\big[B|C\big] \cdot L(C, 1-C)\Big] =$$
$$\int_0^1 \Big[E\big[B|C=c\big] \cdot f_c(c) \cdot L(c, 1-c)\Big]dc =$$
$$\int_0^1 w(c)L(c, 1-c)dc. \tag{11}$$

Buja jt. (2005) tõestatu põhjal defineerib rangelt korraliku kaofunktsiooni integraal keskmise hinna kaost ja igast positiivsest kaalust $w(c)$ [2]. Täpselt seda kujutab ka eelnev tulemus (11). Seega on $E\big[L(C_{FP}, C_{FN})\big]$ rangelt korralik kaofunktsioon iga positiivse $w(c)$ korral. Ehk kui võtta erinevaid positiivseid funktsioone $w(c)$ on võimalik jõuda kõikide erinevate rangelt korralikkude kaofunktsioonideni. Siin töös mainitud kaofunktsioonidest on mõlemad Brier'i skoor ja ristentroopia rangelt korralikud kaofunktsioonid, millele vastavad kaalud on $w(c) = 2$ ja $w(c) = \frac{1}{c(1-c)}$. Brier'i skoori puhul saab kaalu seda näidata järgneva tuletusega [6, 8]:

$$\int_0^1 2L(c, 1c)dc =$$
$$\int_0^1 \Big[2 \cdot \frac{1}{n}\sum_{i=1}^{n} I[\hat{p}_i > c, y_i = 0]c + I[\hat{p}_i \leq c, y_i = 1](1-c)\Big]dc =$$
$$\frac{2}{n}\sum_{i=1}^{n}\Big[\int_0^1 I[\hat{p}_i > c, y_i = 0]c\,dc + \int_0^1 I[\hat{p}_i \leq c, y_i = 1]c\,dc\Big] =$$
$$\frac{2}{n}\sum_{i=1}^{n}\Big[\int_0^{\hat{p}_i} I[y_i = 0]c\,dc + \int_{\hat{p}_i}^1 I[y_i = 1]c\,dc\Big] =$$
$$\frac{2}{n}\sum_{i=1}^{n}\Big[I[y_i = 0]\frac{\hat{p}_i^2}{2} + I[y_i = 1]\frac{(1-\hat{p}_i)^2}{2}\Big] =$$
$$\frac{1}{n}\sum_{i=1}^{n}\Big[I[y_i = 0]\hat{p}_i^2 + I[y_i = 1](1-\hat{p}_i)^2\Big] =$$
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}_i)^2. \tag{12}$$

1.7 Brier'i kõverad

Brier'i kõverad põhinevad Drummond ja Holte (2006) välja pakutud hinnakõveratel. Hinnakõverates erinesid hindade suhe c ja klassifikaatori lävend t, millega määrati ennustatud tõenäosuse põhjal andmepunktile klass [3]. Hernández-Orallo jt. (2013) välja pakutud Brier'i kõverates määrati lävend t samaväärseks hindade suhtega c [6].

Üldiselt on teada, et integraali võtmisel a-st b-ni funktsioonist leitakse selle funktsiooni joone ja x-telje vaheline pindala, mis on piiratud sirgetega vasakult poolt $x = a$ ja paremalt poolt $x = b$. Seega eelneva põhjal (11) tähendab see, et keskmise hinna keskväärtust on võimalik kujutada pindalana funktsiooni $w(c)L(c, 1-c)$ all. Kui võtta, et $w(c) = 2$, siis saab Brier'i kõverad, mille on välja pakkunud Hernández-Orallo jt. (2013) [6]. Autorid tõestasid ka, et Brier'i kõverate alune pindala on võrdne Brier'i skooriga [6]. Varasemas tuletuses (12) on ka näidatud, kuidas $w(c)L(c, 1-c)$ alusest pindalast Brier'i skoor tuleb [6, 8].

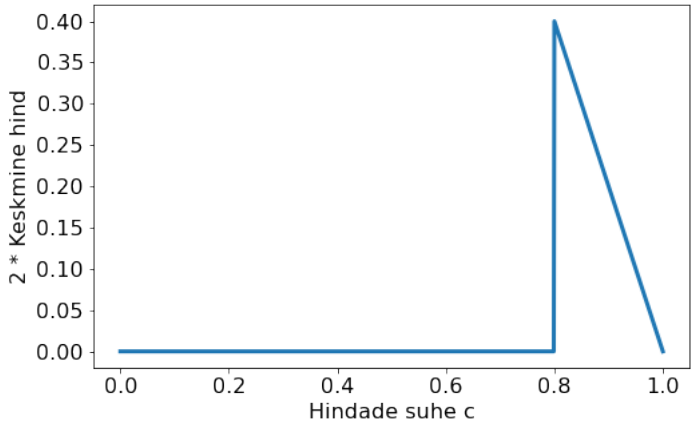
Võttes nüüd funktsiooni

$$2L(c, 1-c) = \frac{2}{n} \sum_{i=1}^{n} I[\hat{p}_i > c, y_i = 0]c + I[\hat{p}_i \leq c, y_i = 1](1-c), \tag{13}$$

võib proovida joonistada seda kõverat üle kõikide hindade suhete c, et näha milline selle alune pindala välja näeb ja kuidas see tekib. Saab võtta ühe positiivse andmepunkti, mis tähendab, et selle väärtus on $y_i = 1$. Kui klassifikaator ennustab andmepunktile tõenäosuseks $\hat{p}_i = 0.8$, siis peab vaatama keskmise hinna (13) summa aluse liitmise paremat poolt $I[\hat{p}_i \leq c, y_i = 1](1-c)$, sest $y_i = 1$. Nüüd käies läbi kõik hindade suhted 0-ist 1-ni väärtustub $\hat{p}_i \leq c$ tõeseks siis, kui jõutakse väärtuseni 0.8 (Joonis 1). Selle väärtuse juures on $2L(c, 1-c) = 2(1-0.8) = 0.4$. Nüüd korratakse sama kuni $c = 1$.

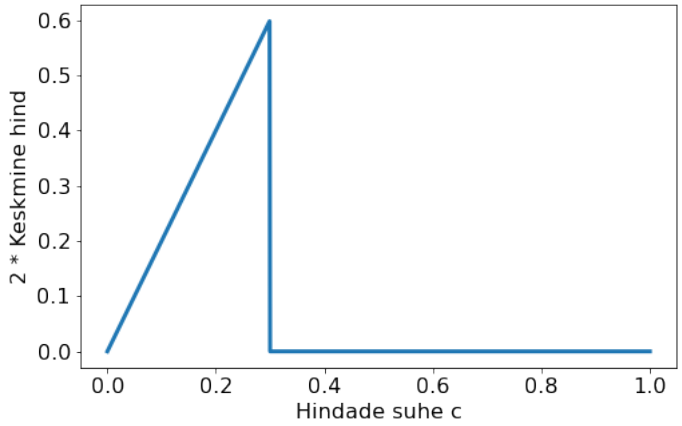
Kuna hindade suhte piirkonnas $c < 0.8$ on $\hat{p}_i > c$, siis on ennustatud klass positiivne, ehk samaväärne tegeliku klassiga. Sellepärast puudub joonisel 1 selles piirkonnas kõvera alune pindala, ehk pole kadu. Kui aga $c > 0.8$, siis $\hat{p}_i < c$ ja joonisel on selles piirkonnas kõvera all pindala, ehk kaol on nullist suurem väärtus. On näha, et joonisel 1 olev kõver on kolmnurk. Selle pindala arvutamiseks saab kasutada lihtsalt kolmnurga pindala valemit, kus kõrgus on $2L(c, 1-c) = 2(1-0.8) = 0.4$ ja alus on $1 - \hat{p}_i = 1 - 0.8 = 0.2$. Seega kolmnurga pindalaks saab $\frac{0.4 \cdot 0.2}{2} = 0.04$. Ka eelnevast tuletatud Brier'i skoorist (12) on näha, et selle kõvera alune pindala on $BS = (1-\hat{p}_i)^2 = (1-0.8)^2 = 0.04$.

Nüüd kui võtta andmepunktiks negatiivne punkt ennustatud tõenäosusega $\hat{p}_i = 0.3$, siis



Joonis 1. Brier'i kõver ühe positiivse andmepunktiga ja tõenäosusega $\hat{p}_i = 0.8$.

tuleb vaadata keskmise hinna summa aluse liitmise vasakut poolt $I[\hat{p}_i > c, y_i = 0]c$, sest $y_i = 0$. Kui alustada hindade suhete läbimist 0-st 1-ni väärtustub $\hat{p}_i > c$ tõeseks kuni jõutakse väärtuseni $c = 0.3$ (Joonis 2). Selles punktis on $2L(c, 1-c) = 2 \cdot 0.3 = 0.6$. Jällegi tulemuse (12) põhjal võib öelda, et kõvera alune pindala on $BS = (0 - 0.3)^2 = 0.09$.

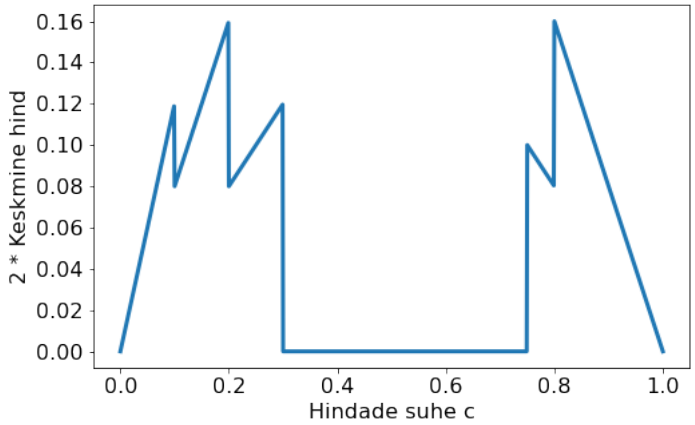


Joonis 2. Brier'i kõver ühe negatiivse andmepunktiga ja tõenäosusega $\hat{p}_i = 0.2$.

Sama saab teha ka mitme andmepunktiga. Viieks andmepunktiks võib võtta $z_1 = (x_1, y_1)$, $z_2 = (x_2, y_2)$, $z_3 = (x_3, y_3)$, $z_4 = (x_4, y_4)$ ja $z_5 = (x_5, y_5)$, kus nende tegelikud väärtused on $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$, ning neile suvalise klassifikaatori poolt ennustatud tõenäosused on $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$, $\hat{p}_4 = 0.75$ ja $\hat{p}_5 = 0.8$. Saab joo-

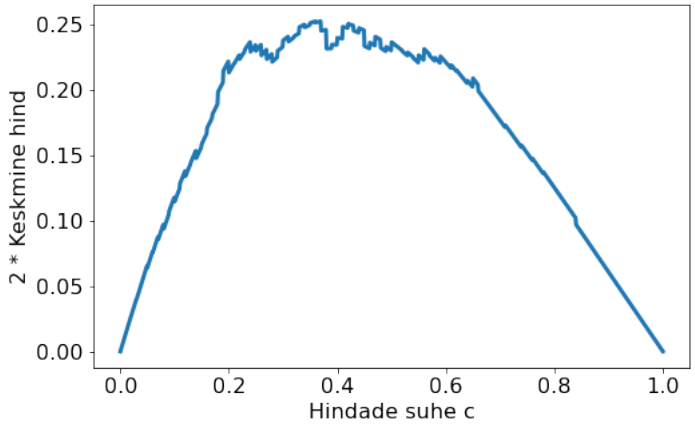
nistada keskmise hinna funktsiooni kõvera üle kõikide hindade suhete neid andmepunkte arvestades (Joonis 3).

Joonisel tuleb üks punkt, ehk ühele hindade suhtele keskmine hind jällegi eelnevast valemist (13). Seega näiteks hindade suhte $c = 0.25$ juures arvutatakse keskmise hinna sees olev summa üle iga andmepunkti $(0 \cdot 0.25 + 0 \cdot 0.25) + (0 \cdot 0.25 + 0 \cdot 0.25) + (1 \cdot 0.25 + 0 \cdot 0.25) + (0 \cdot 0.25 + 0 \cdot 0.25) + (0 \cdot 0.25 + 0 \cdot 0.25) = 0.25$. Lõpliku tulemuse, ehk kahekordse keskmise hinna saamiseks tuleb tulemus korrutada kaaluga $w(c) = 2$ ja jagada kõikide andmepunktide arvuga $n = 5$, mille järel on tulemuseks $\frac{2}{5} \cdot 0.25 = 0.1$. Kasutades Brier'i skoori (12), saab leida selle klassifikaatori kao, ehk kõvera aluse pindala, nendel andmetel: $BS = \frac{1}{5}\left[(0-0.1)^2 + (0-0.2)^2 + (0-0.3)^2 + (1-0.75)^2 + (1-0.8)^2\right] = 0.0485$.



Joonis 3. Brier'i kõver andmepunktide tegelike väärtustega $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$ ning ennustatud tõenäosustega $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$, $\hat{p}_4 = 0.75$ ja $\hat{p}_5 = 0.8$. Kõvera alune pindala ja klassifikaatori Brier'i skoor $BS = 0.0485$ on võrdsed.

Töös on kasutatud krediidiriski andmestikku [7]. Selle andmestiku klassifitseerimise eesmärgiks on leida, kas pangal on inimesele rahaliselt kasulik laenu anda või mitte. Igal inimesel on 20 tunnust, mille kohta on ülevaade UCI Machine Learning Repository veebilehel [7]. Negatiivsed andmepunktid on inimesed, kellele võiks laenu anda ja positiivsed on inimesed, kellele ei peaks. Andmestikkus on 1000 andmepunkti, nendest 700 kasutati scikit-learni otsustusmetsa ja naiivse Bayes'i klassifikaatorite õpetamiseks. Ülejäänud 300 andmepunkti kasutatakse ülejäänud töös krediidiriski andmestiku visualisatsioonidel mõlema klassifikaatori hindamiseks.



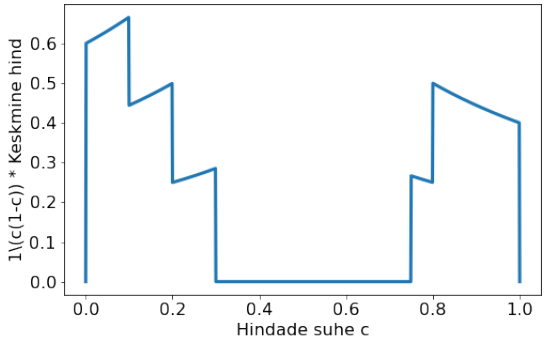
Joonis 4. Brier'i kõver otsustusmetsa klassifikaatori ennustatud tõenäosustega krediidiriski andmestikus. Kõvera alune pindala on Brier'i skoor $BS = 0.1641$.

Brier'i kõverat saab joonistada ka krediidiriski andmetel kaasates veel rohkem andmepunkte (Joonis 4). Nüüd kasutatakse mitte ettemääratud tõenäosuseid, vaid otsustusmetsa klassifikaatori poolt igale andmepunktile ennustatud tõenäosust.

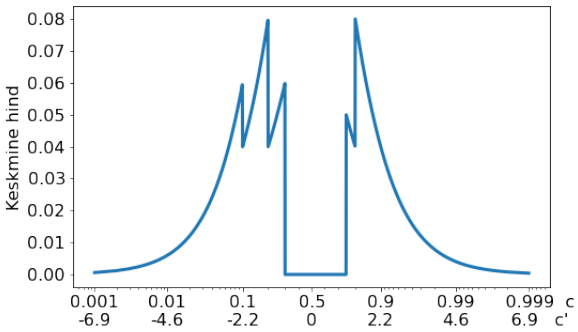
Joonisel 4 kujutatud Brier'i kõvera alune pindala on seega krediidiriski andmetel otsustusmetsa klassifikaatori Brier'i skoor $BS = 0.1641$. Seega keskmise hinna kaofunktsiooni joonistamisel kaaluga $w(c) = 2$ üle kõikide hindade suhete c saab kõvera, mille alune pindala on Brier'i skoor . Kui aga hindade suhe on $c = 0.5$ ja kaal on $w(c) = 1$, siis on ka jooniselt võimalik välja lugeda klassifikaatori veamäär (*error rate*), mis vastab selle hindade suhte juures keskmisele hinnale. Sellisel juhul saab kätte ka klassifikaatori täpsuse (*accuracy*), kuna see on samaväärne lahutusega $1 -$ veamäär.

1.8 Ristentroopia kujutamine Brier'i kõvera aluse pindalana

Brier'i kõveratega oli näha, et kõvera alla tekkinud pindala oli Brier'i skoor. Kõverate all saab kujutada ka teist kaofunktsiooni väärtust, milleks on ristentroopia. Kui kasutada kaalufunktsioonina $w(c) = \frac{1}{c(1-c)}$, siis on kõvera alune pindala ristentroopia. Saadud kõver on väga moonutatud (Joonis 5) ja seda on keeruline võrrelda kõveraga, mille alune pindala oli Brier'i skoor [8].



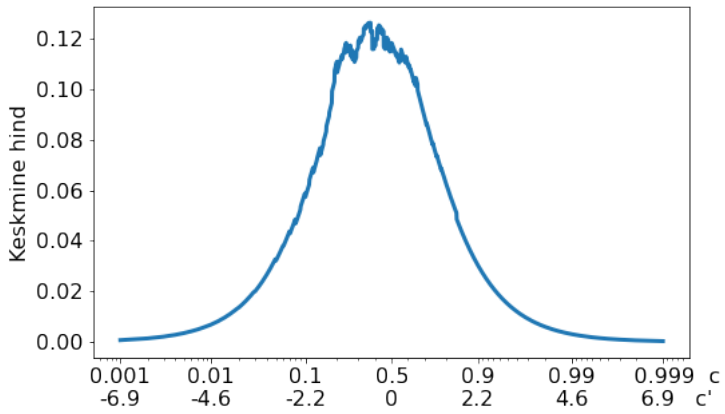
Joonis 5. Moonutatud ristentroopia kõver andmepunktide tegelike väärtuste $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$ ja ennustatud tõenäosustega $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$. Pindala on võrdne ristentroopiaga $CE = 0.2392$.



Joonis 6. Ristentroopia kõver andmepunktide tegelike väärtuste $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$ ja ennustatud tõenäosustega $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$. Pindala on võrdne klassifikaatori ristentroopiaga $CE = 0.2392$.

Ristentroopia kujutamiseks alternatiivse viisi pakub välja Kull (2020) [8], kus Brier'i kõvera, mille kaaluks on $w(c) = 1$, x-telje suunalisel venitamisel saab kõvera aluseks pindalaks ristentroopia [8]. Nimelt peaks x-telge kujutama logitina, kus iga hindade suhte uueks väärtuseks saab $c' = \log \frac{c}{1-c}$ [8]. See tähendab seda, et mida kaugemal on hindade suhte väärtused väärtusest 0.5, seda suuremad on x-teljel olevate väärtusete vahelised vahed. Kasutades joonisel 3 kasutatud andmeid tegelike väärtustega $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$ ja ennustatud tõenäosustega $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$, $\hat{p}_4 = 0.75$ ja $\hat{p}_5 = 0.8$ on võimalik kujutada mudeli ristentroopiat nendel andmetel, mida on näha joonisel 6. Nii joonise 5 kui ka joonise 6 kõverate alune pindala on võrdne ristentroopiaga $CE = 0.2392$, mille leidmiseks on varasemalt tuttav funktsioon (6).

Ristentroopia kõver võib anda hea ettekujutuse ka sama klassifikaatori Brier'i kõverast ja selleläbi ka Brier'i skoorist [8]. Joonisel 6 kujutatud ristentroopia kõveras ja joonisel 3 kujutatud Brier'i kõveras on kasutatud samu andmepunkte ja neile ennustatud tõenäosuseid. Ristentroopia kõveralt saab enamvähem ette kujutada, milline võiks olla Brier'i kõver ja selleläbi Brier'i skoor [8]. Moonutatud ristentroopia kõveralt (Joonis 5) on seda keerulisem teha. Lisaks Brier'i skoorile annab ristentroopia kõvera alune pindala ülevaate ka ristentroopiast.



Joonis 7. Ristentroopia kõver krediidiriski andmete otsustusmetsa klassifikaatoriga. Kõvera alune pindala on võrdne ristentroopiaga $CE = 0.49562$.

Krediidiriski andmetega otsustusmetsa klassifikaatori ristentroopiat on samuti võimalik kujutada pindalana. Selle otsustusmetsa klassifikaatori ristentroopia krediidiriski andmetel on $CE = 0.49562$. Sama ristentroopia on pindalana kujutatud joonisel 7. Seda kõverat kutsutakse ristentroopia kõveraks [8]. Seega on nüüd kõverate all võimalik kujutada kahte erinevat üldlevinud kadu. Nendeks on ristentroopia ja Brier'i skoor.

2 Andmepunktide panuse visualiseerimine

Töö eesmärgiks oli visualiseerida andmepunktide panust kaos. Selles töö osas tutuvustatakse selleks kahte erinevat viisi Brier'i kõverate ja tulpdiagrammi abil.

2.1 Kasutatud vahendid

Visualiseeringute tegemiseks kasutati Pythonit. Joonistamiseks kasutati teeki matplotlib ja erinevate matemaatiliste toimingute ja andmete käsitlemise jaoks NumPy, Scipy ja Pandas teeke. Otsustusmetsa ja naiivse Bayes'i mudelid said tehtud teegi scikit-learn abiga. Selleks, et visualiseeringuid oleks võimalik joonistada ka teistel andmetel, sai töö käigus valmis dokumenteeritud teek, mille repositoorium on kätte saadav lisast I. Installeerimise juhend ja kasutusjuhend on repositooriumi README.md failis. Repositooriumis on ka Jupyteri Notebookid, milles on töös kasutatud visualisatsioonide joonistamiseks kirjutatud kood.

2.2 Visualiseerimine Brier'i kõveratega

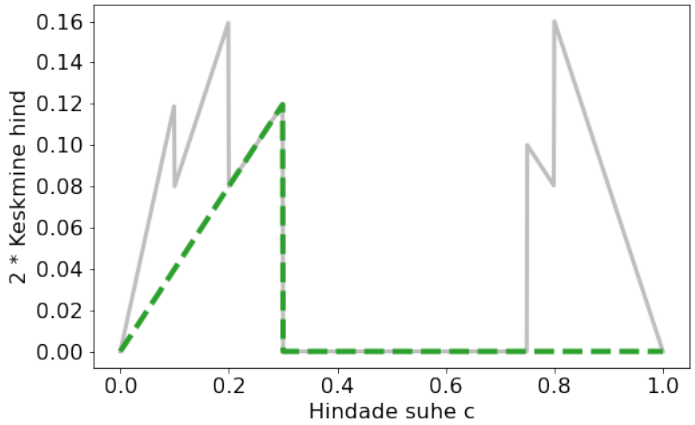
Võttes joonisel 3 kasutatud andmed saab alustada uurimist, kuidas iga andmepunkt Brier'i joone alusesse pindalasse panustab. Viieks andmepunktiks olid z_1, z_2, z_3, z_4, z_5, kus nende tegelikud väärtused olid $y_1 = 0$, $y_2 = 0$, $y_3 = 0$, $y_4 = 1$ ja $y_5 = 1$, ning neile suvalise klassifikaatori poolt ennustatud tõenäosused olid $\hat{p}_1 = 0.1$, $\hat{p}_2 = 0.2$, $\hat{p}_3 = 0.3$, $\hat{p}_4 = 0.75$ ja $\hat{p}_5 = 0.8$.

Kõigepealt peaks joonistama igale andmepunktile kõvera, et näha, milline on nende panustatud kadu terves andmestikus. Selleks tuleks vaadata uuesti keskmise hinna funktsiooni, millega Brier'i kõverad joonistatakse. Seal võetakse keskmine hind üle kõikide andmepunktide. Selleks, et leida ühele andmepunktile vastav kõver, tuleks keskmine hind leida ühe andmepunkti jaoks. Seda saab teha, kui eemaldada kahekordse keskmise hinna, ehk Brier'i kõvera joonistamiseks kasutatavast funktsioonist (13) summa. Seejärel peaks kasutama ühte vaadeldavat andmepunkti z_i ja sellele vastavat väärtust y_i, ning joonistama

järgmise funktsiooni abil (14) kõvera üle kõikide hindade suhete.

$$\frac{2}{n}\sum_{i=1}^{n}\Big[I[\hat{p}_i > c, y_i = 0]c + I[\hat{p}_i \leq c, y_i = 1](1-c)\Big] \rightarrow \frac{2}{n}\Big[I[\hat{p}_i > c, y_i = 0]c + I[\hat{p}_i \leq c, y_i = 1](1-c)\Big] \tag{14}$$

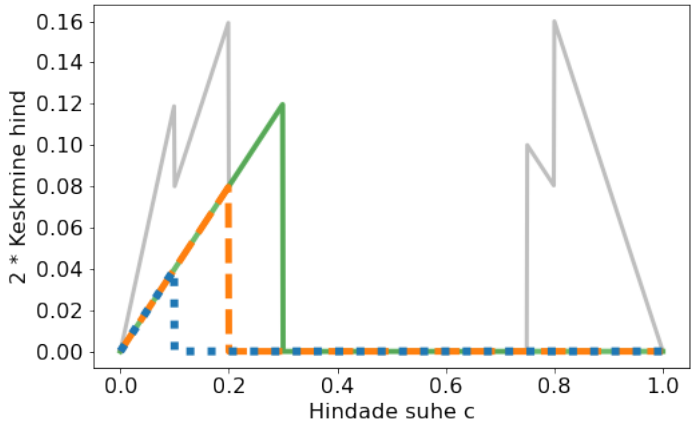
Kui andmetest võtta andmepunkt z_3, siis sellele vastav tõenäosus on $p(x_3) = 0.3$ ja tegelik väärtus $y_3 = 0$. Joonisel 8 on eelmise tulemuse (14) põhjal joonistatud üle kõikide hindade suhete c, kõikides andmetes andmepunktile z_3 vastav kõver ja selle all olev pindala. Halli pidevjoonega kõver on varasemalt tuttav kõikidel andmetel joonistatud Brier'i kõver (Joonis 3). Roheline kriipsjoonega kõver on aga andmepunktile z_3 vastav kõver kõikides nendes andmetes. Selle kõvera alune pindala on terves kaos selle andmepunkti panustatud kadu.



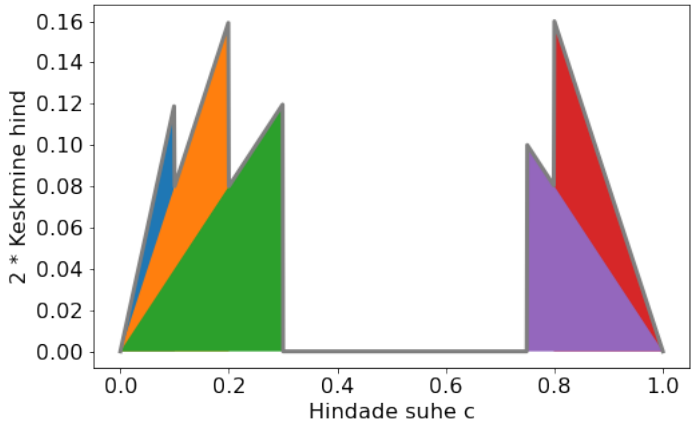
Joonis 8. Andmepunkti z_3 Brier'i kõver (roheline, kriipsjoon) kõikides andmetes, mille alune pindala on selle andmepunkti panus terves kaos. Kõikide andmepunktide (z_1, z_2, z_3, z_4 ja z_5) Brier'i kõver on kujutatud halli pidevjoonega.

Sama saab teha ka teiste andmepunktide jaoks. Joonisel 9 on näha ka andmepunktidele z_1 ja z_2 vastavad kõverad. Nüüd on lisatud oranž kriipsjoonega kõver, mis vastab andmepunktile z_2 ja sinine punktiirjoonega kõver, mis vastab andmepunktile z_1. Jooniselt on näha, et vastavad pindalad moodustuvad hallis pidevas Brier'i kõveras üksteise otsa. Joonisel 10 on need iga andmepunkti kõvera alused pindalad paigutatud üksteise otsa. Nüüd on roheline pindala andmepunkti z_3 Brier'i skoori panustatud kadu, oranž pindala andme-

punkti z_2 panustatud kadu ja sinine pindala andmepunkti z_1 panustatud kadu. Joonisele on lisatud ka positiivsete andmepunktide panustatud kadu. Sellel joonisel on värvitud pindaladena kujutatud terves andmestikus iga andmepunkti panus Brier'i skoori.



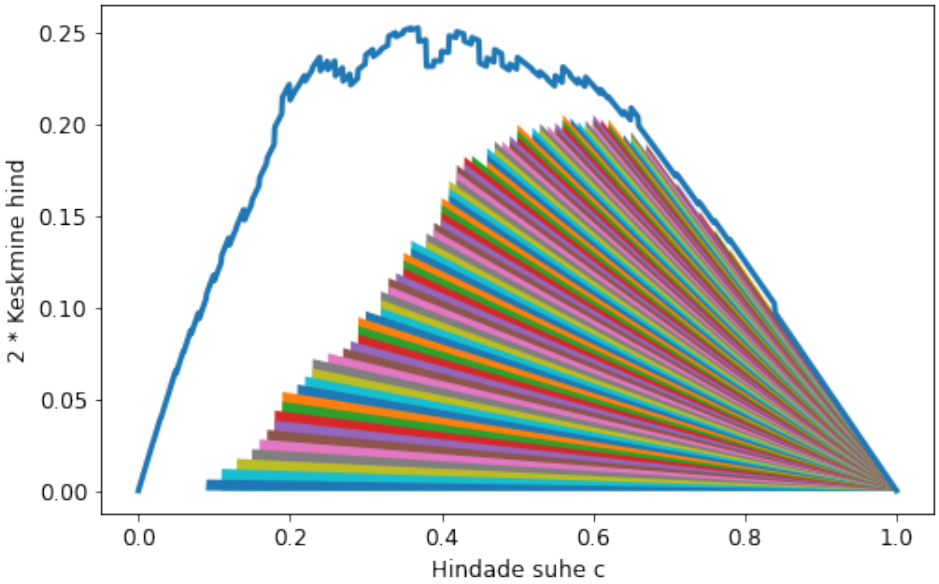
Joonis 9. Andmepunktide z_1 (sinine, punktiirjoon), z_2 (oranž, kriipsjoon) ja z_3 (roheline, pidevjoon) Brier'i kõverad, mille alused pindalad on nende andmepunktide panused terves kaos.



Joonis 10. Kõikide andmepunktide z_1 (sinine), z_2 (oranž) ja z_3 (roheline), z_4 (lilla), z_5 (punane) pindalad, mis on võrdsed nende andmepunktide panustatud kaoga.

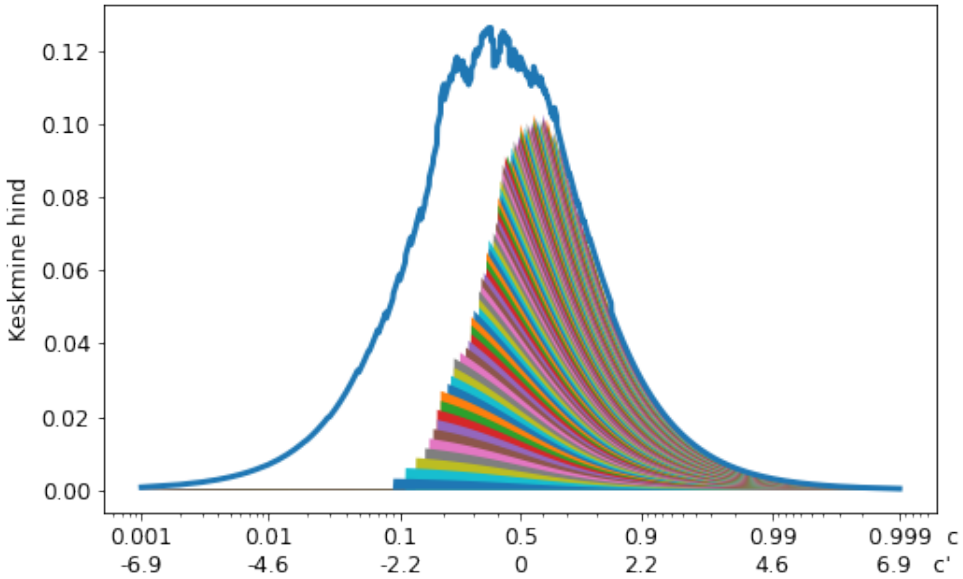
Joonisel 11 on kasutatud krediidiriski andmestikku. Andmepunktidele tõenäosuse ennustamiseks on kasutatud otsustusmetsa algoritmi. Sinise joone alune pindala on selle mudeli Brier'i skoor krediidiriski andmestikul. Värviliselt on välja joonistatud ainult tegelikult

negatiivsete andmepunktide pindalad. Ülejäänud valge pindala sinise kõvera all on tegelikult negatiivsete andmepunktide panustatud, mida sellel joonisel pole eraldi kujutatud.



Joonis 11. Kõikide positiivsete andmepunktide Brier'i skoor kujutatud pindaladena krediidiriski andmestikus, kus iga eraldi värvitud ala on andmepunktide eristamiseks.

Joonisel 12 on Brier'i kõvera asemel kasutatud ristentroopia kõverat krediidiriski andmepunktide kujutamiseks. Seal on iga värviline pindala iga tegelikult negatiivse andmepunkti panustatud ristentroopia. Nagu alampeatükis 1.8 on kirjeldatud, siis peaks ristentroopia kõvera joonistamisel x-telge kujutama logitina, kus iga hindade suhte uueks väärtuseks saab $c' = \log \frac{c}{1-c}$. Selleks tuli kõigepealt joonistada Brier'i kõver ja kaalu $w(c) = 2$ asemel kasutama kaalu $w(c) = 1$. Siis sai joonistatud matplotlib joonisel *plt* kasutada funktsiooni *plt.xscale("logit")*, mis kujutas x-telje logitina. Täpselt sama lahendust rakendati ka iga andmepunkti pindala joonistamisel.



Joonis 12. Kõikide positiivsete andmepunktide ristentroopia kujutatud ristentroopia pindaladena krediidiriski andmestikus.

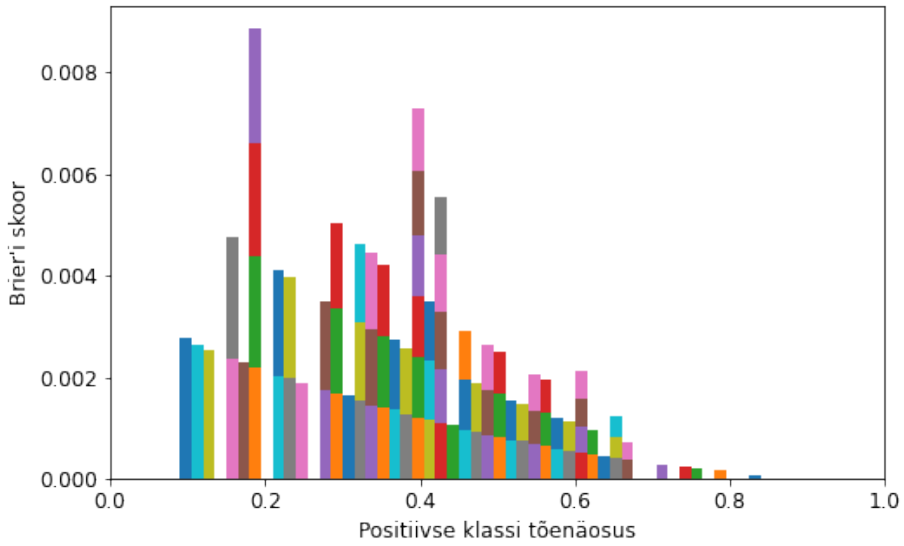
Andmepunktide pindalade eraldi välja joonistamine Brier'i ja ristentroopia kõvera all on üks viis, kuidas tõlgendada kaofunktsioonide väärtuseid andmepunktide kaudu. Võimalik on välja lugeda visuaalselt kui suure osakaalu moodustab tervest kaost iga andmepunkt, aga pole võimalik välja lugeda arvuliselt palju iga andmepunkt kaosse panustab. Selleks vaatleme teist viisi, kuidas andmepunktide kadu kujutada.

2.3 Visualiseerimine tulpdiagrammidega

Teine viis kuidas kujutada iga andmepunkti kadu on joonistada see tulpdiagrammile (Joonis 13). See kujutis aitab mõista ka Brier'i kõverate abil kujutatud iga andmepunkti pindala.

Joonise x-teljel on positiivse klassi tõenäosus ja y-teljel on Brier'i skoor. Iga tulba kõrgus vastab selles tõenäosuse piirkonnas olevate andmepunktide panustatud Brier'i skoorile. Iga tulp on omakorda jaotatud erivärvides tulpadeks, mis vastavad eri andmepunktidele. Joonisel 13 on kasutatud samu krediidiriski andmeid, mis joonisel 11. Iga värv mõlemal joonisel vastab ka samadele andmepunktidele. Ehk kui võtta jooniselt 11 kõige vasak-

poolsem sinine andmepunkt, millele ennustatud tõenäosus on umbes 0.1, siis saab leida ka jooniselt 13 selle tõenäosuse juures lõppeva sinise pindala.



Joonis 13. Kõikide tegelikult positiivsete andmepunktide panustatud Brier'i skoor.

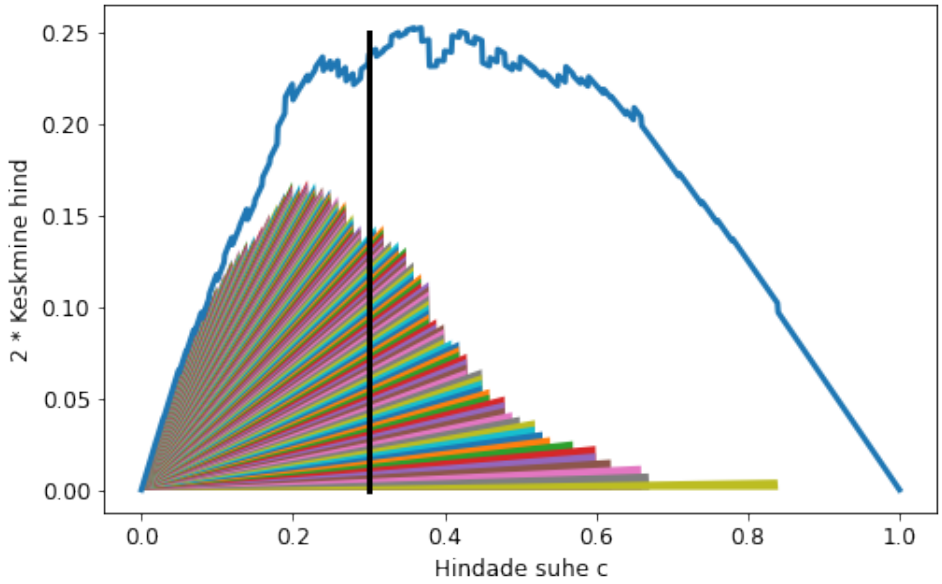
Jooniselt 11 on seega võimalik paremini välja lugeda selle andmepunkti panustatud kao osakaalu terves kaos. Joonis 13 täiendab eelnevat joonist ja sealt saab välja lugeda täpselt kui suur on numbriliselt selle andmepunkti panustatud kadu.

3 Visualisatsioonide uurimine

Selles töö osas uuritakse eelnevat tutvustatud visualiseeringuid sissejuhatuses püstitatud küsimuste abil.

3.1 Andmepunktide võrdlemine ennustatud täpsuse järgi

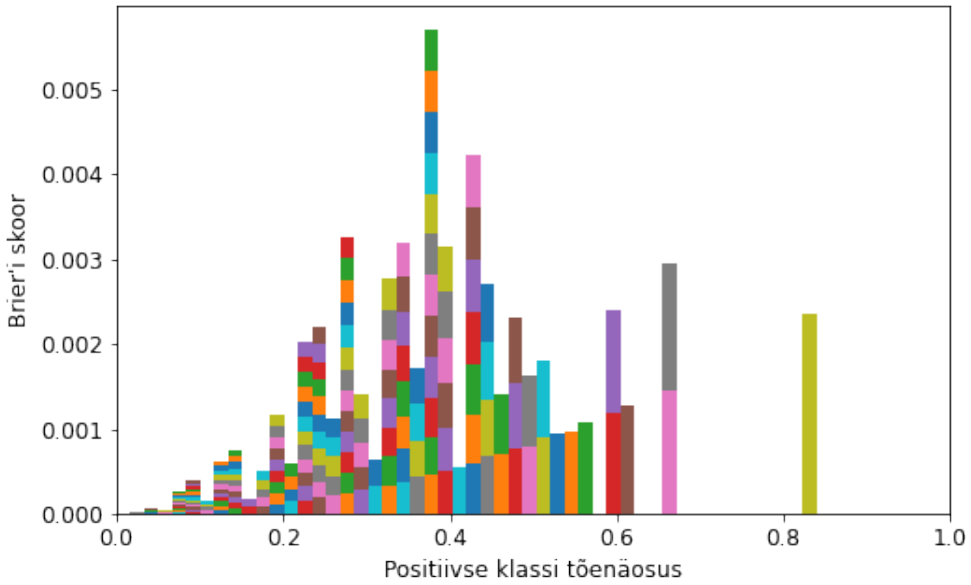
Visualisatsioonide abil on võimalik üldiselt uurida, kuidas iga andmepunkt kaosse panustab. Brier'i kõverate iga andmepunkti panusest ja tulpdiagrammidelt on võimalik välja lugeda, kui palju panustavad üldiselt tegelikule väärtusele lähedale ennustatud andmepunktid kaosse ja kui palju tegelikust väärtusest kaugel olevad punktid. Järgnevatel joonistel on kasutatud ainult krediidiriski andmestikku [7], millel on treenitud otsustusmetsa klassifikaator. Kui vaadata joonist 14, kus on kujutatud iga tegelikult negatiivse andmepunkti panus Brier'i skoori, siis kolmnurgad või andmepunktid, mis lõppevad lähemale hindade suhtele 0 on paremini ennustatud, kui need, mis on sellest kaugemal.



Joonis 14. Kõikide negatiivsete andmepunktide Brier'i skoor kujutatud pindaladena krediidiriski andmestikus.

Andmepunktid, mis on hindade suhtest $c = 0.3$ vasakul pool (mille pindalad lõppevad pärast musta pidevat joont), ehk need, millele ennustatud tõenäosus on väiksem kui $\hat{p}_i = 0.3$, on ennustatud suhteliselt lähedale tegelikule väärtusele. Sellest lävendist paremale

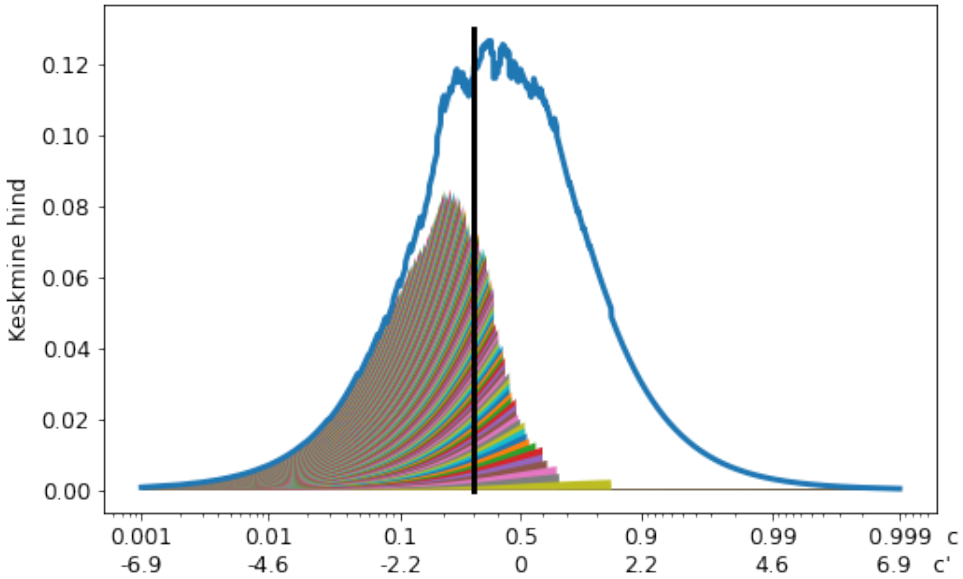
poole jäävad andmepunktid on ennustatud üldiselt kehvemini ja neid on kõikidest 209-st negatiivsest andmepunktist 70. Kui neid andmepunkte võrrelda, siis on võimalik väga hästi näha kuidas tegelikust väärtusest kaugemal ja tegelikule väärtusele lähedamal olevad andmepunktide kogu pindalad erinevad. Paremini ennustatud punktide pindala on palju väiksem kehvemini ennustatud punktide pindalast, kuigi neid on peaaegu poole rohkem ehk 139. See erinevus on ka paremini hoovatavam tulpdiagrammilt joonisel 15. Seal on näha, et paremini ennustatud negatiivsed andmepunktid, ehk tulbad, mis on lähemal nullile panustavad väga vähe kaost võrreldes andmepunktidega, mis on kaugemal nullist.



Joonis 15. Kõikide krediidiriski testandmete andmepunktide panustatud Brier'i skoor.

Kui vaadata ristentroopia kõverat (Joonis 16), kus on kujutatud iga andmepunkti panus ristentroopiana, siis on näha täpsemini ennustatud andmepunktide ja kehvemini ennustatud andmepunktide pindalade vahe, ehk panustatud kadude vahe. Joonise 14 hindade suhtest $c = 0.3$ (tähistatud musta püstise joonega) paremale poole jäävad tegelikult negatiivsete punktide pindala on palju suurem sellest suhtest vasakule jäävate punktide pindalast. Vasakul olevaid punkte on ka palju rohkem, kuigi need panustavad vähem ristentroopiasse.

Kui näiteks kehvasti ennustatud andmepunktid panustavad kogu pindalast kordades rohkem, kui hästi ennustatud, siis võib see klassifikaatori tegijale olla märk sellest, et neid



Joonis 16. Kõikide negatiivsete andmepunktide ristentroopia kujutatud ristentroopia pindaladena krediidiriski andmestikus.

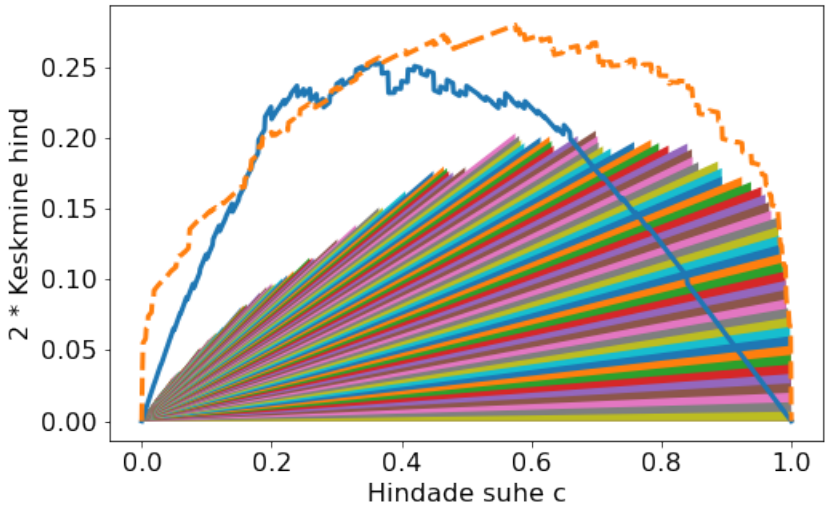
andmepunkte peaks paremini optimeerima. Kas peaks mudelit tegema paremaks nende andmepunktide ennustamisel või uurima, kas andmeid nende andmepunktide jaoks on piisavalt või kas andmed on piisavalt kvaliteetsed. Krediidiriski puhul tähendab see seda, et kas klassifikaatori kao jaoks on üldiselt halvemad pigem inimesed, kes on määratud valesse klassi või need kes on määratud õigesti.

Alternatiivselt saaks vaadata palju andmepunkte ennustati väga mööda. Sealt aga ei tule panustatud kadu välja. Võibolla on nende mööda ennustatud punktide kadu on nii väike, et see ei ole kuigi suur probleem klassifikaatori jaoks. Pigem panustavad näiteks suurema osa kaost siiski paremini ennustatud andmepunktid ja tuleks hoopis neid optimeerida.

3.2 Klassifikaatorite võrdlemine

Visualiseeringud on ka abiks erinevate klassifikaatorite ja andmetöötluste võrdlemisel. Näiteks, kui mõne klassifikaatori kadu on suurem kui teise oma, siis saab uurida, kui palju igast andmepunktist kadu tuleb ja võrrelda seda teise klassifikaatori andmepunktide kaoga.

Joonisel 17 on kujutatud sinise pidevjoonega tähistatud kõveraga eelnevalt krediidiriski andmestikul kasutatud otsustusmetsa klassifikaatorit. Nüüd on ka oranži kriipsjoonega lisatud naiivse Bayes'i klassifikaator. Kolmnurkadena on välja joonistatud tegelikult negatiivsete andmepunktide Brier'i skoor pindalana naiivset Bayes'i klassifikaatorit kasutades. Enamik hindade suhete juures on otsustusmetsa kahekordne keskmine hind parem kui naiivsel Bayes'il, aga hindade suhte $c = 0.2$ piirkonnas on naiivne Bayes veidike parem. Andmepunktide kao visualiseerimisel on võimalik uurida, milliste väärtustega on selles-se naiivse Bayesi ja otsustusmetsa vahesse panustavad andmepunktid. Jooniselt näeb, et

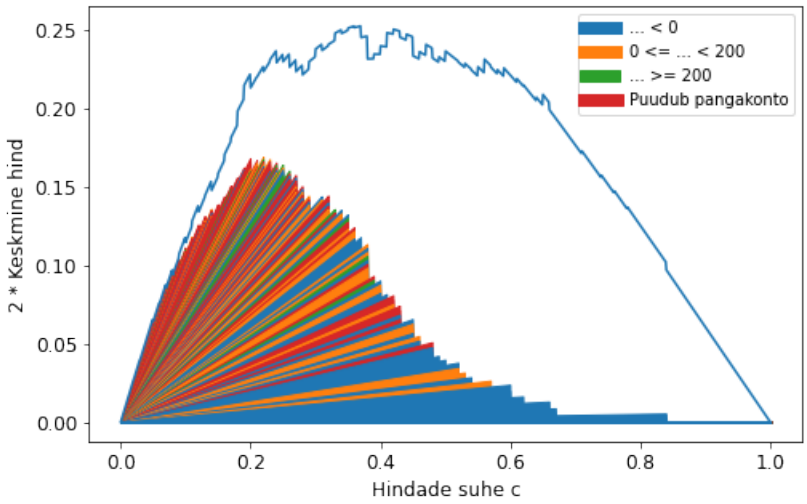


Joonis 17. Kõikide negatiivsete andmepunktide Brier'i skoor kujutatud Brier'i skoori pindaladena krediidiriski andmestikus naiivse Bayes'i klassifikaatoriga. Sinise pidevjoonega kõver on otsustusmetsa klassifikaatori Brier'i kõver ja oranži kriipsjoonega kõver on naiivse Bayes'i klassifikaatori Brier'i kõver.

oranži, ehk naiivse Bayes'i kõver on alates hindade suhtest 0.3 kuni 1 kahe kordse keskmise hinna poolest igal suhtel halvem kui otsustusmets. Tegelikult negatiivsete andmepunktide pindalasid välja joonistades võib näha, et enamus selles piirkonnas panustatud kaost tuleb negatiivsetest andmepunktidest, mille pindalad lõppevad hindade suhte piirkonnas $c = 1$, ehk nende punktidele ennustatud tõenäosus on $\hat{p}_i = 1$. Ehk suurema osa sellest pindalast panustavad tegelikult negatiivsed andmepunktid, mille tõenäosus on ennustatud täpselt $\hat{p}_i = 1$, ehk täiesti mööda. Seega on visualiseeringute abil võimalik võrrelda klassifikaatoreid ja näha, millistest andmepunktidest tuleneb nende kadu.

3.3 Andmepunktide tunnuste eraldamine

Viualiseeringutega on võimalik uurida, kas mõne tunnusega andmepunkt panustab kaosse ebaproportsionaalselt palju võrreldes teisi tunnuseid omavate andmepunktidega. Selleks tuleks jagada andmepunkti panustatud pindalad tunnuste järgi, et seda visualiseerida. Joonisel 18 on kujutatud krediidiriski andmestiku ühte tunnust. Iga andmepunkt andmestikus oli inimene ja eesmärgiks oli vaja ennustada, kas sellele inimesele on pangal rahaliselt mõistlik laenu anda.



Joonis 18. Kõikide negatiivsete andmepunktide Brier'i skoor kujutatud pindaladena krediidiriski andmestikus.

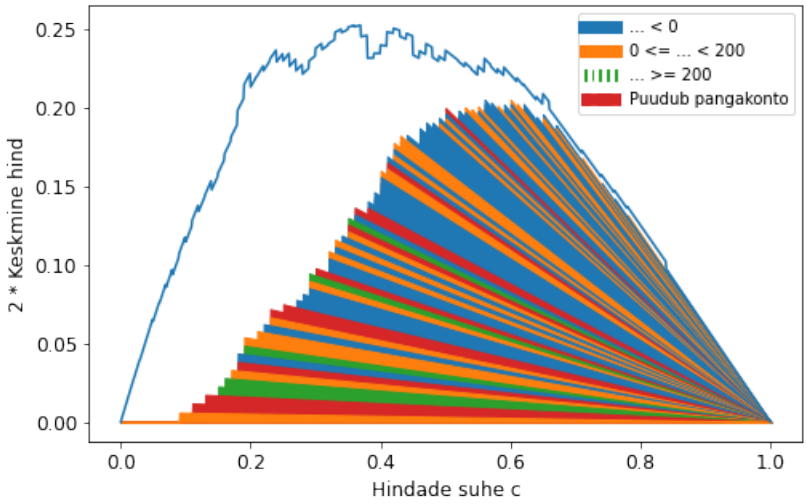
Joonisel 18 on kujutatud ainult inimesed, kellele oli mõistlik laenu anda. Visualiseeritud on üks tunnus, milleks on inimese pangakonto seis. Eraldi värvidega on kujutatud iga selle tunnuse väärtus. Sinised on inimesed, kelle pangakonto seis on alla 0 ühiku, kollane 0 ja 200 vahel, roheline üle 200 ja punane inimesed, kellel pole pangas kontot. Negatiivseid andmepunkte on joonisel 209.

Joonisel 18 ja tabelist 2 on näha, et inimesi, kelle pangakonto seis oli alla nulli, on klassifikaatoril keeruline ennustada. Selle tunnuse väärtusega andmepunkte on 300-st punktist 44, aga nende pindala ehk kadu on 0.025 ja need moodustavad väga suure osa kaost arvestades nende arvu. Punaseid, ehk inimesi kellel puudub pangakonto ja keda on üle

Tabel 2. Krediidiriski andmestiku negatiivsete andmepunktidele, millel on erinev pangakonto seis, vastavad kadu ja andmepunktide arv.

Tunnus	Kadu	Andmepunktide arv
... <0	0.025	44
$0 \geq$ ...<200	0.015	45
... ≥ 200	0.003	14
Puudub pangakonto	0.015	108

poole rohkem, ehk 108, panustavad kaosse 0.015, seega mõnevõrra vähem. See võib olla tingitud kas mudelist endast või ka sellest, et nende andmepunktide kohta pole andmed piisavalt head. Inimesi, kellel polnud varem kontot pangas ja kellele oli mõistlik laenu anda, oskab mudel ennustada suhteliselt täpselt võrreldes teiste tunnuste väärtustega.



Joonis 19. Kõikide positiivsete andmepunktide Brier'i skoor kujutatud ristentroopia pindaladena krediidiriski andmestikus.

Joonisel 19 on kujutatud ainult inimesed, kellele ei olnud mõistlik laenu anda. Sinised, ehk konto seisuga alla nulli on enamjaolt ennustatud täpsemalt kui joonisel 18. Siin ei tulnud aga pindalade erinevus nii hästi välja, ning see on kujutatud ka tabelis 3.

Tabel 3. Krediidiriski andmestiku positiivsetele andmepunktidele, millel on erinev pangakonto seis, vastavad kadu ja andmepunktide arv.

Tunnus	Kadu	Andmepunktide arv
... <0	0.045	46
0 $\geq$ ...<200	0.032	29
... $\geq$ 200	0.01	5
Puudub pangakonto	0.019	11

3.4 Puudujäägid ja edasiarendamise võimalused

Kuigi on täidetud põhi eesmärk, et leida viis kuidas kadu tõlgendada andmepunktide kaudu, on autori silmis tööl mõningad puudused. Need võivad olla ka tulevikus töö edasiarendamise võimalused.

Tööst on puudu viualiseeringute praktiline rakendus. Edasi võiks uurida, kuidas jooniste abil saab klassifikaatoreid või andmestikku kvaliteeti parandada. Alampeatükis 3.3 tutvustati andmepunktide tunnuste järgi eraldamist jooonisetel. Inimesed, kellele oli mõistlik laenu anda ja kelle tunnuseks oli pangakonto negatiivne seis olid klassifikaatori poolt kehvemini ennustatud. Uurida saaks kuidas andmete täiendamisel selle tunnusega inimestega visualisatsioon muutuks. Samuti võiks proovida klassifikaatorit täiendada või kasutada otsustusmetsa asemel muid mudeleid.

Ka visualisatsioonidel on autori arvates puudujäägid. Kui klassifikaatori tegija jaoks on oluline ainult, et mudel piisavalt täpselt ennustaks, siis ei pruugi visualisatsioonitest abi olla, sest võib lihtsalt vaadata, kui palju on kehvasti ennustatud andmepunkte. Kui aga tähtis on ka kadu, siis võib kehvasti ennustatud andmepunktide kadu olla piisavalt väike, et suurema osa kaost panustavad paljud väikese kaoga õigemini ennustatud andmepunktid. Sellisel juhul oleks võibolla mõistlikum uurida, kuidas neid paremini ennustatud andmepunkte saaks optimeerida.

Ka erinevate andmepunktide pindalade eraldamine värvidega võib olla segadusse ajav. Kui pindalade vahed on liiga väiksed, siis võib pindalade võrdlemine olla raskendatud.

Selle nõrkuse leevendamiseks tehti ka tulpdiagramm, aga sealt on keerulisem lugeda välja andmepunktide kogu panust kaos.

Kokkuvõte

Töö eesmärgiks oli uurida, kuidas kaofunktsiooni väljundit saab tõlgendada andmepunktide kaudu. Kaofunktsiooni väljund on ainult üks numbriline väärtus ja selle laiemaks tõlgendamiseks ja visualisatsioonide loomise jaoks püstitati kolm küsimust.

Varasemalt on kaofunktsiooni väljundit, ehk kadu tõlgendatud pindaladena kõverate all [3, 6, 8]. Töö esimeses osas anti algteadmised hinna, Brier'i ja ristentroopia kõverate mõistmiseks. Seejärel tutvustati neid kõveraid. Teises osas vaadeldi, kuidas oleks võimalik nende kõverate abil kujutada iga andmepunkti panustatud kadu nende kõverate aluses pindalas. Kolmandas töö osas uuriti visualisatsioone ja vastati sissejuhatuses püstitatud küsimustele. Viimases alampeatükis toodi välja töö ja tutvustatud visualisatsioonide arvatavad nõrkused. Peamiselt on tööst puudu tutvustatud meetodi praktiline rakendus. Selleks, et visualisatsioone oleks võimalik joonistada ka teistel andmetel, valmis Pythonis teek (Lisa I).

Esiteks on visualiseeringute abil on võimalik hõlpsamini uurida millised andmepunktid panustavad suurema osa kaost, kas need, mille ennustatud tõenäosus on lähemal tegelikule väärtusele ja mida on rohkem või need, mille ennustatud tõenäosus on kaugemal tegelikust väärtusest ja mida on vähem.

Teiseks saab võrrelda erinevaid klassifikaatoreid. Kui mõnes hindade suhte piirkonnas on üks klassifikaator nõrgem kui teine, siis saab uurida millised andmepunktid on sellest lisanduvast kaost suurema osa panustanud.

Kolmandaks saab eraldada andmepunktid mõne konkreetse tunnuse põhjal ja uurida, kas mõni tunnus panustab kaosse ebaproportsionaalselt palju võrreldes teisi tunnuseid omavate andmepunktidega. See võib olla märk sellest, et klassifikaator ei oska selle tunnusega andmepunkte piisavalt hästi ennustada ja viga võib olla klassifikaatoris endas või andmetes.

Viidatud kirjandus

[1] Glenn W Brier. „Verification of forecasts expressed in terms of probability“. *Monthly weather review* 78.1 (1950), lk. 1–3.

[2] Andreas Buja, Werner Stuetzle ja Yi Shen. „Loss functions for binary class probability estimation and classification: Structure and applications“. *Working draft, November* 3 (2005).

[3] Chris Drummond ja Robert C Holte. „Cost curves: An improved method for visualizing classifier performance“. *Machine learning* 65.1 (2006), lk. 95–130.

[4] Charles Elkan. „The foundations of cost-sensitive learning“. Teoses: *International joint conference on artificial intelligence*. Köide 17. 1. Lawrence Erlbaum Associates Ltd. 2001, lk. 973–978.

[5] Peter Flach. *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York: Cambridge University Press, 2012.

[6] José Hernández-Orallo, Peter A Flach ja Cèsar Ferri Ramirez. „Brier curves: a new cost-based visualisation of classifier performance“. Teoses: *Proceedings of the 28th International Conference on Machine Learning*. 2011, lk. 585–592.

[7] Hans Hofmann. *UCI Machine Learning Repository*. 1994. URL: `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`. (07. 05. 2021).

[8] Meelis Kull. „Do we need to estimate the calibration loss of probabilistic classifiers? If yes, then how?“ 2020. URL: `https://drive.google.com/file/d/1jxi5vyHG9EtC35PHZt577W_PXToWC-Jl/view?usp=sharing`. (07. 05. 2021).

[9] Edgar C Merkle ja Mark Steyvers. „Choosing a strictly proper scoring rule“. *Decision Analysis* 10.4 (2013), lk. 292–304.

[10] Bianca Zadrozny ja Charles Elkan. „Learning and making decisions when costs and probabilities are both unknown“. Teoses: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. 2001, lk. 204–213.

Lisad

I Koodi repositoorium

Töö käigus valmis Python'i teek Brier'i kõverate, tulpdiagrammide ja iga andmepunkti panuse visualiseerimiseks kaos. Teek ja töös joonistatud visualisatsioonide joonistamiseks kasutatud kood Jupyter Notebookides on leitav Githubi repositooriumist: https://github.com/magnuspaal/datapointloss.

II Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Magnus Paal,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Andmepunktide panuse visualiseerimine binaarklassifitseerija kaos,

   mille juhendaja on Meelis Kull,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Magnus Paal
07.05.2021