TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Kaspar Valk

Klassifitseerija kalibreerituse testi võimsuse suurendamine

Bakalaureusetöö (9 EAP)

Juhendaja: Meelis Kull

Tartu 2020

Klassifitseerija kalibreerituse testi võimsuse suurendamine

Lühikokkuvõte: Masinõppes nimetatakse klassifitseerijat kalibreerituks, kui selle ennustatud klasside tõenäosused vastavad ka tegelikule andmete klassijaotusele. Ohutust nõudvates klassifitseerimisülesannetes on oluline, et klassifitseerija ennustused ei oleks enese- või ebakindlad, vaid et need oleksid kalibreeritud. Kalibreeritust on võimalik hinnata mõõdu ECE abil ning ECE väärtuse põhjal on omakorda võimalik konstrueerida kalibreerituse test: statistiline test, milllega kontrollida hüpoteesi, et klassifitseerija on kalibreeritud. Töös otsiti katseliselt optimaalseid parameetreid ECE arvutamisel, et selle põhjal sooritatud kalibreerituse test oleks võimalikult võimas. See tähendab, et möödakalibreeritud klassifitseerija korral suudaks test võimalikult tihti ümber lükata nullhüpoteesi, et klassifitseerija on kalibreeritud. Töös jõuti tulemuseni, et võimalikult võimsa kalibreerituse testi jaoks on ECE arvutamisel mõistlik paigutada iga andmepunkt eraldi vahemikku. Kui on eeldus, et esineb andmepunkte, mille jaoks klassifitseerija on väga möödakalibreeritud, siis on parim kasutada Kullback-Leibleri kaugusest inspireeritud logaritmkaugust. Muudel juhtudel on mõistlikum kasutada absoluut- või ruutkaugust. Need soovitused erinevad oluliselt varasemas teaduskirjanduses tavaks olnud ECE arvutamisel kasutatud parameetritest. Töös leitu võimaldab paremini tuvastada klassifitseerijate mittekalibreeritust.

Võtmesõnad: masinõpe, klassifitseerija kalibreeritus, eeldatav kalibreerimisviga, testi võimsus

CERCS: P176 Tehisintellekt

Increasing the Power of a Classifier’s Calibration Test

Abstract: In machine learning, a classifier is called to be calibrated if its predicted class probabilities match with the actual class distribution of the data. In classification tasks where safety is necessary, it is important that the classifier’s predictions would not be over- or underconfident but instead would be calibrated. Calibration can be evaluated using the measure ECE, and based on its value it is possible to construct a calibration test: a statistical test which allows to check if the hypothesis that the model is calibrated holds. In the thesis, experiments were performed to find optimal parameters for

calculating ECE, so that the calibration test based on this would be as powerful as possible. That is, for a miscalibrated classifier the test would be able to reject the null hypothesis that the model is calibrated as frequently as possible. The work concluded that to make the calibration test as powerful as possible, the datapoints should be placed into separate bins when calculating ECE. If the dataset is expected to contain datapoints for which the classifier is largely miscalibrated, then it is best to use a variant of ECE with the logarithmic distance measure inspired by Kullback-Leibler divergence. Otherwise, it is more reasonable to use absolute or square distance. These recommendations differ significantly from conventional parameter values used when calculating ECE in previous scientific literature. The results of this thesis allow for improved identification of miscalibration in classifiers.

Keywords: machine learning, classifier's calibration, expected calibration error, power of a test

CERCS: P176 Artificial intelligence

Sisukord

Sissejuhatus

Klassifitseerijaid ehk masinõppe mudeleid, mis ennustavad andmete klassimärgendeid, kasutatakse mitmetes valdkondades erinevate ülesannete lahendamiseks: näiteks isejuhtivates autodes objektituvastuseks, meditsiinis diagnooside määramiseks või veebilehtedel kahtlaste sisselogimiste leidmiseks. Klassifitseerijatel on rakendusvaldkondi, kus andmete põhjal ei ole mõistlik väljastada ainult kõige tõenäosema klassi märgend, vaid iga klassi kohta sinna kuulumise tõenäosus (Flach, 2012). Näiteks võib vaadata klassifitseerijat, mille ülesandeks on tuvastada inimese nahka kujutava pildi põhjal, kas inimesel on pahaloomuline kasvaja, healoomuline kasvaja või ei ole kasvajat. Kui seda klassifitseerijat kasutada nahaarsti töö kiirendamiseks, siis on oluline, et selle väljundiks oleks igasse klassi kuulumise tõenäosus, mitte ainult kõige tõenäosem klassimärgend. Sellisel juhul saaks arst ise üle kontrollida ka juhud, kus mudel on üpris kindel, et kasvajat ei ole, ent kasvaja jaoks ennustatud tõenäosus on siiski märkimisväärne.

Ülesannetes, kus klassifitseerija ennustab igasse klassi kuulumise tõenäosust, võib olla oluline, et klassifitseerija väljastatavad tõenäosused vastaksid ka tegelikule mõõdetud klassijaotusele ehk oleksid kalibreeritud (Guo, Pleiss, Sun ja Weinberger, 2017). Kui kasvajat tuvastav klassifitseerija on suure koguse piltide jaoks ennustanud väljundi (0.1, 0.0, 0.9), kus klassid on vastavalt: *pahaloomuline kasvaja, healoomuline kasvaja, ei ole kasvaja*, siis klassifitseerija on kalibreeritud, kui nende piltide seas keskmiselt ühel juhul kümnest esines pahaloomuline kasvaja ning keskmiselt üheksal juhul kümnest ei esinenud kasvajat.

Kui soovime klassifitseerija kalibreeritust kontrollida, siis praktikas ei saa seda teha otseselt kalibreerituse definitsioonist lähtuvalt (Guo jt, 2017). Seda seetõttu, et klassifitseerija väljundiks olevad tõenäosused on pidevad suurused, ent teadaolevate märgenditega andmete arv, mille põhjal kalibreeritust otsustada, on piiratud (Guo jt, 2017). Klassifitseerija kalibreeritust on aga võimalik hinnata ECE (*expected calibration error*) ehk eeldatava kalibreerimisvea abil (Naeini, Cooper ja Hauskrecht, 2015). Klassifitseerija jaoks mõõdetud ECE väärtuse abil on omakorda võimalik sooritada

kalibreerituse test: statistiline test, millega mingi olulisusnivoo juures ümber lükata hüpotees, et mudel on perfektselt kalibreeritud (Vaicenavicius jt, 2019).

Bakalaureusetöö eesmärk on kalibreerituse testi võimsuse suurendamine. See tähendab, et kalibreerituse test suudaks kalibreerimata klassifitseerija korral võimalikult tihti ümber lükata hüpoteesi, et mudel on perfektselt kalibreeritud (Möls, 2013). Selleks otsitakse, milline on parim ECE arvutamisel kasutatav vahemike arv, vahemike paigutusviis ning kaugusfunktsioon.

Töö esimeses osas kirjeldatakse vajalikke taustateadmisi mõistmaks klassifitseerija kalibreeritust, ECEd ning statistilist testi klassifitseerija kalibreerituse kontrolliks. Töö teises osas kirjeldatakse läbi viidud katsete tulemusi ja meetodeid, et leida ECE arvutamisel kalibreerituse testi kontekstis optimaalset vahemike paigutusviisi, vahemike arvu ning kaugusfunktsiooni.

1 Mõisted ja definitsioonid

Selles töö osas kirjeldatakse põhilisi vajalikke taustateadmisi mõistmaks klassifitseerija kalibreeritust, ECEd ning selle põhjal sooritatud kalibreerituse testi.

1.1 Tõenäosuslik klassifitseerija

Klassifitseerimisülesannete kontekstis koosneb andmestik andmepunktidest X ehk tunnuste väärtustest ja andmepunktidele vastavatest klassimärgenditest Y, mis kirjeldavad, millisesse klassi andmepunkt kuulub (Flach, 2012). Näiteks võib vaadata ülesannet tuvastada inimese nahka kujutava pildi põhjal, kas inimesel on pahaloomuline kasvaja, healoomuline kasvaja või ei ole kasvajat. Üheks andmepunktiks on konkreetse pildi pikslite värvikanalite väärtused ning selle andmepunkti klassimärgendiks tegelik pildil kujutatu: pahaloomulise kasvajaga, healoomulise kasvajaga või kasvajata nahk.

Klassifitseerijaks nimetatakse funktsiooni, mis iga andmepunkti $x \in X$ jaoks tagastab klassimärgendi $c(x)$ (Flach, 2012). Tõenäosuslikuks klassifitseerijaks nimetatakse funktsiooni, mis iga andmepunkti $x \in X$ jaoks tagastab tõenäosusvektori $(p_1, \dots, p_k)$ üle kõigi võimalike klasside, kus p_i kirjeldab ennustatud tõenäosust, et andmepunkt kuulub i-ndasse klassi (Flach, 2012).

Bakalaureusetöös kasutatakse edaspidi terminit klassifitseerija tähenduses tõenäosuslik klassifitseerija. Terminiga sisend tähistatakse edaspidi andmepunkti $x \in X$ ning terminiga väljund tähistatakse edaspidi klassifitseerija sisendile ennustatud tõenäosusvektorit $(p_1, \dots, p_k)$.

1.2 Klassifitseerija kalibreeritus

Kalibreeritus on klassifitseerija omadus, mis kirjeldab, kas sisendi jaoks ennustatud tõenäosused vastavad tegelikule sisendi klassijaotusele (Vaicenavicius jt, 2019). Kull jt (2019) on klassifitseerija kalibreeritust täpsemalt defineerinud kolmel eri viisil: mitmeklassiliselt kalibreeritud (*multiclass-calibrated*) ehk kalibreeritud, enesekindluse järgi kalibreeritud (*confidence calibrated*) ja klassikaupa kalibreeritud (*classwise-*

calibrated). Kõige rangem neist on mitmeklassiliselt kalibreerituse definitsioon, kusjuures mitmeklassiliselt kalibreeritud klassifitseerija on kalibreeritud ka enesekindluse järgi ning klassikaupa (Kull jt, 2019). Järgnevalt on toodud need kolm definitsiooni (Kull jt, 2019):

- Klassifitseerijat nimetatakse mitmeklassiliselt kalibreerituks, kui iga tõenäosusvektori $(p_1, \dots, p_k)$ korral on kõigi sisendite seas, mille väljundiks on vektor $(p_1, \dots, p_k)$, tegelik klassijaotus ka $(p_1, \dots, p_k)$.
- Klassifitseerijat nimetatakse enesekindluse järgi kalibreerituks, kui iga tõenäosuse $p \in [0,1]$ korral on kõigi sisendite seas, mille väljundis esinev maksimaalne tõenäosus on p, maksimaalse tõenäosusega klassi kuuluvate sisendite osakaal p.
- Klassifitseerijat nimetatakse klassikaupa kalibreerituks, kui iga väljundklassi i jaoks kehtib järgnev: kõigi sisendite seas, mille väljundis on klassi i kuulumise tõenäosus p_i, on klassi i kuuluvate sisendite osakaal p_i.

Näide 1. Olgu meil kolmeklassiline klassifitseerija, mille kogu sisendiruum koosneb tabelis 1 toodud kümnest sisendist. Sel juhul on see klassifitseerija mitmeklassiliselt kalibreeritud, kuna

- tõenäosusvektori (0.2, 0.2, 0.6) puhul on viie sisendi väljundiks see vektor ning neist täpselt üks kuulub esimesse, üks teise ning kolm kolmandasse klassi (ehk klasside jaotuse proportsioonid on $1:1:3 = 0.2:0.2:0.6$);
- tõenäosusvektori (0.2, 0.0, 0.8) puhul on viie sisendi väljundiks see vektor ning neist täpselt üks kuulub esimesse, null teise ning neli kolmandasse klassi.

Samuti on see klassifitseerija enesekindluse järgi kalibreeritud, kuna

- tõenäosuse 0.6 puhul on viie sisendi jaoks väljundis maksimaalne tõenäosus 0.6 ning kolm neist kuuluvad maksimaalse tõenäosusega klassi;
- tõenäosuse 0.8 puhul on viie sisendi jaoks maksimaalne tõenäosus 0.8 ning neli neist kuuluvad maksimaalse tõenäosusega klassi.

Samuti on see klassifitseerija klassikaupa kalibreeritud, kuna

- esimese klassi jaoks on kümme sisendit, mis omavad tõenäosust 0.2 ning neist täpselt kaks kuuluvad esimesse klassi;

- teise klassi jaoks on viis sisendit, mis omavad tõenäosust 0.2 ning neist üks kuulub teise klassi, ja viis sisendit, mis omavad tõenäosust 0.0 ning neist null kuulub teise klassi;
- kolmanda klassi jaoks on viis sisendit, mis omavad tõenäosust 0.6 ning neist kolm kuuluvad kolmandasse klassi, ja viis sisendit, mis omavad tõenäosust 0.8 ning neist neli kuuluvad kolmandasse klassi.

Tabel 1. Klassifitseerija sisendid näites 1

Sisend	**Tegelik klass**	**Väljund**
x_1	1	(0.2, 0.2, 0.6)
x_2	2	(0.2, 0.2, 0.6)
x_3	3	(0.2, 0.2, 0.6)
x_4	3	(0.2, 0.2, 0.6)
x_5	3	(0.2, 0.2, 0.6)

Sisend	**Tegelik klass**	**Väljund**
x_6	1	(0.2, 0.0, 0.8)
x_7	3	(0.2, 0.0, 0.8)
x_8	3	(0.2, 0.0, 0.8)
x_9	3	(0.2, 0.0, 0.8)
x_{10}	3	(0.2, 0.0, 0.8)

Praktikas aga ei saa peaaegu ühegi klassifitseerija kalibreeritust kontrollida otseselt kalibreerituse definitsioonist lähtuvalt, kuna klassifitseerija väljundiks olevad tõenäosused on pidevad suurused, ent teadaolevate märgenditega andmete arv on piiratud (Guo jt, 2017). Tabelis 1 kirjeldatud kümne sisendiga klassifitseerija on vaadeldaval kümne andmepunktiga andmestikul küll kalibreeritud, kuid kas see võis olla tingitud ka juhuslikkusest? Kui kõigi sisendite $x_6, \dots, x_{10}$ tegelik klassimärgend oleks olnud 3, siis kas on võimalus, et klassifitseerija on siiski kalibreeritud? Samuti, kuidas anda kalibreeritusele hinnangut, kui sisendite $x_6, \dots, x_{10}$ väljundid oleksid olnud $(0.21, 0.0, 0.79)$ või hoopis $(0.15, 0.13, 0.72)$? Kalibreeritud klassifitseerija korral ei oleks selliste väljundite ja nii väheste sisendite korral kunagi võimalik definitsiooni põhjal järeldada, et klassifitseerija on kalibreeritud. Seetõttu on olemas ka järgnev meetod kalibreerituse mõõtmiseks.

1.3 Eeldatav kalibreerimisviga

ECE (*expected calibration error*) ehk eeldatav kalibreerimisviga on meetod kalibreerituse mõõtmiseks (Naeini jt, 2015). Kuigi sõltuvalt kalibreerituse definitsioonist on ECE leidmine erinev, on see siiski kõigil juhtudel intuitiivselt sarnane. Klassifitseerija ennustatud tõenäosused jagatakse vahemikesse ning igas vahemikus leitakse keskmine ennustatud tõenäosus ja tegelik vaadeldavasse klassi kuuluvate sisendite osakaal. Lõpuks leitakse, kui palju erinevad need väärtused keskmiselt üle kõigi vahemike. Töös nimetatakse enesekindluse järgi kalibreerituse korral leitud ECEd kui cf-ECE (*confidence*-ECE) ning klassikaupa kalibreerituse korral leitud ECEd kui cw-ECE (*classwise*-ECE). Järgnevalt on ECE leidmise algoritmi täpsemalt kirjeldatud.

1.3.1 ECE leidmise algoritm

Guo jt (2017) on kirjeldanud cf-ECE mõõtmise algoritmi absoluutkaugusega varianti. Üldistatuna igasugusele kaugusfunktsioonile on see järgnev (Guo jt, 2017):

1. Olgu meil n andmepunkti, millest igaühe kohta on teada tema tegelik klassimärgend $c \in \{1, \dots, k\}$.
2. Vaatleme iga andmepunkti jaoks klassifitseerija väljundiks oleva tõenäosusvektori $\boldsymbol{pred} = (p_1, \dots, p_k)$ kõrgeimat tõenäosust $p_m = \max(p_1, \dots, p_k)$. Samuti leiame tõeväärtuse L, mis näitab, kas p_m on ennustatud samale klassile, mis tegelik klassimärgend: $L = \begin{cases} 1, \text{ kui } m = c \\ 0, \text{ kui } m \neq c \end{cases}$.
3. Vaatleme eelneval sammul leitud tõenäosuseid p_m ning neile vastavaid tõeväärtuseid L ning jagame nad b vahemikku vastavalt tõenäosustele p_m ning mingile vahemikeks jaotamise algoritmile. Vahemikeks jaotamise algoritme tutvustatakse täpsemalt peatükis 1.3.2.
4. Igas vahemikus B_i vaatame sinna paigutatud tõenäosuseid ${p_m}^{(1)}, \dots, {p_m}^{(j)}$ ning neile vastavaid tõeväärtuseid $L^{(1)}, \dots, L^{(j)}$, kus j tähistab vahemikku B_i langenud punktide arvu ehk $j = |B_i|$. Leiame keskmise vahemikus ennustatud tõenäosuse

$$p(B_i) = \frac{{p_m}^{(1)} + \cdots + {p_m}^{(j)}}{j}.$$

   Leiame tegeliku vaadeldavasse klassi kuuluvate sisendite osakaalu vahemikus

$$y(B_i) = \frac{L^{(1)} + \cdots + L^{(j)}}{j}.$$

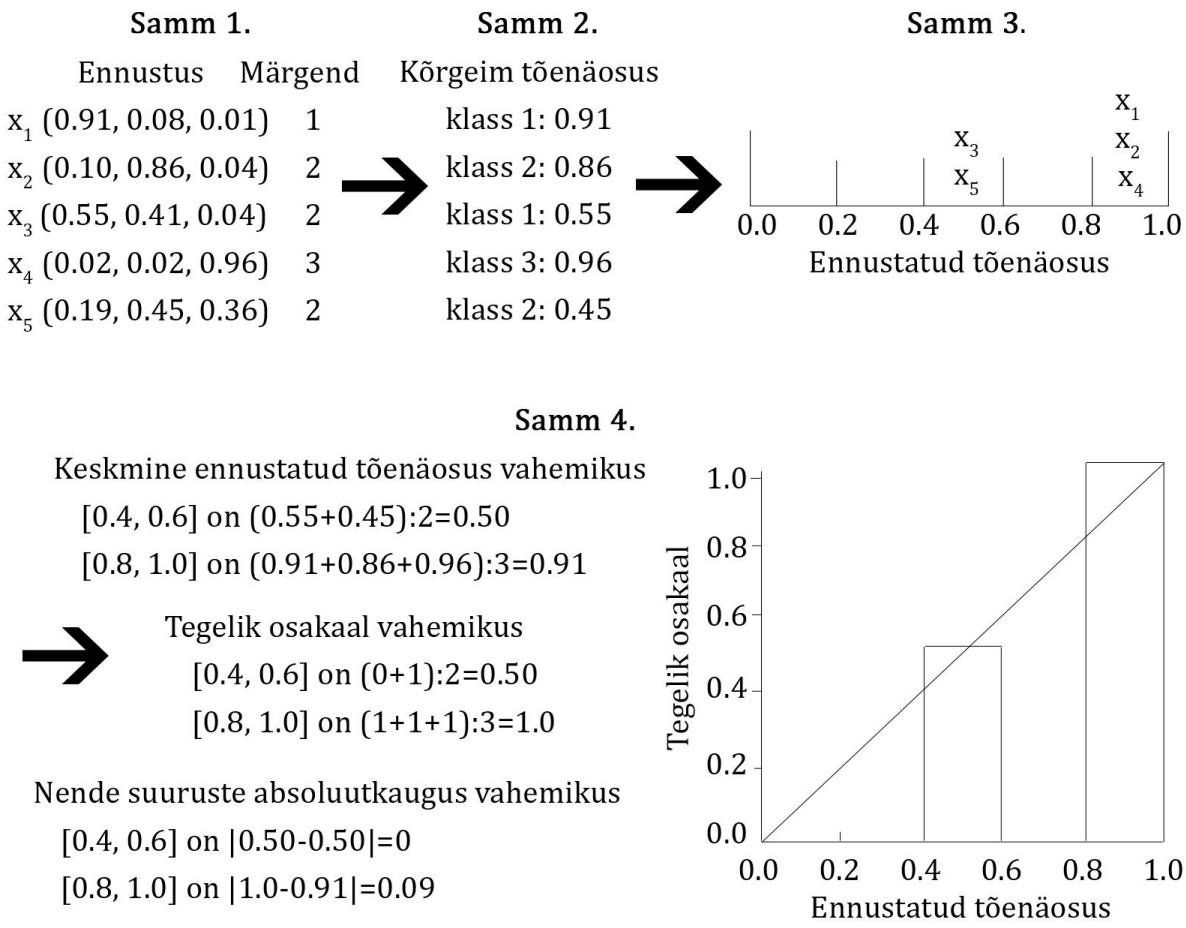
Leiame nende suuruste kauguse $d\big(y(B_i), p(B_i)\big)$ mingi kaugusfunktsiooni põhjal. Võimalikke kaugusfunktsioone kirjeldatakse täpsemalt peatükis 1.3.3.

5. Igas vahemikus leiame selle vahemiku andmepunktide osakaalu kõigi andmepunktide seast $\frac{|B_i|}{n}$ ning kaalume kaugusfunktsiooni väärtust vahemikus selle osakaaluga: $\frac{|B_i|}{n} \cdot d\big(y(B_i), p(B_i)\big)$.
6. cf-ECE väärtus on kõigi vahemike kaalutud kaugusfunktsioonide väärtuste summa.

Matemaatilises kirjapildis on see

$$\text{cf-ECE} = \sum_{i=1}^{b} \frac{|B_i|}{n} \cdot d(y(B_i), p(B_i)),$$

kus b on vahemike arv, $|B_i|$ on andmepunktide arv vahemikus B_i, $y(B_i)$ on maksimaalse tõenäosusega klassi kuuluvate sisendite tegelik osakaal vahemikus B_i, $p(B_i)$ on keskmine ennustatud tõenäosus vahemikus B_i. Joonisel 1 on kujutatud cf-ECE leidmine viie andmepunkti põhjal.



Joonis 1. cf-ECE leidmine viie andmepunkti põhjal

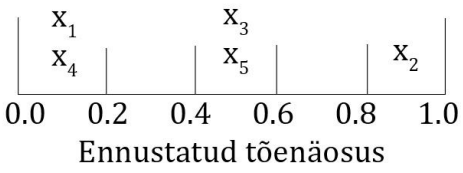
Kull jt (2019) on kirjeldanud klassikaupa ECE ehk cw-ECE leidmist ning järgnev lõik põhineb nende kirjeldatul. Klassikaupa kalibreerituse puhul on ECE leidmine sarnane, ent algoritmi korratakse iga klassi jaoks. Algoritmi teisel sammul vaadeldakse iga andmepunkti jaoks maksimaalse tõenäosuse asemel kindla klassi j tõenäosust $p_m = \boldsymbol{pred}[j]$ ehk $m = j$. Algoritmi lõpptulemuseks on iga klassi jaoks leitud ECE väärtuste aritmeetiline keskmine. Matemaatilises kirjapildis on see

$$\text{cw-ECE} = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=1}^{b}\frac{|B_{i,j}|}{n}\cdot d\left(y\left(B_{i,j}\right), p\left(B_{i,j}\right)\right),$$

kus k on klasside arv, b on vahemike arv, $\left|B_{i,j}\right|$ on andmepunktide arv vahemikus $B_{i,j}$, $y\left(B_{i,j}\right)$ on klassi j kuuluvate sisendite tegelik osakaal vahemikus $B_{i,j}$, $p(B_{i,j})$ on keskmine ennustatud klassi j kuulumise tõenäosus vahemikus $B_{i,j}$. Joonisel 2 on kujutatud cw-ECE leidmise etapp klassi 2 jaoks sama viie andmepunkti põhjal, mis joonisel 1. Täieliku cw-ECE leidmiseks tuleks korrata seda protsessi ka klassi 1 ja 3 jaoks ning leida kolme klassi ECEde aritmeetiline keskmine.

Samm 1. **Samm 2.** **Samm 3.**

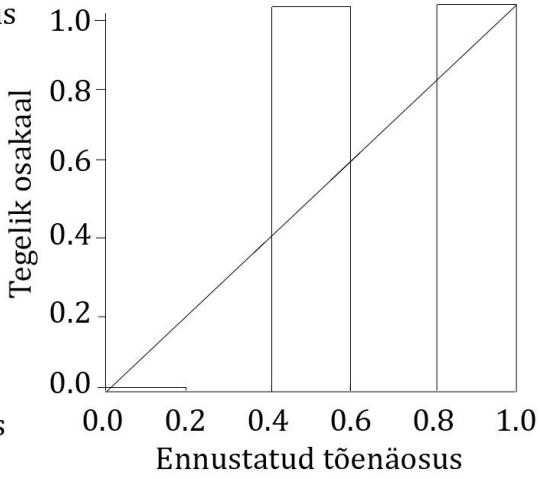
	Ennustus	Märgend	Klassi 2 tõenäosus
x_1	(0.91, 0.08, 0.01)	1	0.08
x_2	(0.10, 0.86, 0.04)	2	0.86
x_3	(0.55, 0.41, 0.04)	2	0.41
x_4	(0.02, 0.02, 0.96)	3	0.02
x_5	(0.19, 0.45, 0.36)	2	0.45



Samm 4.

Keskmine ennustatud tõenäosus vahemikus
[0.0, 0.2] on (0.08+0.02):2=0.05
[0.4, 0.6] on (0.41+0.45):2=0.43
[0.8, 1.0] on 0.86:1=0.86

Tegelik klassi 2 osakaal vahemikus
[0.0, 0.2] on (0+0):2=0.0
[0.4, 0.6] on (1+1):2=1.0
[0.8, 1.0] on 1:1=1.0



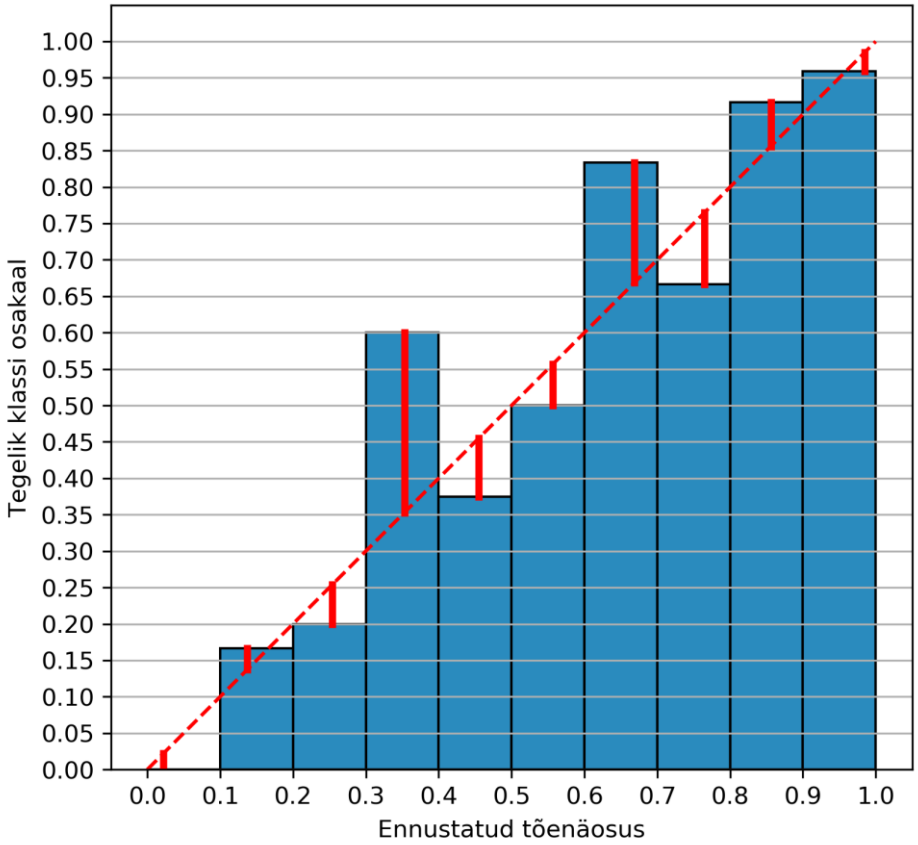
Nende suuruste absoluutkaugus vahemikus
[0.0, 0.2] on |0.05-0.0|=0.05
[0.4, 0.6] on |0.43-1.0|=0.57
[0.8, 1.0] on |0.86-1.0|=0.14

Samm 5. **Samm 6.**

Kaugusfunktsioonide väärtused kaalutud vastava vahemiku suurusega vahemikus
[0.0, 0.2] on $^2/_5 \cdot 0.05 = 0.02$
[0.4, 0.6] on $^2/_5 \cdot 0.57 = 0.228$
[0.8, 1.0] on $^1/_5 \cdot 0.14 = 0.028$

ECE klassi 2 jaoks on
0.02+0.228+0.028=0.276

Joonis 2. cw-ECE leidmise etapp klassi 2 jaoks viie andmepunkti põhjal

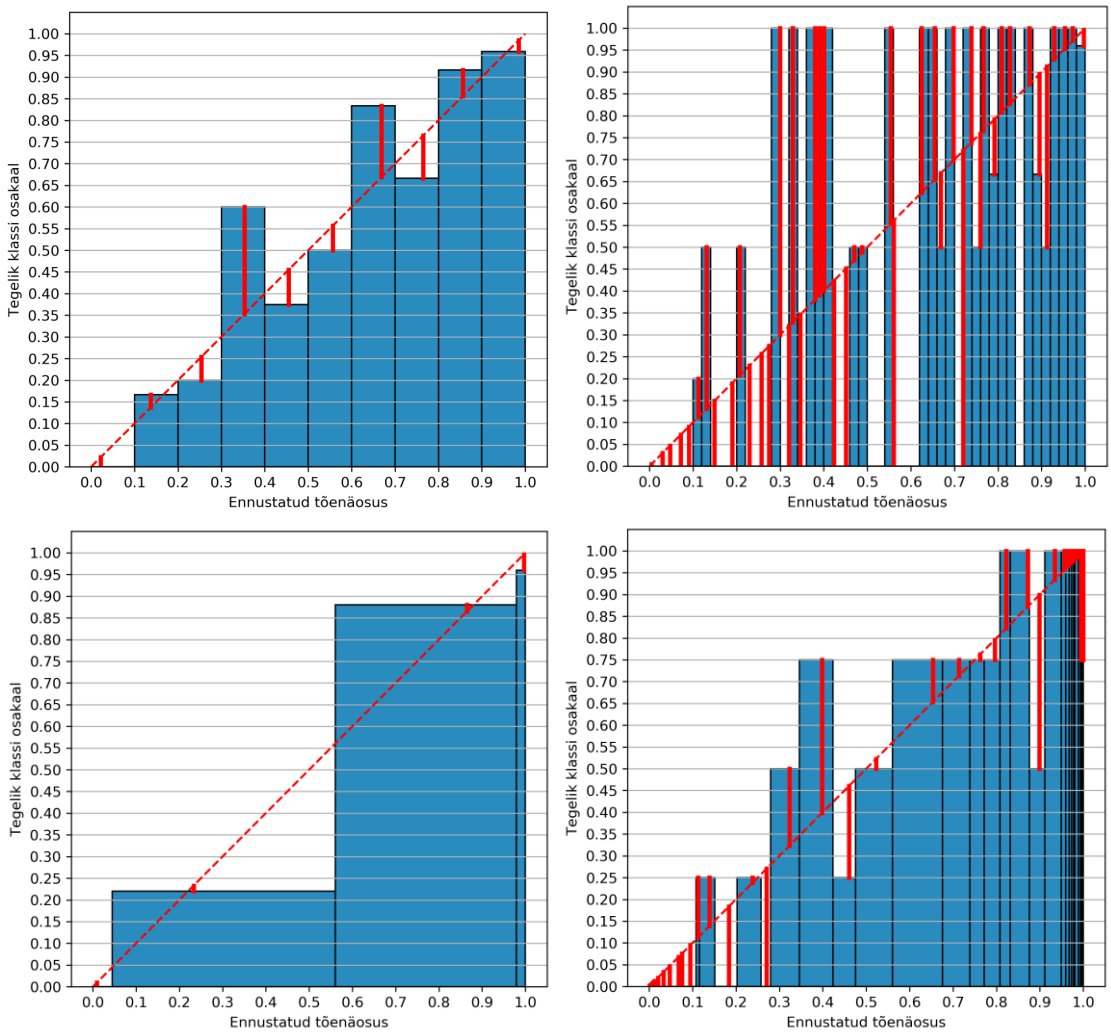


Joonis 3. Usaldusdiagramm cw-ECE puhul ühe klassi kohta

Joonisel 3 on kujutatud cw-ECE jaoks usaldusdiagramm (*reliability diagram*) ühe klassi kohta ehk üks võimalus ECE visualiseerimiseks (Kull jt, 2019). Joonisel on kujutatud cw-ECE leidmine 2-klassilise klassifitseerija teise klassi jaoks 200 andmepunkti põhjal, kus andmed ja klassfitseerija on tehislikult koostatud vastavalt töö eksperimentaalses osas kirjeldatud katsele 2. Ennustused on jagatud kümne võrdse laiusega vahemiku vahel. Joonise x-teljel on klassifitseerija ennustatud tõenäosus klassi 2 jaoks, y-teljel keskmine klassi 2 kuulumise osakaal vahemikus. Punase ribaga on tähistatud iga vahemiku jaoks kaugus keskmise ennustuse ja tegeliku vaadeldava klassi osakaalu vahel. Keskmine ennustus vahemikus ei pruugi olla alati vahemiku keskel: ennustused võivad olla koondunud ka vahemiku ühte serva. Seetõttu pole ka punase ribaga tähistatud kaugus mõõdetud tingimata vahemiku keskelt. Klassikaupa ECEst võib mõelda kui keskmisest vahemiku kaugusest diagonaalist keskmistatult omakorda üle kõigi klasside usaldusdiagrammide, kus iga vahemiku kaugus on kaalutud andmepunktide osakaaluga vahemikus. Perfektselt kalibreeritud klassifitseerija korral vastavad piisavalt suure andmete hulga juures x-telje väärtused y-telje väärtustega ehk iga vahemiku tipp on täpselt diagonaalil ning punaste ribade pikkused on olematud.

1.3.2 Vahemike paigutusviis ja arv ECE algoritmis

ECE arvutamise algoritmi kolmandal sammul kasutatakse vahemikeks jagamiseks kaht viisi. Guo jt (2017) on kasutanud võrdse laiusega vahemikke (*equal width bins*), kus iga vahemik hõlmab võrdse laiusega ala piirkonnas [0,1]. Kumar, Liang ja Ma (2019) on kasutanud võrdse suurusega vahemikke (*equal size bins*), kus vahemikud valitakse nii, et igasse vahemikku langeks võrdne kogus andmepunkte. Kull jt (2019), Widmann, Lindsten ja Zachariah (2019), Vaicenavicius jt (2019) on kasutanud mõlemat viisi. Puudub ka universaalselt kasutatud arv mitme vahemiku vahel ennustusi jagada: artiklites on kasutatud vahemike arvuna nii kümmet kui viitteist vahemikku.



Joonis 4. Erinevad usaldusdiagrammid cw-ECE jaoks. Esimeses reas on diagramm tehtud esimesel graafikul kümne ning teisel graafikul viiekümne võrdse laiusega vahemikuga. Teises reas on diagramm tehtud esimesel graafikul nelja ja teisel graafikul viiekümne võrdse suurusega vahemikuga

Joonisel 4 on kujutatud neli usaldusdiagrammi, mis on saadud kasutades erinevaid vahemike paigutusviise ECE algoritmis. Igal diagrammil on kujutatud cw-ECE leidmine 2-klassilise klassifitseerija teise klassi jaoks 200 andmepunkti põhjal, kus andmed ja klassifitseerija on tehislikult koostatud vastavalt töö eksperimentaalses osas kirjeldatud katsele 2. Igal diagrammil on cw-ECE leitud samade andmete põhjal, kuid tulemused on erinevad: ECE teise klassi jaoks esimese rea esimesel graafikul on 0.044, teisel graafikul 0.108, teise rea esimesel graafikul 0.018 ning teisel graafikul 0.070. Esimeses reas asuvatel graafikutel on kujutatud kümne ning viiekümne võrdse laiusega vahemikuga ECE leidmine. Teises reas asuvatel graafikutel on kujutatud nelja ja viiekümne võrdse suurusega vahemikuga ECE leidmine.

1.3.3 Kaugusfunktsioon ECE algoritmis

ECE arvutamise algoritmi neljandal sammul on kaugusfunktsioonina kasutatud kaht viisi. Kull jt (2019), Guo jt (2017), Naeini jt (2015) on kasutanud absoluutkaugust $d_{abs}(y,p) = |p-y|$. Kumar jt (2019) on kasutanud ruutkaugust $d_{sq}(y,p) = (p-y)^2$. Widmann jt (2019), Vaicenavicius jt (2019) on kasutanud mõlemat viisi.

Bakalaureusetöös vaadatakse uudse lähenemisena võimaliku kaugusfunktsioonina ka logaritmkaugust, mis on inspireeritud Kullback-Leibleri kaugusest. Kullback-Leibleri kaugus mõõdab, kui suurel määral erinevad üksteisest juhuslikud suurused jaotustega Q ja P ning on defineeritud kui

$$d_{KL}(Q,P) = \sum_{x \in X} g\big(Q(x), P(x)\big), \text{ kus}$$

$$g(a,b) = \begin{cases} 0, & \text{kui } a = 0 \\ \infty, & \text{kui } a \neq 0 \; ja \; b = 0 \\ a \ln \dfrac{a}{b} & \text{muidu} \end{cases} \text{ (Lember, 2018).}$$

Täpsemalt mõõdab Kullback-Leibleri kaugus, kui ootamatu on, et juhusliku suuruse jaotus on Q, kui eeldati, et see on P (Lember, 2018).

ECE leidmise algoritmis on iga vahemiku jaoks võimalik vaadelda binaarset juhuslikku suurust, mille väärtus näitab, kas vahemikku kuuluv andmepunkt kuulub vaadeldavasse klassi või ei kuulu. Eeldatakse, et seda sündmust kirjeldab Bernoulli jaotus P, kus

positiivse sündmuse tõenäosus on keskmine klassifitseerija ennustus vahemikus ehk p ja negatiivse sündmuse tõenäosus on $1-p$. Tegelikult kirjeldab seda sündmust aga Bernoulli jaotus Q, kus positiivse sündmuse tõenäosus on tegelik vaadeldav osakaal vahemikus ehk y ja negatiivse sündmuse tõenäosus on $1-y$. Seega on iga vahemiku jaoks võimalik leida suuruste Q ja P Kullback-Leibleri kaugus

$$d_{KL}(Q,P) = \sum_{x \in X} g\big(Q(x), P(x)\big) = g(y,p) + g(1-y, 1-p).$$

Bakalaureusetöös käsitletav logaritmkaugus ongi defineeritud analoogselt kujul

$$d_{\log}(y,p) = g(y,p) + g(1-y, 1-p).$$

Kui mitte vaadata erijuhte, kus $y = 0$, $p = 0$, $y = 1$ või $p = 1$, siis logaritmkauguse lihtsustatud kuju on

$$d_{\log}(y,p) = y \ln \frac{y}{p} + (1-y) \ln \frac{1-y}{1-p}.$$

1.3.4 ECE väärtuse kasutamine

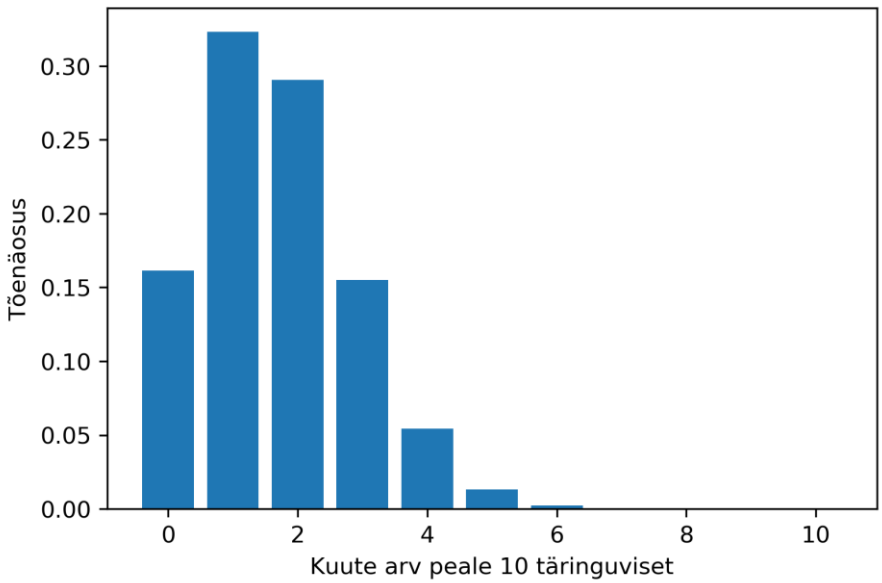
ECE väärtust kasutatakse teadusartiklites tihti mõõduna, kui kaugel kalibreeritusest klassifitseerija on. Näiteks on seda teinud Kull jt (2019), Vaicenavicius jt (2019), Guo jt (2017), Widmann jt (2019) ning Naeini jt (2015).

Samas kasutatakse ECE väärtust ka teistmoodi: selle põhjal on võimalik konstrueerida statistiline test, millega mingi olulisusnivoo korral kontrollida, kas kehtib nullhüpotees, et klassifitseerija on perfektselt kalibreeritud (Vaicenavicius jt, 2019). Bakalaureusetöös nimetatakse sellist statistilist testi kui kalibreerituse test. Käesolevas töös keskendutakse ECE algoritmile just kalibreerituse testi võimsuse parandamise kontekstis. Seetõttu kehtivad kõik töö katselises osas tehtud järeldused ECE arvutamise kohta ainult ECE arvutamisel kalibreerituse testi kontekstis ning ei pruugi kehtida juhtudel, kui ECEd kasutatakse mõõdikuna, kui kaugel kalibreeritusest klassifitseerija on.

1.4 Hüpoteeside statistiline testimine

Selleks, et mõista paremini ECE põhjal konstrueeritud statistilist testi, on kirjeldatud järgnevas peatükis hüpoteeside statistilist testimist Märt Mölsi (2013) loengu-materjalide põhjal.

Kui kellelgi on mingi protsessi toimumise kohta hüpotees, siis selle hüpoteesi paikapidamise tõenäosust on võimalik hinnata hüpoteesi statistilise testimisega. Esmalt sõnastatakse eeldatud nullhüpotees ning talle vastanduv alternatiivne hüpotees. Seejärel tehakse eeldus, et nullhüpotees kehtib ning see kirjeldab protsessi toimumist. Siis hinnatakse, milliseid väärtuseid ning missuguse tõenäosusega võidakse protsessi vaatlemisel saada, kui eeldatud nullhüpotees protsessi toimumist tõepoolest kirjeldab. Seejärel vaadeldakse protsessi tegelikku toimumist ning hinnatakse, kui tõenäoline on saada sellist või veel ekstreemsemat vaatlustulemust. Seda väärtust nimetatakse p-väärtuseks. Kui p-väärtus on väike ning väiksem kui enne testimist kehtestatud lubatud vea piir ehk olulisusnivoo, siis lükatakse nullhüpotees ümber ning tehakse järeldus, et valitud olulisusnivoo juures kehtib tõenäoliselt alternatiivne hüpotees. Vastasel juhul tehakse järeldus, et valitud olulisusnivoo juures ei anna vaatlustulemus alust nullhüpoteesi ümber lükata ning jäädakse nullhüpoteesi juurde.



Joonis 5. Eri tulemuste tõenäosus 10 järjestikusel täringuviskel

Näide 2. Vaatleme protsessina kümmet järjestikust täringuviset. Meil on kahtlus, et täringud võivad olla ebaausad ning anda tulemuseks liiga tihti numbrit 6. Seega soovime kontrollida nullhüpoteesi, et täringud on ausad ning et kuue tulemise tõenäosus on 1/6. Sõnastame alternatiivse hüpoteesi, et täringud on ebaausad ning kuue tulemise tõenäosus on suurem kui 1/6. Valime olulisusnivoo 0.05. Teame, et ausate täringute korral kirjeldab joonisel 5 kujutatu tõenäosuseid saada erinevaid koguseid numbrit 6. Seejärel vaatleme kümmet järjestikust täringuviset ning märgime üles vaatluse tulemuse. Olgu selleks tulemuseks kaheksa kuut kümnest viskest. Ausate täringute korral on aga sellise tulemuse saamine väga ebatõenäoline: vähemalt kaheksa kuue viskamise tõenäosus on umbes 0.00002. Kuna see tõenäosus ehk p-väärtus on väiksem kui meie valitud olulisusnivoo 0.05, siis lükkame nullhüpoteesi ümber ning teeme järelduse, et valitud olulisusnivoo korral kehtib tõenäoliselt alternatiivne hüpotees.

1.5 Kalibreerituse test

ECE väärtuse põhjal konstrueeritud kalibreerituse test töötab oma olemuselt samamoodi kui eelnevas peatükis kirjeldatud näide 2 täringutel. Vaicenavicius jt (2019) kirjeldavad sellist statistilist testi, millega kontrollida, kas klassifitseerija on enesekindluse järgi kalibreeritud, kasutades selleks nullhüpoteesi, et klassifitseerija on mitmeklassiliselt kalibreeritud. Mitmeklassiliselt kalibreeritud klassifitseerija on seda ka enesekindluse järgi ning tugevama nullhüpoteesi kasutamine võimaldab testi läbiviimiseks vajalikke andmeid genereerida (Kull jt, 2019). Kuna mitmeklassiliselt kalibreeritud klassifitseerija on kalibreeritud ka klassikaupa, siis on võimalik sellist statistilist testi kasutada ka selleks, et kontrollida, kas klassifitseerija on klassikaupa kalibreeritud (Kull jt, 2019). Järgnevalt on kirjeldatud kalibreerituse testi läbiviimist (Vaincenavicius jt, 2019):

1. Olgu meil n andmepunkti $x_1, \dots x_n$, mille kohta on teada nende tegelik märgend $y_1, \dots y_n$ ning olgu klassifitseerija ennustus iga andmepunkti jaoks vektor $p(x_i)$, $1 \leq i \leq n$.
2. Sõnastame nullhüpoteesi, et klassifitseerija on perfektselt mitmeklassiliselt kalibreeritud ehk et iga andmepunkti x_i märgend y_i on pärit jaotusest $p(x_i)$. Sellisel juhul on klassifitseerija ka perfektselt enesekindluse järgi ja perfektselt

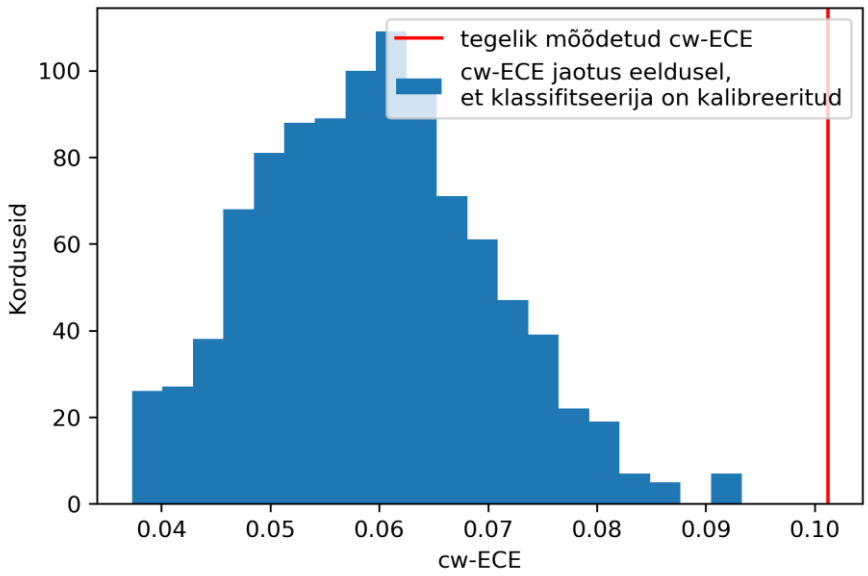
klassikaupa kalibreeritud. Sõnastame alternatiivse hüpoteesi, et klassifitseerija ei ole kalibreeritud.

3. Valime olulisusnivoo α ja oletame, et kehtib nullhüpotees.
4. Leiame ECE jaotuse klassifitseerija väljundite $p(x_i)$ korral eeldusel, et nullhüpotees kehtib. Selleks kordame m korda järgnevaid samme a ja b:
    a. Kasutame järjepideva taasvaliku (*consistency resampling*) meetodit, et genereerida andmepunktidele uued märgendid eeldusel, et klassifitseerija on kalibreeritud. See tähendab, et iga andmepunkti x_i jaoks valime jaotusest $p(x_i)$ juhuslikult uue märgendi y'_i.
    b. Leiame ECE väärtuse uute märgendite y'_i ja $p(x_i)$ põhjal, $1 \leq i \leq n$.
5. Leiame klassifitseerija väljundite $p(x_i)$ ja tegelike märgendite y_i põhjal ECE väärtuse e.
6. Hindame empiiriliselt tõenäosust p, et perfektselt kalibreeritud klassifitseerija korral saada vähemalt sama kõrge tulemus kui e. Selleks leiame proportsiooni sammul 4 vaadeldud juhtudest, mis on vähemalt sama kõrge väärtusega kui e.
7. Kasutame seda väärtust kui p-väärtust ning kui $p \leq \alpha$, siis lükkame ümber nullhüpoteesi, et klassifitseerija on perfektselt kalibreeritud. Vastasel juhul jääme nullhüpoteesi juurde.

Nii on võimalik cf-ECE väärtuse põhjal mingi statistilise olulisusega ümber lükata hüpotees, et klassifitseerija on enesekindluse järgi kalibreeritud, ning cw-ECE väärtuse põhjal hüpotees, et klassifitseerija on klassikaupa kalibreeritud (Vaicenavicius jt, 2019).

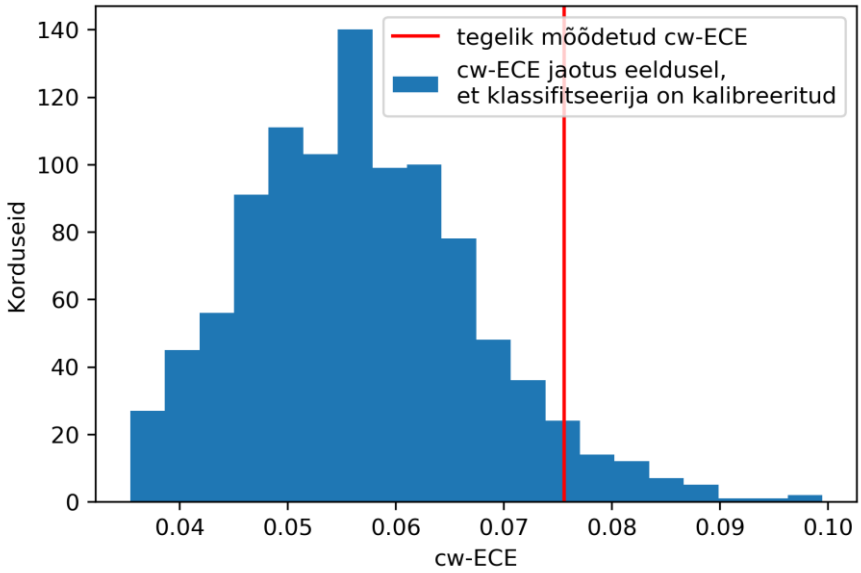
Näide 3. Vaatleme tehislikult koostatud möödakalibreeritud klassifitseerijat ja 100 juhuslikult valitud andmepunktiga andmestikku, mis on loodud vastavalt töö eksperimentaalses osas kirjeldatud katsele 1. Viime sellel klassifitseerijal läbi kalibreerituse testi. Olgu ECE leitud klassikaupa, absoluutkaugusega ning kümne võrdse laiusega vahemikuga. Valime olulisusnivoo 0.05. Eeldame, et kehtib nullhüpotees, et klassifitseerija on kalibreeritud. Hindame andmestikku kuuluva 100 andmepunkti põhjal cw-ECE jaotust eeldusel, et klassifitseerija on kalibreeritud, leides selleks 1000 väärtust sellest jaotusest. Leiame esialgsete märgendite põhjal cw-ECE ning hindame, kui tõenäone on saada nii ekstreemset tulemust. Sooritatud kalibreerituse testi tulemus on näha joonisel 6 kujutatud histogrammil. Arvuliselt saime p-väärtuseks 0. Kuna see on

väiksem kui valitud olulisusnivoo, siis lükkame nullhüpoteesi ümber ja järeldame, et klassifitseerija ei ole tõenäoliselt kalibreeritud.



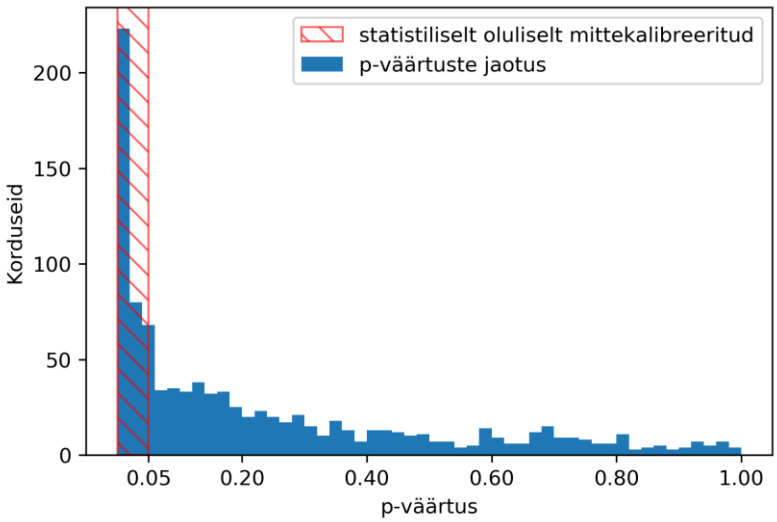
Joonis 6. Kalibreerituse testi näide

Samas sõltub ühe konkreetse ECE väärtus andmepunktidest ja nende märgenditest, mille põhjal seda arvutatakse (Vaicenavicius jt, 2019). Kuna praktikas on andmepunktid pärit juhuslikult kogu võimalikust sisendiruumist, siis ühe kindla ECE väärtuse puhul on tegemist juhusliku suurusega (Vaicenavicius jt, 2019). Korrates näites 3 tehtud kalibreerituse testi samast andmejaotusest võetud uutel andmetel, saame seetõttu uue ning tõenäoliselt teistsuguse tulemuse. Joonisel 7 on kujutatud näites 3 toodud kalibreerituse testi kordus, kus kõik muu on jäetud samaks, aga andmestik on valitud uus. Seekord on p-väärtus 0.055, mis on suurem kui valitud olulisusnivoo. Seega jääme nullhüpoteesi juurde, et klassifitseerija on kalibreeritud. Tegime testi põhjal vale järelduse, kuna tegelikult oli klassifitseerija möödakalibreeritud ning konstrueeritud töö eksperimentaalses osas katses 1 kasutatud viisil, mida on täpsemalt kirjeldatud peatükis 2.3.3.



Joonis 7. Kalibreerituse testi näite kordus uutel andmetel

On selge, et praktikas tahaksime, et kalibreerituse test oleks võimalikult usaldusväärne ning tegelikult möödakalibreeritud klassifitseerija korral jõuaksime võimalikult harva järelduseni, et tõenäoliselt on klassifitseerija kalibreeritud. See tähendab, et kalibreerituse test oleks võimalikult võimas: tõenäosus lugeda tõestatuks alternatiivne hüpotees oleks võimalikult kõrge, kui tegelikult kehtibki alternatiivne hüpotees (Möls, 2013). Korrates kalibreerituse testi alati uute andmetega on võimalik koostada nende testide p-väärtuste jaotus. Selle jaotuse põhjal saab omakorda hinnata testi võimsust. Selline kalibreerituse testi kordamine testi võimsuse hindamiseks on töös võimalik, kuna testi sooritatakse sünteetilistel andmetel, mistõttu on andmeid võimalik piiramatult genereerida. Joonisel 8 on kujutatud näites 3 sooritatud kalibreerituse testi 1000 korduse p-väärtuste jaotus, kus igas testis on kasutatud uut andmestikku. Joonisel kujutatud jaotusest saab hinnata, kui suur osa testidest viivad valitud olulisusnivoo juures vale järelduseni. Olulisusnivoo 0.05 korral vaatame jaotuse kvantiili kohal 0.05: joonisel 8 on see 0.34-kvantiil ehk 34% kordadest sooritame kalibreerituse testis korrektse järelduse. Seega on selle kalibreerituse testi võimsus olulisusnivoo 0.05 jaoks 0.34. Kalibreerituse testi võimsust on võimalik mõjutada muutes ECE arvutamisel kasutatavaid parameetreid: kaugusfunktsiooni, vahemike paigutusviisi, vahemike arvu.



Joonis 8. Tuhande kalibreerituse testi tulemuseks olevate p-väärtuste jaotus

Kalibreerituse testi võimsuse mõõtmiseks vaatasime selle testi korduva sooritamise tulemuseks oleva p-väärtuste jaotuse kvantiili väärtust usaldusnivoo väärtuse kohal. Joonisel 8 oli selleks triibutatud alasse ehk vahemikku $[0.0, 0.05]$ jäävate väärtuste osakaal.

Sõltuvalt klassifitseerija olemusest ja vahemike paigutusviisist võib aga juhtuda, et kalibreeritud klassifitseerija jaoks leitud ECE on keskmiselt suurema väärtusega kui tegelik kalibreerimata klassifitseerija jaoks leitud ECE, mis võib esialgu tunduda ebaintuitiivne. Et näitlikustada sellist ebaintuitiivset olukorda, kus kalibreeritud klassifitseerija ECE on keskmiselt suurem, vaatame järgnevat näidet.

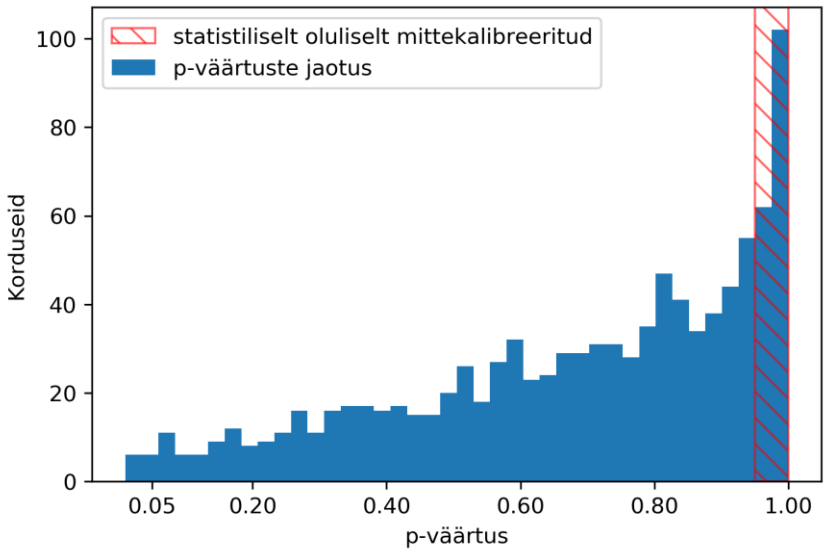
Näide 4. Oletame, et meil on 2-klassiline konstantne klassifitseerija, mis ennustab alati tõenäosust $(0.8, 0.2)$. Olgu iga andmepunkti tegelik märgend aga pärit jaotusest $(0.9, 0.1)$. Selliselt on 90% juhtudest tegelik märgend $(1, 0)$ ning 10% juhtudest tegelik märgend $(0, 1)$. Taasvaliku meetodiga valitud perfektselt kalibreeritud klassifitseerija märgendid on aga 80% kordadest $(1, 0)$ ja 20% kordadest $(0, 1)$. Olgu ECE leitud klassikaupa absoluutkaugusega nii, et iga andmepunkt on eraldi vahemikus. Seega tegelik ECE on keskmiselt

$$ECE_{tegelik} = 0.5 \cdot \big(0.9 \cdot (|1 - 0.8| + |0 - 0.2|) + 0.1 \cdot (|0 - 0.8| + |1 - 0.2|)\big) = 0.26.$$

Perfektselt kalibreeritud klassifitseerija jaoks on ECE aga keskmiselt

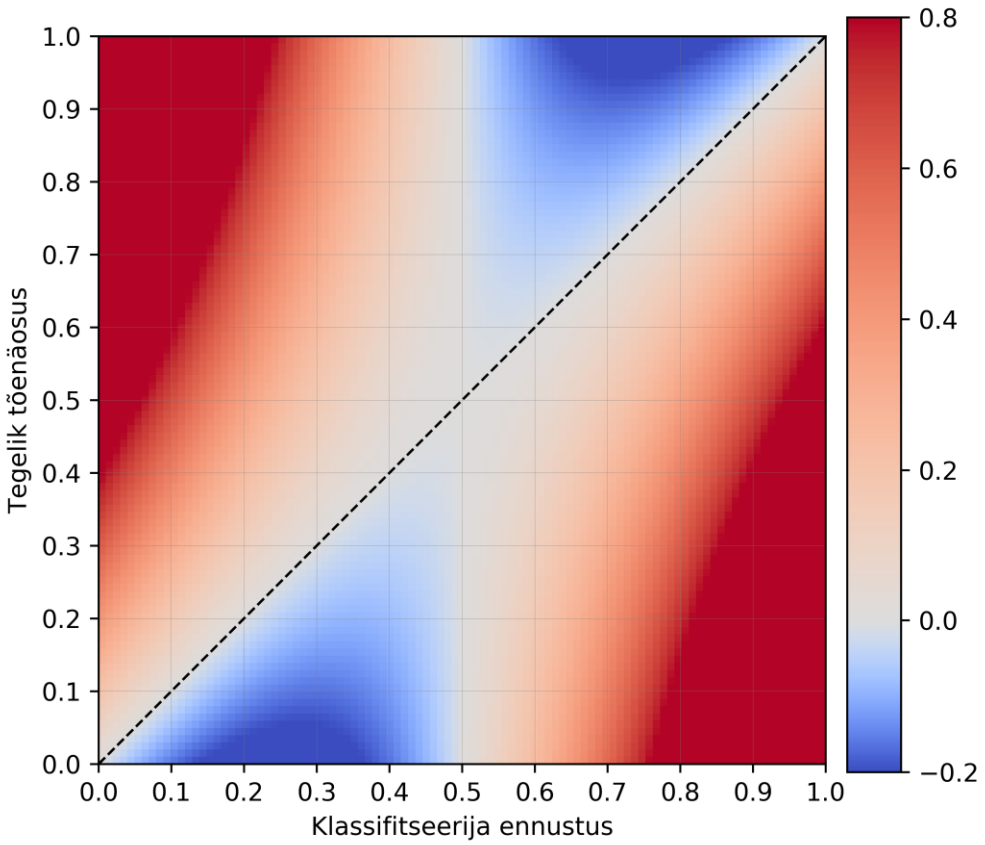
$$ECE_{perf} = 0.5 \cdot \big(0.8 \cdot (|1 - 0.8| + |0 - 0.2|) + 0.2 \cdot (|0 - 0.8| + |1 - 0.2|)\big) = 0.32.$$

Näitest on näha, et sellisel juhul on tegelik ECE keskmiselt väiksem kui kalibreeritud klassifitseerija ECE.



Joonis 9. Kalibreerituse testide tulemuseks olevate p-väärtuste jaotus, kui tegelik ECE on keskmiselt madalam kui kalibreeritud klassifitseerija ECE

Kui tegelik ECE on keskmiselt madalam kui kalibreeritud klassifitseerija ECE, siis kalibreerituse testide korduvast jooksutamisest saadud p-väärtuste jaotus on teistpidise kaldega kui joonisel 8 ning testi võimsuse hindamiseks tuleb vaadata hoopis vahemikku [0.95, 1.0] jäävate väärtuste osakaalu. See tähendab väärtust $1 - q$, kus q on kvantiil kohal (1 − usaldusnivoo). Joonisel 9 on kujutatud näide p-väärtuste jaotusest juhul, kui möödakalibreeritud klassifitseerija tegelik ECE on keskmiselt madalam kui kalibreeritud klassifitseerija ECE ning testi võimsuse hindamiseks peab usaldusnivoo 0.05 korral vaatama väärtust 1 − (kvantiil kohal 0.95). Joonisel kujutatud p-väärtuste jaotus on saadud korrates 500 andmepunktiga kalibreerituse testi klassifitseerijal, mis on loodud vastavalt töö eksperimentaalses osas katses 9 kirjeldatule. ECE on leitud klassikaupa, enesekindluse järgi ning iga andmepunkt on eraldi vahemikus. Kalibreerituse testis on kalibreeritud klassifitseerija ECEde jaotuses 200 väärtust ning joonisel kujutatud p-väärtuste jaotuses on 1000 testi tulemus. Alternatiivselt oleks võimalik kalibreerituse testi vaadata kahepoolse statistilise testina, mis juhul ei tuleks sõltuvalt olukorrast testi võimsuse hindamiseks vaadata kaht erinevat väärtust.



Joonis 10. Keskmine vahe tegeliku cw-ECE ja taasvaliku meetodiga leitud cw-ECE vahel 2-klassilise klassifitseerija jaoks ühe andmepunkti põhjal, kus ECE on leitud absoluutkaugusega

Kaheklassilise klassifitseerija jaoks on võimalik ühe andmepunkti põhjal leitud ECE korral visualiseerida, millistel juhtudel on tegelik ECE keskmiselt väiksem kalibreeritud klassifitseerija ECEst. Joonisel 10 on seda tehtud absoluutkaugusega leitud cw-ECE jaoks. Joonise x-teljel on kujutatud andmepunkti jaoks ennustatud tõenäosus esimesse klassi kuuluda ning y-teljel andmepunkti tegelik tõenäosus esimesse klassi kuuluda. Jooniselt on näha, et 2-klassilise klassifitseerija korral, kui iga andmepunkt on paigutatud eraldi vahemikku, on tegelik cw-ECE keskmiselt väiksem olukordades, kus klassifitseerija on liiga ebakindel: kõrgeima tõenäosusega klassi ennustus on madalam kui tegelik tõenäosus. Tegelik cw-ECE on keskmiselt suurem olukordades, kus klassifitseerija on liiga enesekindel: kõrgeima tõenäosusega klassi ennustus on liiga kõrge.

2 Optimaalsete parameetrite otsing

Selles töö osas kirjeldatakse läbi viidud katseid ning nende tulemusi.

2.1 Töö eesmärk

Bakalaureusetöö eesmärk on kalibreerituse testi võimsuse suurendamine. Selleks otsitakse, millist vahemike paigutusviisi, vahemike arvu ning kaugusfunktsiooni peaks cf-ECE ja cw-ECE leidmisel kasutama, kui tahetakse testida, kas klassifitseerija on vastavalt enesekindluse järgi kalibreeritud või klassikaupa kalibreeritud.

2.2 Kasutatud tarkvara

Katsete tegemiseks kasutati Pythoni versiooni 3.6 ning teekide NumPy versiooni 1.16.5 ja scikit-learn versiooni 0.21.2. Visualiseeringute loomiseks kasutati teeki matplotlib. Koodirepositoorium on toodud töö lisas I.

2.3 Katsemeetodi kirjeldus

Kindla parameetrite kombinatsiooni hindamiseks koostati p-väärtuste jaotus kalibreerituse testide korduvast jooksutamisest. Kalibreerituse testi võimsust hinnati olulisusnivoo 0.05 jaoks, kuna 0.05 on praktikas tihti kasutust leidev olulisusnivoo (Möls, 2013). See tähendab, et võimsuse hindamiseks vaadati p-väärtuste jaotuse kvantiili väärtust kohal 0.05 või mõnel juhul ka väärtust $1 - q$, kus q on kvantiil kohal 0.95, kui katse käigus selgus, et möödakalibreeritud klassifitseerija tegelik ECE on keskmiselt madalam kui kalibreeritud klassifitseerija ECE. Alternatiivina oleks võimalik kalibreerituse testi sooritada kahepoolse statistilise testina, mille korral saaks vältida kahe erineva väärtuse vaatamist sõltuvalt olukorrast.

Katsetes leiti klassifitseerija ECE 100 andmepunkti alusel, kalibreerituse testis koosnes perfektselt kalibreeritud klassifitseerija ECEde jaotus 100 ECEst, p-väärtuste jaotus koosnes 1000 kalibreerituse testi tulemusest. Kuna töös läbi viidud katsed on

arvutusmahukad ning seetõttu ka ajamahukad, siis jaotuste suurused valiti kompromissina katsete jooksmiseks kuluva aja ning valimi väiksusest tuleneva mõõtmisvea vahel. Suuremate valimite korral saaks veel pisut täpsemaid tulemusi, kuid kuluks ka rohkem aega.

Katseid sooritati cf-ECE ja cw-ECE jaoks absoluut-, ruut- ja logaritmkaugusega. Katseid sooritati nii võrdse suurusega vahemikega kui ka võrdse laiusega vahemikega. Võrdse suurusega vahemikega sooritati katseid vahemike arvudega 5, 10, 15, ... 95, 100. Võrdse laiusega vahemikega sooritati katseid vahemike arvudega 5, 10, 15, 20, 25, 50, 75, 100, 125, 150, 175, 200, 500. Vahemike arvud valiti lähtuvalt katsetes kasutatud andmepunktide arvust. Kuna katseid sooritati 100 andmepunkti põhjal, siis ei ole võrdse suurusega vahemike korral mõtet võtta rohkem kui 100 vahemikku, kuna 100 vahemiku korral on juba iga andmepunkt eraldi vahemikus. Võrdse laiusega vahemike korral valiti tihedamalt väikeseid vahemike arvi, kuna varasemas teaduskirjanduses on ECE leidmises tavaks valida pigem vähe vahemikke. Võrdse laiusega vahemike korral hakkab väga suurte vahemike arvude korral andmepunktide jaotumine vahemikesse muutuma sarnaseks juhule, kus kasutatakse võrdse suurusega vahemikke, mistõttu valiti maksimaalseks vahemike arvuks võrdse laiusega vahemike korral 500.

2.3.1 Katsetes kasutatud klassifitseerijad

Katseid viidi läbi üheksa erineva klassifitseerija jaoks, millest seitse olid koostatud tehislikel andmetel. Katsetes kasutatud klassifitseerijate kirjeldused on toodud tabelis 2. Tehislikel andmetel koostatud klassifitseerijate loomise meetodid on seletatud lahti järgnevates peatükkides 2.3.3 ja 2.3.4. Katses 7 kasutati Covertype andmestikul[*] treenitud otsustuspuud ning katses 8 kasutati Covertype andmestikul treenitud otsustusmetsa. Otsustuspuu õigsus testandmetel on 84% ja otsustusmetsa õigsus testandmetel on 95%. Täpsem info otsustuspuu ja -metsa treenimise kohta on toodud töö lisas II.

[*] https://archive.ics.uci.edu/ml/datasets/Covertype

Tabel 2. Katsetes kasutatud klassifitseerijad

Katse number	Klasside arv	Klassifitseerija loomise meetod	Meetodis kasutatud jaotus	Meetodi parameeter β
1	2	Meetod 1	Dir(0.1, 0.1)	0.95
2	2	Meetod 1	Dir(0.2, 0.18)	0.95
3	2	Meetod 2	Dir(0.1, 0.1)	0.30
4	5	Meetod 1	Dir(0.1, 0.1, … ,0.1)	0.95
5	5	Meetod 1	Dir(0.2, 0.18, 0.16, 0.14, 0.12)	0.95
6	5	Meetod 2	Dir(0.1, 0.1, … ,0.1)	0.30
7	7	Otsustuspuu Covertype andmestikul	-	-
8	7	Otsustusmets Covertype andmestikul	-	-
9	10	Meetod 2	Dir(0.1, 0.1, … ,0.1)	-0.1

Suurem osa katseid viidi läbi tehislikult koostatud klassifitseerijatel kahel põhjusel. Esiteks nõudis katsete läbiviimine suurt andmestikku: kui ECE on leitud 100 andmepunkti põhjal ning p-väärtuste jaotuses on 1000 kalibreerituse testi tulemus, siis testi võimsuse hindamiseks on tarvis $100 \cdot 1000 = 10^5$ andmepunkti. Lisaks on reaalsetel andmetel treenitud klassifitseerijate korral tarvis ka andmeid klassifitseerija treenimiseks. Teiseks taheti määrata, kui möödakalibreeritud ning millisel viisil möödakalibreeritud klassifitseerijad on, et katsete põhjal tehtud järeldused oleksid võimalikult informatiivsed.

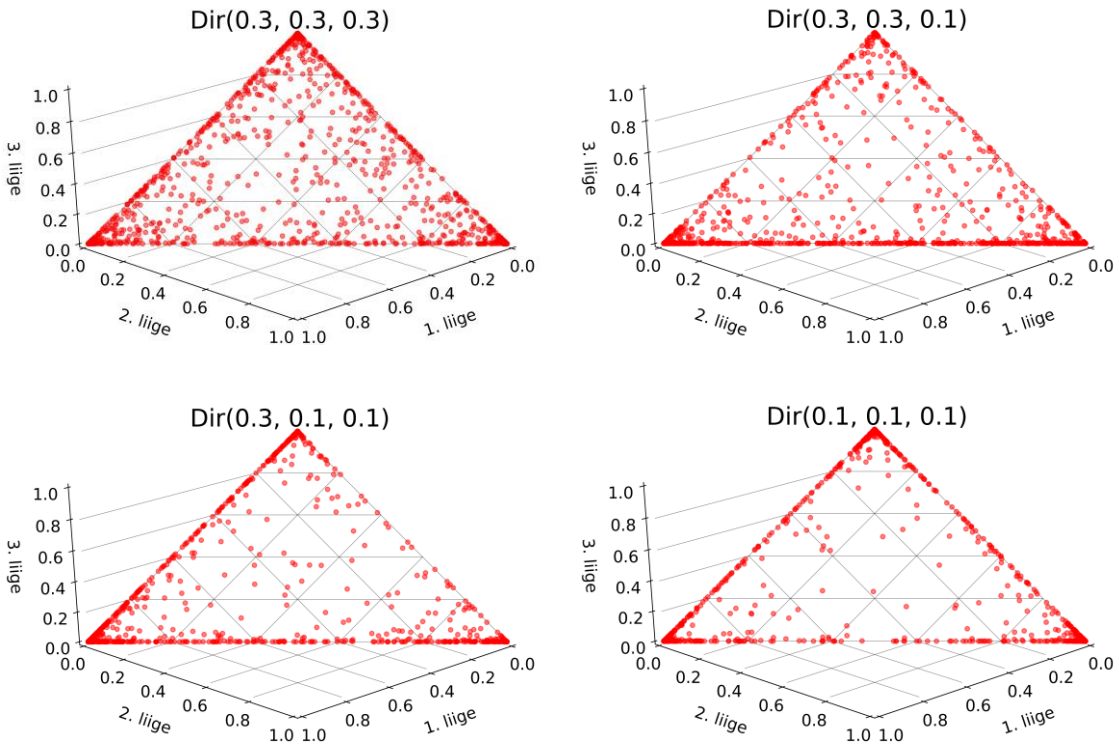
Katsetes 1-6 on tehislikult loodud klassifitseerija liiga enesekindel, katses 9 liiga ebakindel. Katseid viidi läbi rohkematel enesekindlatel klassifitseerijatel kui ebakindlatel, kuna praktikas on klassifitseerijad pigem liialt enesekindlad (Kull jt, 2019). Katsetes kasutatud tehislikult koostatud klassifitseerijate loomise parameetrid valiti nii, et imiteerida reaalsetel andmetel treenitud klassifitseerijate ennustuste jaotuseid, mida on kirjeldanud näiteks Vaicenavicius jt (2019). Katseid viidi läbi võimalikult paljudel erinevatel klassifitseerijatel, et näha, kas katses kasutatud

parameetrite kombinatsiooni võimsus sõltub viisist, kuidas möödakalibreeritus klassifitseerija ennustustes esineb, ning kas võimsus sõltub klassifitseerija klasside arvust või klassifitseerija ennustuste jaotusest. Katsete tulemuste põhjal tehtud järeldused on seda usaldusväärsemad, mida rohkematel erinevatel klassifitseerijatel neid kinnitada. Seetõttu on töös kasutatud katsete arv valitud kompromissina katsete sooritamiseks kuluva ajaga.

Kuna mõlemas tehisliku klassifitseerija loomise meetodis kasutatakse ennustuste genereerimiseks Dirichlet' jaotust, siis enne nende meetodite kirjeldust kirjeldatakse järgnevas peatükis Dirichlet' jaotust.

2.3.2 Dirichlet' jaotus

Järgnev Dirichlet' jaotuse kirjeldus põhineb Frigyik, Kapila ja Gupta (2010) artiklil. Dirichlet' jaotus kirjeldab k-liikmeliste vektorite jaotust, kus vektori iga liige on mittenegatiivne ning vektori summa on 1. Ühte k-klassilise klassifitseerija ennustust saab vaadata kui sellist vektorit. See tähendab, et Dirichlet' jaotusega on võimalik kirjeldada klassifitseerija ennustuste jaotust.



Joonis 11. 1000 juhuslikult valitud vektorit nelja Dirichlet' jaotuse jaoks

Dirichlet' jaotuse $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ määravad tema parameetrid $\alpha_1, \dots, \alpha_k$. Ühte k parameetriga Dirichlet' jaotust saab kirjeldada $(k-1)$-dimensioonilise kujundiga. Näiteks joonisel 11 on kujutatud nelja erineva kolme parameetriga Dirichlet' jaotuse jaoks 1000 juhuslikult valitud vektorit nendest jaotustest. Kui Dirichlet' jaotuse parameetrid on $\alpha_1, \dots, \alpha_k < 1$, siis on jaotuse tihedus koondunud seda jaotust kirjeldava kujundi tippude ja servade ümber, nagu on näha joonisel 11. Kui parameetrid $\alpha_1, \dots, \alpha_k$ on võrdsed, siis on jaotus sümmeetriline. Kui $\alpha_1, \dots, \alpha_k < 1$, kuid parameetrid ei ole võrdsed, siis on ennustused koondunud rohkem nende tippude ümber, mille jaoks parameeter α on suurem. Töös kasutatakse vaid selliste parameetritega Dirichlet' jaotusi, mille jaoks $\alpha_1, \dots, \alpha_k < 1$, et imiteerida reaalsetel andmetel treenitud klassifitseerijate ennustuste jaotuseid.

2.3.3 Tehisliku klassifitseerija loomise meetod 1

Esimene meetod, millega analoogset on kasutanud ka Widmann jt (2019) tehisliku klassifitseerija loomiseks, võtab sisendiks mingi Dirichlet' jaotuse $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ ja parameetri $\beta \in [0,1]$. Meetodi käigus vaadeldakse kujuteldavaid andmepunkte ning genereeritakse neile tegelikud märgendid ning klassifitseerija ennustused. Andmepunktide tunnuseid ei genereerita, kuna seda pole ECE leidmiseks tarvis.

Klassifitseerija ennustused genereeritakse jaotusest $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ ning andmepunktide märgendid parameetri β põhjal. Kui klassifitseerija väljastab ennustused k andmepunkti jaoks, siis neist $\beta \cdot k$ andmepunkti jaoks genereeritakse märgendid kalibreeritult ehk jaotusest, mida klassifitseerija ennustas, ning ülejäänud $(1-\beta) \cdot k$ andmepunkti jaoks valitakse märgendid ühtlaselt juhuslikult võrdse tõenäosusega $\frac{1}{k}$ iga klassi jaoks.

Nii on parameetri $\beta = 1$ korral loodud klassifitseerija perfektselt kalibreeritud. Parameetri $0 \leq \beta < 1$ korral loodud klassifitseerija on möödakalibreeritud, vastavalt seda rohkem, mida väiksem on β. Kuna jaotusest $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ valitud vektori suurim tõenäosus on alati $\geq \frac{1}{k}$, siis on meetodi väljundiks olev klassifitseerija alati liiga enesekindel: kõrgeima tõenäosusega klassi ennustus on liiga kõrge. Meetodiga 1 loodud klassifitseerijad on töös kasutatud meetodi parameetrite korral enamiku

andmepunktide jaoks kalibreeritud, kuid ülejäänud andmepunktide jaoks suurel määral liiga enesekindlad oma ennustuses.

2.3.4 Tehisliku klassifitseerija loomise meetod 2

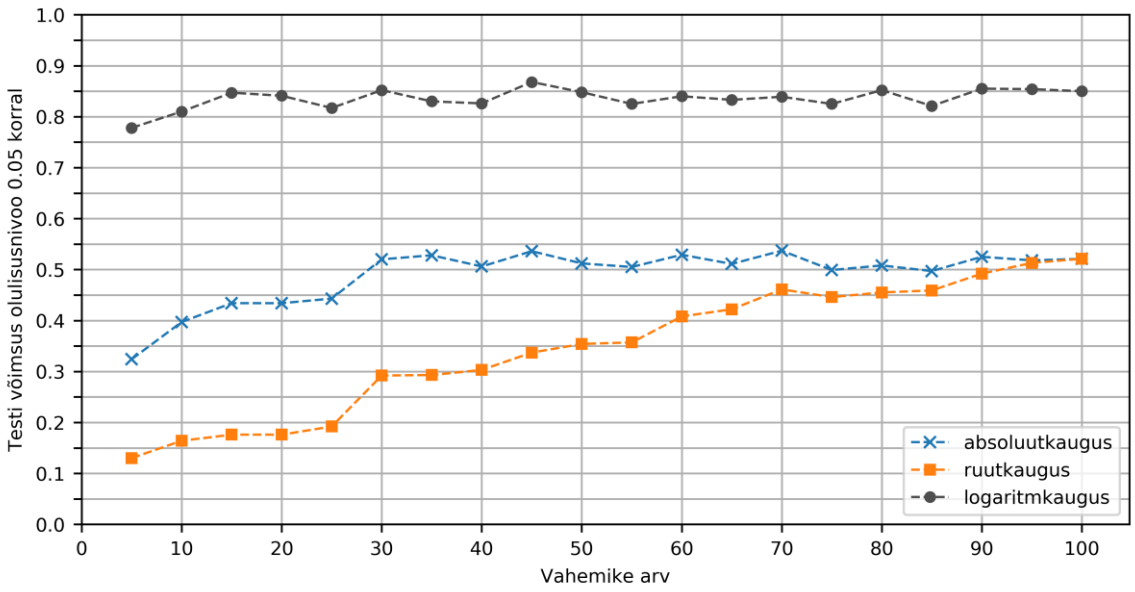
Teine meetod, mida töös kasutatakse, võtab sisendiks mingi Dirichlet' jaotuse $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ ja parameetri $\beta \in (-1, \infty]$. Meetodi käigus genereeritakse sarnaselt meetodiga 1 vaid kujuteldavate andmepunktide tegelikud märgendid ja klassifitseerija ennustused ning andmepunktide tunnuseid ei genereerita.

Esmalt genereeritakse andmepunktide jaoks jaotusest $\mathrm{Dir}(\alpha_1, \dots, \alpha_k)$ märgendite tõenäosused ning valitakse nendest tõenäosustest juhuslikult märgendid. Klassifitseerija ennustused andmepunktidele saadakse märgendite tõenäosuste modifitseerimisest. Iga andmepunkti märgendi jaoks genereeritud tõenäosusjaotuse suurimale tõenäosusele liidetakse parameeter β ning modifitseeritud tõenäosusvektori iga liige jagatakse vektori summaga, et tõenäosuste summa oleks endiselt 1. Näiteks, kui andmepunkti märgendile on genereeritud tõenäosusjaotus $(0.8, 0.16, 0.04)$ ning meetodi parameeter on $\beta = 1$, siis klassifitseerija ennustus sellele andmepunktile on $\left(\frac{0.8+1}{2}, \frac{0.16}{2}, \frac{0.04}{2}\right) = (0.9, 0.08, 0.02)$.

Kui $\beta = 0$, siis on klassifitseerija perfektselt kalibreeritud. Iga $\beta > 0$ korral saadakse klassifitseerija, mis on liiga enesekindel: kõrgeima tõenäosusega klassi ennustus on liiga kõrge. Iga $\beta < 0$ korral saadakse klassifitseerija, mis on liiga ebakindel: kõrgeima tõenäosusega klassi ennustus on liiga madal. Meetod 2 töötab, kui parameeter β ei ole väiksem kui vähim võimalik maksimaalne ennustus $\frac{1}{k}$, sest muidu võib saada tõenäosusvektoris negatiivse tulemuse. Töös valitud parameetrite korral on seda järgitud: katses 9 on meetodi parameeter -0.1 ning klassifitseerija vähim võimalik maksimaalne ennustus on 0.1, kuna klassifitseerija on kümneklassiline. Meetodiga 2 loodud klassifitseerijad on töös kasutatud meetodi parameetrite korral iga andmepunkti jaoks vähesel määral möödakalibreeritud.

2.4 Tulemuste analüüs cw-ECE jaoks

Joonisel 12 on kujutatud katse 1 tulemused cw-ECE jaoks võrdse suurusega vahemike korral. Joonise x-teljel on kujutatud vahemike arv ning y-teljel kalibreerituse testi võimsus olulisusnivoo 0.05 korral. Kuna katsetes arvutati ECE saja andmepunkti põhjal, siis kümne võrdse suurusega vahemiku korral on igas vahemikus kümme andmepunkti, 100 vahemiku korral on aga iga andmepunkt eraldi vahemikus. Jooniselt on näha, et selles katses oli kalibreerituse test kõige võimsam logaritmkaugusega leitud ECE korral ning testi võimsus ei sõltunud logaritmkauguse jaoks suuresti vahemike arvust. Ruutkauguse jaoks oli võimsaim tulemus 100 vahemiku korral ning absoluutkauguse jaoks läksid tulemused vahemike arvu suurendamisel paremaks kuni 30 vahemikuni.

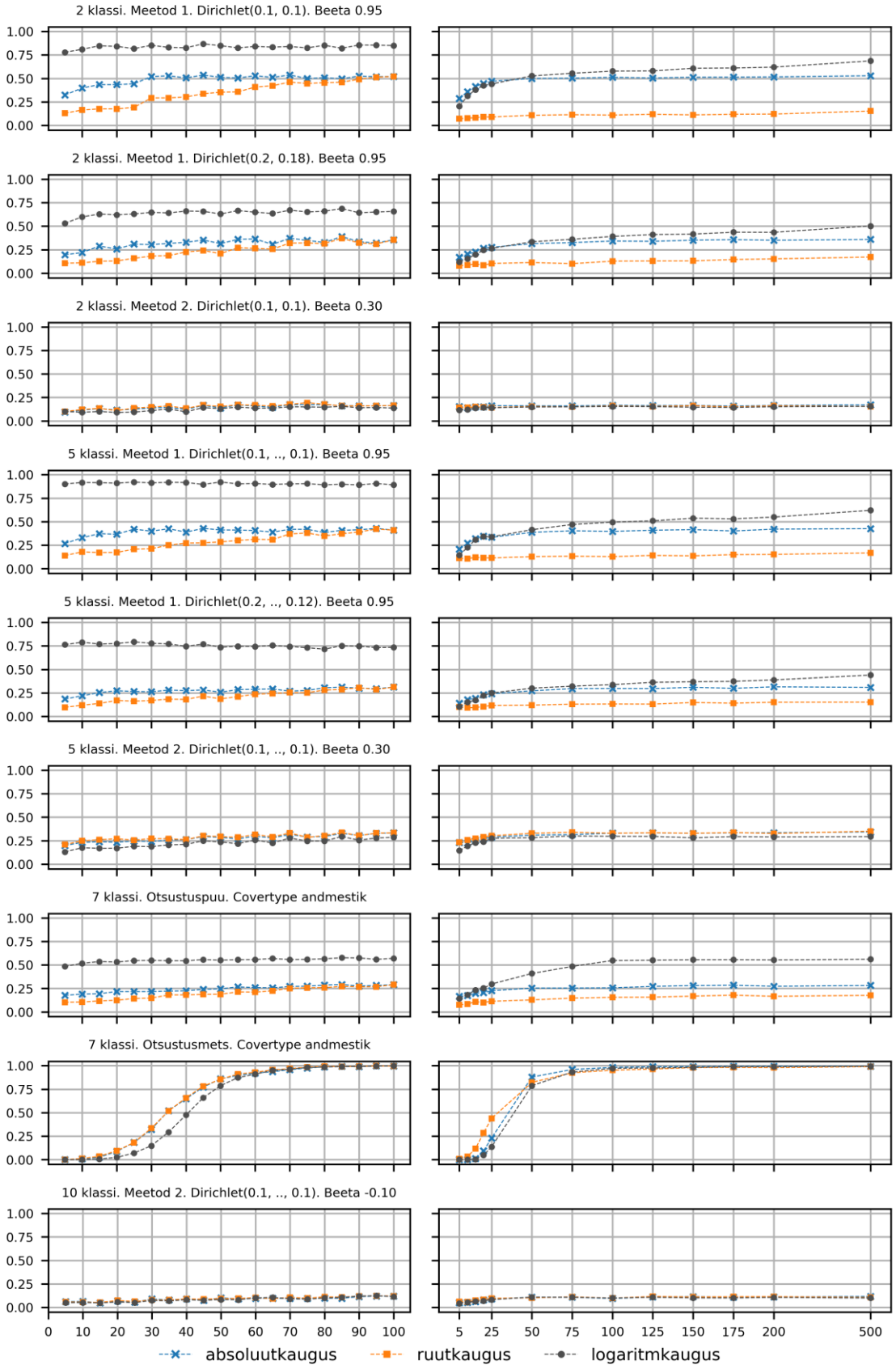


Joonis 12. Katse 1 tulemused cw-ECE jaoks võrdse suurusega vahemike korral

Joonisel 13 on kujutatud kõigi katsete tulemused cw-ECE jaoks. Esimeses veerus olevatel graafikutel on kujutatud tulemused võrdse suurusega vahemike jaoks, teises veerus olevatel graafikutel on kujutatud tulemused võrdse laiusega vahemike jaoks. Graafikute x-teljel on kujutatud vahemike arv ning y-teljel kalibreerituse testi võimsus olulisusnivoo 0.05 korral. Katsetes 8 ja 9 oli tegelik ECE keskmiselt madalam kui kalibreeritud klassifitseerija ECE.

Mõlema vahemike paigutusviisi jaoks selgub, et kalibreerituse test läheb võimsamaks, mida rohkem vahemikke valida. See viitab, et kalibreerituse test on seda võimsam, mida

vähem andmepunkte ühte vahemikku langeb. Seega peaks iga andmepunkti paigutama eraldi vahemikku. See on üllatav tulemus, kuna vaadates peatükis 1.3.2 toodud usaldusdiagrammide näiteid võrdse laiusega vahemike jaoks, saab visuaalselt oluliselt selgema arusaama klassifitseerija kalibreerituse kohta just väiksema vahemike arvuga leitud ECE korral.



Joonis 13. Testitulemused cw-ECE jaoks. Vasakul on tulemused võrdse suurusega vahemike jaoks, paremal on tulemused võrdse laiusega vahemike jaoks. X-teljel on vahemike arv. Y-teljel testi võimsus olulisusnivoo 0.05 korral

Katsete tulemustest on näha, et absoluut- ja ruutkaugusega sooritatud kalibreerituse testi võimsus on võrdne, kui iga andmepunkt paigutub eraldi vahemikku. Testide võrdset võimsust sellisel juhul põhjendab järgnev teoreem, mille tõestus on toodud töö lisas III.

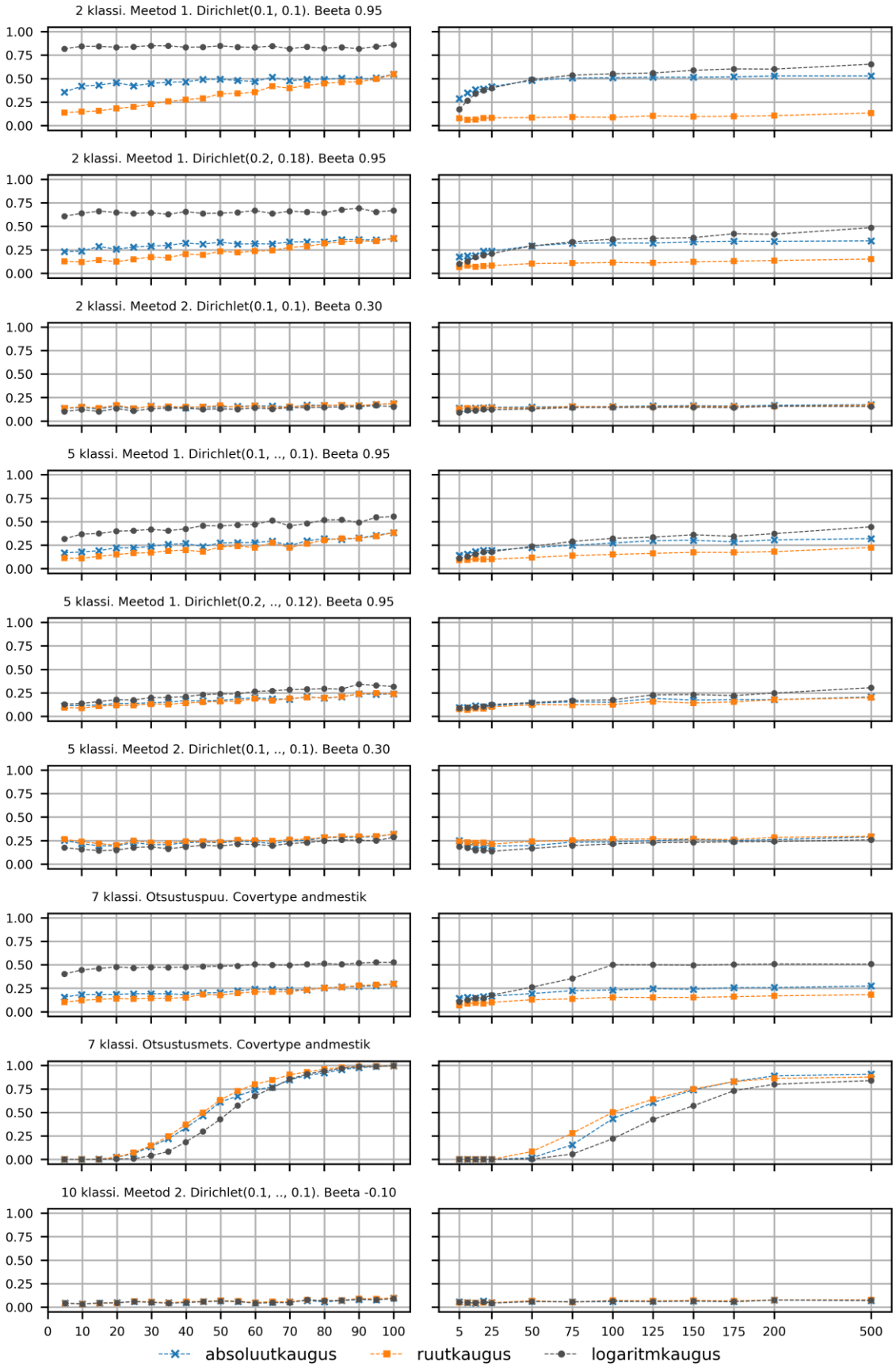
Teoreem 1. Kui cw-ECE leidmises on iga andmepunkt eraldi vahemikus, siis tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe on ruutkaugusega ja absoluutkaugusega mõõdetud ECE puhul võrdne.

Logaritmkaugus on meetodiga 1 loodud klassifitseerijate ning otsustuspuu korral oluliselt parem kui ruut- või absoluutkaugus. Muudel juhtudel on kaugusfunktsioonid ligikaudu võrdsed või logaritmkaugus on väga vähesel määral halvem. Seega on logaritmkaugus oluliselt parem olukordades, kus klassifitseerija ennustuste seas esineb suurel määral möödakalibreeritud andmepunkte. Olukordades, kus klassifitseerija ennustused on kõik vähesel määral möödakalibreeritud, on kaugusfunktsioonid ligikaudu võrdsed.

2.5 Tulemuste analüüs cf-ECE jaoks

Joonisel 14 on kujutatud kõigi katsete tulemused cf-ECE jaoks. Esimeses veerus olevatel graafikutel on kujutatud tulemused võrdse suurusega vahemike jaoks, teises veerus olevatel graafikutel on kujutatud tulemused võrdse laiusega vahemike jaoks. Graafikute x-teljel on kujutatud vahemike arv ning y-teljel kalibreerituse testi võimsus olulisusnivoo 0.05 korral. Katsetes 8 ja 9 oli tegelik ECE keskmiselt madalam, kui kalibreeritud klassifitseerija ECE.

Tulemustest on näha, et kalibreerituse test muutub võimsamaks, mida rohkem vahemikke valida ning mida vähem andmepunkte ühte vahemikku satub. Seega ka cf-ECE puhul on mõistlikum paigutada iga andmepunkt eraldi vahemikku.



Joonis 14. Testitulemused cf-ECE jaoks. Vasakul on tulemused võrdse suurusega vahemike jaoks, paremal on tulemused võrdse laiusega vahemike jaoks. X-teljel on vahemike arv. Y-teljel testi võimsus olulisusnivoo 0.05 korral

Analoogselt cw-ECEga on näha, et kui iga andmepunkt paigutada eraldi vahemikku, siis on absoluut- ja ruutkaugus võrdsed. Testide võrdset võimsust põhjendab järgnev teoreem, mille tõestus on toodud töö lisas III.

Teoreem 2. Kui cf-ECE leidmises on iga andmepunkt eraldi vahemikus, siis tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe on ruutkaugusega ja absoluutkaugusega mõõdetud ECE puhul võrdne.

Meetodiga 1 loodud klassifitseerijatel ja otsustuspuul oli logaritmkaugus parem kui muud kaugused. Muudel juhtudel on kaugusfunktsioonid ligikaudu võrdsed või on logaritmkaugus väga vähesel määral halvem. See on sarnane cw-ECE puhul leitud tulemustega.

Kokkuvõttes võib nii cf-ECE kui ka cw-ECE jaoks katsete tulemustest järeldada, et kalibreerituse testi võimsust on võimalik märkimisväärselt suurendada, kui paigutada iga andmepunkt eraldi vahemikku. Samuti, kui iga andmepunkt paigutada eraldi vahemikku, siis ei ole vahet, kas kasutada absoluut- või ruutkaugust. Tulemustest meetodiga 1 loodud klassifitseerijatel võib järeldada, et nii cf-ECE kui ka cw-ECE jaoks võib töös kasutatud logaritmkaugus testi võimsust märkimisväärselt suurendada, kui esineb andmepunkte, mille jaoks on klassifitseerija ennustus väga suurel määral möödakalibreeritud. Tulemustest meetodiga 2 loodud klassifitseerijatel võib järeldada, et kui klassifitseerija ennustused on kõik vähesel määral möödakalibreeritud, on tulemused erinevatel kaugusfunktsioonidel ligikaudu võrdsed või on logaritmkaugus väga vähesel määral halvem.

Kuna bakalaureusetöös uuriti vaid kaht sümmeetrilist vahemike paigutusviisi, siis uurida saaks veel muid võimalikke ebasümmeetrilisi vahemike paigutusviise. See tähendab viise, kus andmepunktide arv vahemikus või vahemiku laius sõltub vahemiku asukohast tõenäosusjaotuses. Uurida saaks veel muude võimalike kaugusfunktsioonide mõju kalibreerituse testi võimsusele. Töös sooritatud katsete tulemusi saaks kinnitada rohkematel klassifitseerijatel.

Kokkuvõte

Bakalaureusetöös otsiti, milline parameetrite kombinatsioon on cf-ECE ja cw-ECE arvutamisel optimaalne, et suurendada ECE põhjal sooritatud kalibreerituse testi võimsust. Võimalike vahemike paigutusviisidena uuriti võrdse laiusega ja võrdse suurusega vahemikke. Võimaliku kaugusfunktsioonina uuriti absoluut- ja ruutkaugust ning uudse lähenemisena ka Kullback-Leibleri kaugusest inspireeritud logaritmkaugust.

Töös jõuti tulemuseni, et ECE arvutamisel on kalibreerituse testi võimsuse suurendamiseks võimalik kasutada oluliselt paremaid parameetrite kombinatsioone, kui senises ECE kasutuses teaduskirjanduses tavaks on olnud. Töös läbi viidud katsete tulemused näitasid, et iga andmepunkt on mõistlik paigutada eraldi vahemikku, mis on üllatav tulemus, kuna varasemas ECE kasutuses on üldiselt kasutatud väikest arvu vahemikke. Samuti tõestati, et kui iga andmepunkt paigutada eraldi vahemikku, siis ei ole vahet, kas kasutada absoluut- või ruutkaugust. Töös näidati katseliselt, et nii cf-ECE kui ka cw-ECE jaoks on üksikute väga möödakalibreeritud andmepunktide korral logaritmkaugus oluliselt parem nii absoluut- kui ka ruutkaugusest. Samuti näitasid katsete tulemused, et juhtudel, kus klassifitseerija ennustused on kõik vähesel määral möödakalibreeritud, on tulemused erinevatel kaugusfunktsioonidel ligikaudu võrdsed või on logaritmkaugus väga vähesel määral halvem. Töös leitu võimaldab tulevikus paremini tuvastada klassifitseerijate mittekalibreeritust.

Töö edasi arendamiseks saaks uurida veel erinevaid ebasümmeetrilisi vahemike paigutusviise ning erinevaid võimalikke kaugusfunktsioone. Samuti saaks töös läbi viidud katsete tulemusi kinnitada rohkematel reaalsetel andmetel.

Viidatud kirjandus

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms That Make Sense of Data*. New York: Cambridge University Press.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). Introduction to the Dirichlet Distribution and Related Processes. *UWEE Technical Report, UWEETR-2010-0006*.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *International Conference on Machine Learning*. arXiv: 1706.04599.

Kull, M., Perello-Nieto, M., Kängsepp, M., Filho, T. S., Song, H., & Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration. *Advances in Neural Information Processing Systems.* arXiv: 1910.12656.

Kumar, A., Liang, P., & Ma T. (2019). Verified Uncertainty Calibration. *Advances in Neural Information Processing Systems.* arXiv: 1909.10155.

Lember, J. (2018). *Informatsiooniteooria.* Loengukonspekt ja ülesanded. Kasutatud 30.04.2020, https://courses.ms.ut.ee/MTMS.02.040/2018_spring/uploads/Main/konsp2018.pdf

Möls, M. (2013). *Hüpoteeside statistiline kontrollimine*. Biomeetria loengukonspekt. Kasutatud 30.04.2020, http://www.ms.ut.ee/mart/biomeetria2013/loeng3_4.pdf

Naeini, M. P, Cooper, G. F, & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,* 2901-2907.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., & Schön, T. B. (2019). Evaluating model calibration in classification. K. Chaudhuri, & M. Sugiyama (Eds). *Proceedings of Machine Learning Research*, vol 89, 3459–3467. arXiv: 1902.06977.

Widmann, D., Lindsten, F., & Zachariah, D. (2019). Calibration tests in multi-class classification: A unifying framework. *Advances in Neural Information Processing Systems*. arXiv: 1910.11385.

Lisad

I. Koodi repositoorium

Töös jooksutatud katsete ning loodud visualiseeringute kood on kättesaadav aadressil https://github.com/kaspar98/kalibreerituse-testi-optimeerimine.

II. Otsustuspuu ja otsustusmetsa treenimine

Kõik otsustuspuu ja otsustusmetsa treenimiseks kasutatud meetodid on pärit teegist scikit-learn ning kõik täpsustamata parameetrid on vaikeväärtusega.

Katses 7 kasutatud otsustuspuu on treenitud mudelil DecisionTreeClassifier parameetritega $min\_leaf\_samples = 50$ ja $random\_state = 999$.

Katses 8 kasutatud otsustusmets on treenitud mudelil RandomForestClassifier parameetritega $n\_estimators = 200$ ja $random\_state = 999$.

Katsetes kasutati treening- ja testandmete jaotamiseks scikit-learn'i meetodit train_test_split parameetritega $test\_size = 0.2$ ja $random\_state = 999$. Treening ja testandmetel kasutati tunnuste valimiseks mudelit ExtraTreesClassifier ja transformaatorit SelectFromModel.

III. Tõestused

Teoreem 1. Kui cw-ECE leidmises on iga andmepunkt eraldi vahemikus, siis tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe on ruutkaugusega ja absoluutkaugusega mõõdetud ECE puhul võrdne.

Teoreemi 1 tõestus. Olgu meil n andmepunkti, millest igaühe kohta on teada k-klassilise klassifitseerija ennustus vektorina $\boldsymbol{p_i} = (p_{i,1}, \dots, p_{i,k})$ ning tegelik märgend üks-mitmest vektorina (*one-hot vector*[*]) $\boldsymbol{y_i} = (y_{i,1}, \dots, y_{i,k})$, $0 \leq i \leq n$. Olgu meil iga andmepunkti kohta teada ka klassifitseerija ennustusest genereeritud uus märgend üks-mitmest vektorina $\boldsymbol{y'_i} = (y'_{i,1}, \dots, y'_{i,k})$. Olgu ECE leitud klassikaupa võrdse suurusega vahemikega, kus vahemike arvuks on n.

1) Leiame esmalt absoluutkaugusega mõõdetud ECE jaoks tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe.

Tegelike märgendite põhjal leitud ECE on

$$ECE_{abs1} = \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} |p_{i,j} - y_{i,j}|.$$

Genereeritud märgendite põhjal leitud ECE on

$$ECE_{abs2} = \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} |p_{i,j} - y'_{i,j}|.$$

Nende suuruste vahe on

$$v_{abs} = ECE_{abs1} - ECE_{abs2}$$

[*] Nullidest koosnev vektor, milles esineb üks 1

$$= \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} |p_{i,j} - y_{i,j}| - \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} |p_{i,j} - y'_{i,j}|$$

$$= \frac{1}{kn} \cdot \sum_{i=1}^{n} \sum_{j=1}^{k} \left(|p_{i,j} - y_{i,j}| - |p_{i,j} - y'_{i,j}| \right).$$

Paneme tähele, et

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \left(|p_{i,j} - y_{i,j}| - |p_{i,j} - y'_{i,j}| \right)$$

on summa üle kõigi andmepunktide. Vaatame seda summat mingi andmepunkti i jaoks. Selgema tähistuse jaoks loobume ajutiselt alaindeksist i.

$$\sum_{j=1}^{k} \left(|p_j - y_j| - |p_j - y'_j| \right)$$

Olgu selle andmepunkti tegelik märgend 1 a-nda klassi jaoks ning genereeritud märgend 1 b-nda klassi jaoks. Ehk leiduvad $a, b \in \{1, \dots, k\}$ nii, et $y_a = y'_b = 1$ ning iga $c, d \in \{1, \dots, k\}, c \neq a, d \neq b$ korral kehtib $y_c = y'_d = 0$. Seega

$$\begin{aligned}
&\sum_{j=1}^{k} \left(|p_j - y_j| - |p_j - y'_j| \right) = \\
&= |p_1 - y_1| - |p_1 - y'_1| + \cdots + |p_a - y_a| - |p_a - y'_a| + \cdots \\
&\qquad + |p_b - y_b| - |p_b - y'_b| + \cdots + |p_k - y_k| - |p_k - y'_k| \\
&= |p_1 - 0| - |p_1 - 0| + \cdots + |p_a - 1| - |p_a - 0| + \cdots \\
&\qquad + |p_b - 0| - |p_b - 1| + \cdots + |p_k - 0| - |p_k - 0| \\
&= p_1 - p_1 + \cdots + (1 - p_a) - p_a + \cdots + p_b - (1 - p_b) + \cdots + p_k - p_k \\
&= -2p_a + 2p_b.
\end{aligned}$$

Järelikult saame, et

$$v_{abs} = \frac{1}{kn} \cdot \sum_{i=1}^{n} \left(2p_{i,b} - 2p_{i,a} \right).$$

2) Leiame nüüd ruutkaugusega mõõdetud ECE jaoks tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe.

Tegelike märgendite põhjal leitud ECE on

$$ECE_{sq1} = \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \left(p_{i,j} - y_{i,j}\right)^2.$$

Genereeritud märgendite põhjal leitud ECE on

$$ECE_{sq2} = \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \left(p_{i,j} - y'_{i,j}\right)^2.$$

Nende suuruste vahe on

$$\begin{aligned} v_{sq} &= ECE_{sq1} - ECE_{sq2} \\ &= \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \left(p_{i,j} - y_{i,j}\right)^2 - \frac{1}{k} \cdot \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n} \left(p_{i,j} - y'_{i,j}\right)^2 \\ &= \frac{1}{kn} \cdot \sum_{i=1}^{n} \sum_{j=1}^{k} \left(\left(p_{i,j} - y_{i,j}\right)^2 - \left(p_{i,j} - y'_{i,j}\right)^2 \right). \end{aligned}$$

Paneme tähele, et

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \left(\left(p_{i,j} - y_{i,j}\right)^2 - \left(p_{i,j} - y'_{i,j}\right)^2 \right)$$

on summa üle kõigi andmepunktide. Vaatame seda summat mingi andmepunkti i jaoks. Selgema tähistuse jaoks loobume ajutiselt alaindeksist i.

$$\sum_{j=1}^{k} \left(\left(p_j - y_j\right)^2 - \left(p_j - y'_j\right)^2 \right)$$

Olgu selle andmepunkti tegelik märgend 1 a-nda klassi jaoks ning genereeritud märgend 1 b-nda klassi jaoks. Ehk leiduvad $a, b \in \{1, \dots, k\}$ nii, et $y_a = y'_b = 1$ ning iga $c, d \in \{1, \dots, k\}, c \neq a, d \neq b$ korral kehtib $y_c = y'_d = 0$. Seega

$$\sum_{j=1}^{k} \left(p_j - y_j\right)^2 - \left(p_j - y'_j\right)^2 =$$
$$= (p_1 - y_1)^2 - (p_1 - y'_1)^2 + \cdots + (p_a - y_a)^2 - (p_a - y'_a)^2 + \cdots$$
$$+ (p_b - y_b)^2 - (p_b - y'_b)^2 + \cdots + (p_k - y_k)^2 - (p_k - y'_k)^2$$
$$= (p_1 - 0)^2 - (p_1 - 0)^2 + \cdots + (p_a - 1)^2 - (p_a - 0)^2 + \cdots$$
$$+ (p_b - 0)^2 - (p_b - 1)^2 + \cdots + (p_k - 0)^2 - (p_k - 0)^2$$
$$= {p_1}^2 - {p_1}^2 + \cdots + {p_a}^2 - 2p_a + 1 - {p_a}^2 + \cdots$$
$$+ {p_b}^2 - {p_b}^2 + 2p_b - 1 + \cdots + {p_k}^2 - {p_k}^2$$
$$= -2p_a + 2p_b.$$

Järelikult saame, et

$$v_{sq} = \frac{1}{kn} \cdot \sum_{i=1}^{n} \left(2p_{i,b} - 2p_{i,a}\right).$$

Kuna $v_{abs} = v_{sq}$, siis olemegi tõestanud, mida tarvis.■

Teoreem 2. Kui cf-ECE leidmises on iga andmepunkt eraldi vahemikus, siis tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe on ruutkaugusega ja absoluutkaugusega mõõdetud ECE puhul võrdne.

Teoreemi 2 tõestus. Olgu meil n andmepunkti, millest igaühe kohta on teada k-klassilise klassifitseerija ennustus vektorina $\boldsymbol{p_i} = (p_{i,1}, \dots, p_{i,k})$ ning tegelik märgend üks-mitmest vektorina (*one-hot vector*) $\boldsymbol{y_i} = (y_{i,1}, \dots, y_{i,k})$, $0 \leq i \leq n$. Olgu meil iga andmepunkti kohta teada ka klassifitseerija ennustusest genereeritud uus märgend üks-mitmest vektorina $\boldsymbol{y'_i} = (y'_{i,1}, \dots, y'_{i,k})$. Olgu ECE leitud enesekindluse järgi võrdse suurusega vahemikega, kus vahemike arvuks on n.

1) Leiame esmalt absoluutkaugusega mõõdetud ECE jaoks tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe.

Tegelike märgendite põhjal leitud ECE on

$$ECE_{abs1} = \frac{1}{n}\sum_{i=1}^{n}\left|\max(\boldsymbol{p_i}) - y_{i,arg\max(\boldsymbol{p_i})}\right|.$$

Genereeritud märgendite põhjal leitud ECE on

$$ECE_{abs2} = \frac{1}{n}\sum_{i=1}^{n}\left|\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right|.$$

Nende suuruste vahe on

$$\begin{aligned} v_{abs} &= ECE_{abs1} - ECE_{abs2} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left|\max(\boldsymbol{p_i}) - y_{i,arg\max(\boldsymbol{p_i})}\right| - \frac{1}{n}\sum_{i=1}^{n}\left|\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\left|\max(\boldsymbol{p_i}) - y_{i,arg\max(\boldsymbol{p_i})}\right| - \left|\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right|\right). \end{aligned}$$

Paneme tähele, et

$$\sum_{i=1}^{n}\left(\left|\max\left(\boldsymbol{p_i}\right)-y_{i,arg\max\left(\boldsymbol{p_i}\right)}\right|-\left|\max\left(\boldsymbol{p_i}\right)-{y'}_{i,arg\max\left(\boldsymbol{p_i}\right)}\right|\right)$$

on summa üle kõigi andmepunktide. Vaatame seda summat mingi andmepunkti i jaoks. Selgema tähistuse jaoks loobume ajutiselt alaindeksist i.

$$\left|\max\left(\boldsymbol{p}\right)-y_{arg\max\left(\boldsymbol{p}\right)}\right|-\left|\max\left(\boldsymbol{p}\right)-{y'}_{arg\max\left(\boldsymbol{p}\right)}\right|$$

Vaatame nelja erinevat juhtu:

1) kui $y_{arg\max\left(\boldsymbol{p}\right)}=0$ ja ${y'}_{arg\max\left(\boldsymbol{p}\right)}=0$, siis

$$\left|\max\left(\boldsymbol{p}\right)-0\right|-\left|\max(\boldsymbol{p})-0\right|=0.$$

2) kui $y_{arg\max\left(\boldsymbol{p}\right)}=1$ ja ${y'}_{arg\max\left(\boldsymbol{p}\right)}=0$, siis

$$\left|\max\left(\boldsymbol{p}\right)-1\right|-\left|\max(\boldsymbol{p})-0\right|=1-\max(\boldsymbol{p})-\max(\boldsymbol{p})=1-2\max\left(\boldsymbol{p}\right).$$

3) kui $y_{arg\max\left(\boldsymbol{p}\right)}=0$ ja ${y'}_{arg\max\left(\boldsymbol{p}\right)}=1$, siis

$$\left|\max\left(\boldsymbol{p}\right)-0\right|-\left|\max(\boldsymbol{p})-1\right|=\max(\boldsymbol{p})-(1-\max(\boldsymbol{p}))=2\max(\boldsymbol{p})-1.$$

4) kui $y_{arg\max\left(\boldsymbol{p}\right)}=1$ ja ${y'}_{arg\max\left(\boldsymbol{p}\right)}=1$, siis

$$\left|\max\left(\boldsymbol{p}\right)-1\right|-\left|\max(\boldsymbol{p})-1\right|=0.$$

2) Leiame nüüd ruukaugusega mõõdetud ECE jaoks tegelike märgendite põhjal leitud ECE ja klassifitseerija ennustustest genereeritud märgendite põhjal leitud ECE vahe.

Tegelike märgendite põhjal leitud ECE on

$$ECE_{sq1}=\frac{1}{n}\sum_{i=1}^{n}\left(\max(\boldsymbol{p_i})-y_{i,\mathrm{argmax}(\boldsymbol{p_i})}\right)^2.$$

Genereeritud märgendite põhjal leitud ECE on

$$ECE_{sq2} = \frac{1}{n}\sum_{i=1}^{n}\left(\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right)^2.$$

Nende suuruste vahe on

$$\begin{aligned} v_{sq} &= ECE_{sq1} - ECE_{sq2} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\max(\boldsymbol{p_i}) - y_{i,\mathrm{argmax}(\boldsymbol{p_i})}\right)^2 - \frac{1}{n}\sum_{i=1}^{n}\left(\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\left(\max(\boldsymbol{p_i}) - y_{i,\mathrm{argmax}(\boldsymbol{p_i})}\right)^2 - \left(\max(\boldsymbol{p_i}) - y'_{i,arg\max(\boldsymbol{p_i})}\right)^2\right). \end{aligned}$$

Paneme tähele, et

$$\sum_{i=1}^{n}\left(\left(\max(\boldsymbol{p_i}) - y_{i,\mathrm{argmax}(\boldsymbol{p_i})}\right)^2 - \left(\max(\boldsymbol{p_i}) - y'_{i,\mathrm{argmax}(\boldsymbol{p_i})}\right)^2\right)$$

on summa üle kõigi andmepunktide. Vaatame seda summat mingi andmepunkti i jaoks. Selgema tähistuse jaoks loobume ajutiselt alaindeksist i.

$$\left(\max(\boldsymbol{p}) - y_{\mathrm{argmax}(\boldsymbol{p})}\right)^2 - \left(\max(\boldsymbol{p}) - y'_{\mathrm{argmax}(\boldsymbol{p})}\right)^2$$

Vaatame nelja erinevat juhtu:

1) kui $y_{arg\max(\boldsymbol{p})} = 0$ ja $y'_{arg\max(\boldsymbol{p})} = 0$, siis

$$(\max(\boldsymbol{p}) - 0)^2 - (\max(\boldsymbol{p}) - 0)^2 = 0.$$

2) kui $y_{arg\max(\boldsymbol{p})} = 1$ ja $y'_{arg\max(\boldsymbol{p})} = 0$, siis

$$\begin{aligned} (\max(\boldsymbol{p}) - 1)^2 - (\max(\boldsymbol{p}) - 0)^2 &= \max(\boldsymbol{p})^2 - 2\max(\boldsymbol{p}) + 1 - \max(\boldsymbol{p})^2 \\ &= 1 - 2\max(\boldsymbol{p}). \end{aligned}$$

3) kui $y_{arg\max(\boldsymbol{p})} = 0$ ja $y'_{arg\max(\boldsymbol{p})} = 1$, siis

$$(\max(\boldsymbol{p}) - 0)^2 - (\max(\boldsymbol{p}) - 1)^2 = \max(\boldsymbol{p})^2 - \max(\boldsymbol{p})^2 + 2\max(\boldsymbol{p}) - 1$$
$$= 2\max(\boldsymbol{p}) - 1.$$

4) kui $y_{arg\max(\boldsymbol{p})} = 1$ ja $y'_{arg\max(\boldsymbol{p})} = 1$, siis

$$(\max(\boldsymbol{p}) - 1)^2 - (\max(\boldsymbol{p}) - 1)^2 = 0.$$

Kuna igal võimalikul juhul ühe andmepunkti jaoks on summa liige absoluut- ja ruutkauguse jaoks võrdne, siis on selge, et

$$v_{abs} = v_{sq}.$$

Seega olemegi tõestanud, mida tarvis.■

IV. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Kaspar Valk**,

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

   Klassifitseerija kalibreerituse testi võimsuse suurendamine,

   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.

Kaspar Valk

08.05.2020