TARTU ÜLIKOOL
Loodus- ja täppisteaduste valdkond
Arvutiteaduse instituut
Informaatika õppekava

Mart-Mihkel Aun

Masinõppe mudelite hindamine väheste märgenditega andmetel

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur, DSc. (Tech)

Tartu 2023

Masinõppe mudelite hindamine väheste märgenditega andmetel Lühikokkuvõte:

Lühikokkuvõte: Klassifitseerimisülesandeid lahendavate masinõppe mudelite hindamiseks kasutatakse kvaliteedimõõte nagu õigsus täpsus ja saagis. Nimetatud suurused või nende hinnangud avalduvad andmepunktide tegelike klassimärgendite ja meetodi klassifikatsioonide kaudu. Tegelike klassimärgendite leidmiseks peab need manuaalselt üle vaatama. Sageli hinnatakse kvaliteedimõõte üle lõpliku valimi, leitud hinnangud sisaldavad vigu. Antud töö käigus leiti kui suurt valimit on vaja, et mingi kindlusega ei ületaks hinnangu viga selle lubatud piiri. Lisaks peab valimi puhul õigsuse, täpsuse või saagise hindamiseks definitsiooni põhjal leidma kõikide valimi andmepunktide märgendid. Kui lisaks hinnatavale klassifitseerimismeetodile on olemas teine meetod, saab seda kasutada uue hindamiseks. Seejuures on võimalik märgendamiseks vajalikku manuaalset tööd vähendada, uurides uue meetodi kvaliteedimõõdu arvutamise asemel, kui palju on uus meetod vanast parem. Töös uuriti tehnikaid, mis aitavad vähendada märgendamist vajavate andmepunktida arvu kahe klassifitseerimismeetodi kvaliteedimõõtude vahede hindamiseks.

Võtmesõnad: masinõpe, klassifitseerimine, tõenäosusteooria, statistika, õigsus, täpsus, saagis

CERCS:P176 Tehisintellekt

Evaluating machine learning models on data with few labels Abstract:

Machine learning models used to solve classification tasks are evaluated using quality measures such as accuracy, precision, and recall. These measures or their estimates are calculated through the class labels of data points and the classifications made by the method on those data points. To find the actual class labels, they must be manually reviewed. Often, quality measures are evaluated using a finite sample, and the obtained estimates may contain errors. In this thesis, the necessary sample size was determined not to exceed the limit of estimation error with a certain confidence level. Additionally, for a given sample, the definition-based method of determining accuracy, precision, or recall for all the sample data points' labels must be established. If another method exists alongside the one being evaluated, it can be used for a new assessment. In this case, it is possible to reduce the amount of manual work required for labeling by assessing how much better the new method is compared to the old one, rather than calculating the quality measures of the new method. This thesis explored techniques that help reduce the number of data points that require labeling for evaluating the quality measures of the two classification methods.

Keywords: machine learning, classification, probability theory, statistics, accruacy, precision, recall

CERCS:P176 Artificial intelligence

Sisukord

1 Sissejuhatus

Masinõppe mudelite hindamiseks kasutatakse erinevaid viise. Klassifitseerimisülesannete puhul saab ülesannet lahendava masinõppe meetodi headust hinnata kvaliteedimõõtudega. Kolm tihti kasutatud kvaliteedimõõtu on õigsus, täpsus ja saagis. Õigsus, täpsus ja saagis on nulli ja ühe vahelised numbrilised suurused, kus suurem väärtus tähendab paremat mudelit. Nimetatud suurused või nende hinnangud avalduvad mingi hulga andmepunktide tegelike klassimärgendite ja meetodi poolt klassifitseeritud klassimärgendite kaudu. Tegelike klassimärgendite leidmiseks peab need manuaalselt üle vaatama, mis võib paljude andmepunktide puhul muutuda kulukaks. Kvaliteedimõõtude täpsete väärtuste arvutamiseks peaks teadma populatsiooni kõigi andmepunktide märgendeid. Kuna kogu populatsiooni uurimine ei pruugi praktikas olla võimalik, leitakse sageli kvaliteedimõõtudele hinnangud üle lõpliku valimi.

Üle valimi arvutatud suuruste hinnangud ei ühti tavaliselt nende suuruste oodatud väärtusega, ehk üle populatsiooni arvutatud suurustega. Leitud hinnangud sisaldavad valimi juhuslikkuse tõttu vigu. Seega tuleb küsimuse alla kui suur antud viga on ning kui kindel saab olla, et viga ei ole suurem kui võib lubada. Antud töö käigus leitakse kui suurt valimit on vaja (või kui väikest valimit võib võtta), et mingi kindlusega ei ületaks hinnangu viga selle lubatud piiri.

Valimi puhul peab õigsuse, täpsuse või saagise hindamiseks definitsiooni põhjal leidma kõikide valimi andmepunktide märgendid. Ka minimaalse võimaliku valimi puhul võib see olla probleemne. Juhul kui asendada vana mudelit uuega ei pea küsima mis on uue meetodi õigsus, võib ka uurida kas ja kui palju on uue meetodi õigsus suurem vana omast. Selles töös on uuritud tehnikaid, mis aitavad vähendada märgendamist vajavate andmepunktida arvu kahe klassifitseerimismeetodi kvaliteedimõõtude vahede hindamiseks.

Esimese osas on toodud töös kasutatud mõisted ja definitsioonid ning kirjeldatud töö mõistmist abistavad taustteadmised. Töö teises osas on defineeritud klassifitseerimismeetodi kvaliteedimõõdud nende oodatud väärtuste kaudu, kirjeldatud viise nende hindamiseks ning uuritud kui suurt valimit on vaja küllaltki suure kindlusega piisavalt täpse hinnangu leidmiseks. Kolmandas osas uuritakse kuidas hinnata uut klassifitseerimismeetodit võrreldes seda juba olemasolevaga ning selle juures vältida kõikide andmepunktide märgendamist.

2 Taust

Masinõppemeetodite tulemuslikkuse mõõtmiseks kasutatakse mitmeid teoreetilisi kvaliteedimõõtusid. Kvaliteedimõõdud on defineeritud läbi nende ooteväärtuste. Praktikas hinnatakse kvaliteedimõõtusid ligikaudselt, arvutades keskmisi väärtusi üle valimi.

2.1 Lõplikud ja lõpmatud andmekogumid

Statistilises uuringus nimetatakse uuringu all olevat objekti üldkogumiks ehk populatsiooniks. Populatsioon koosneb andmepunktidest. Andmepunktide kogus on määratletud populatsiooni korral tavaliselt lõplik. Kõikse uuringu korral mõõdetakse kõiki populatsiooni andmepunkte. Kõikne uuring on sageli ülemäära kulukas. Lihtsam on mõõta juhuslikku osahulka populatsiooni andmepunktidest. Mõõdetavate andmepunktide hulka nimetatakse valimiks. Valimi põhjal tehakse järeldusi kogu populatsiooni kohta. Tehtud järeldused võivad valimi juhuslikkuse tõttu sisaldada vigu, kuid sellised vead on tõenäosuslikult hinnatavad [4].

Vahel pole populatsioon uuringu tegemise hetkel üheselt fikseeritud või lõplik, nagu näiteks kõik järgmise 24 tunni jooksul kiirabisse pöörduvad inimesed või kõigi Canon 70D fotokaameratega tehtud pildid. Sellisel juhul määrab tulemuse andmeid genereeriv füüsiline protsess ning seda mudeldatakse tihti juhusliku jaotusega. Formaalselt on juhuslik jaotus andmete allikas, millest saab võtta kuitahes palju sõltumatuid andmepunkte. Antud töös vaatame protsesse, millel on lõplik arv väljundväärtusi, see tähendab jaotus on diskreetne. Sel juhul fikseerib jaotus iga konkreetse andmepunkti jaoks selle esinemistõenäosuse, millest võib mõelda kui andmepunkti oodatavat sagedusest valimis, millesse on võetud piisavalt palju andmepunkte jaotusest.

Lõpliku valimi korral saab mingi uuritava omaduse A esinemise tõenäosust defineerida kui suhet omaduse esinemise arvu valimis n_A ning valimi suuruse n vahel

$$\Pr[A] = \frac{n_A}{n} \ .$$

Juhuslikust valimist juhusliku andmepunkti võtmiselt tähendab see tõenäosust, et andmepunkt on omadusega A.

Juhuslikku valimit saab moodustada võttes populatsioonist andmepunkte üksteisest sõltumatult ja juhuslikult. Võib juhtuda, et samad populatsiooni andmepunktid sattuvad valimisse mitmekordselt. Saadud valimi puhul eeldatakse, et kõik selle andmepunktid on sama jaotusega. Tegelikkuses ei pruugi sõltumatuse ja sama jaotuse eeldus olla täidetud. Näiteks Canon 70D fotokaameraga tehtud pildiseeria piltide jaotused on üksteisega korreleeritud, kuna ühest sündmusest tehakse tavaliselt mitu pilti. See-eest üksikute piltide jaotuste puhul võib sageli sõltumatust eeldada. Valimikeskmine on valimi kõiki-

de andmepunktide väärtuste aritmeetiline keskmine

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \ . \tag{1}$$

Juhul kui x_i on indikaator valimi i-nda andepunkti mingi omaduse kohta, läheneb valimikeskmine $\bar{x}$ tõenäosusele, et juhuslikul objektil on see omadus.

Valimidispersioon on kõigi andmepunktide hälvete ruutude aritmeetiline keskmine

$$s^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \ . \tag{2}$$

Dispersioon näitab andmete hajuvust. Valimidispersiooni kaudu saab leida hinnangu valimikeskmise dispersioonile $\frac{s^2}{n}$, millest ruutjuurt nimetatakse standardveaks [4]. Standardviga näitab valimikeskmise $\bar{x}$ fluktuatsiooni tema oodatud väärtusest, ehk mitut tüvenumbrit hinnangust võib usaldada.

Olgu olemas juhuslik protsess, mis genereerib populatsiooni andmepunkte. Protsessi genereeritud andmepunktid on üksteisest sõltumatud ja sama jaotusega. Jaotuse abil defineeritud potentsiaalselt lõpmatu andmestiku uurimiseks saab defineerida valimiüleste valemite (1) ja (2) analoogid, milleks need lõpmata suure valimi korral koonduvad.

Valimikeskmise analoog keskväärtus iseloomustab jaotuse väärtuste paiknevust, mõnikord nimetatakse keskväärtust ka matemaatiliseks ootuseks [3]. Diskreetse juhusliku suuruse X keskväärtus on defineeritud summana

$$\mathbf{E}\left[X\right] = \sum_{i} x_i p_i \ , \tag{3}$$

kus x_i on suuruse üks võimalik väärtus ning p_i tõenäosus, et X selle väärtuse võtab. Pideva juhusliku suuruse X, mille tihedusfunktsioon on $f_X(x)$, keskväärtus leitav määratud integraalina

$$\mathbf{E}\left[X\right] = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \ ,$$

mitmemõõtmelise juhusliku suuruse korral on selle keskväärtus leitav analoogiliselt mitmemõõtmelise integraaliga.

Valimidispersiooni analoog dispersioon on juhusliku suuruse hälve ruudu keskväärtus

$$\mathbf{D}\left[X\right] = \mathbf{E}\left[(X - \mathbf{E}\left[X\right])^2\right] = \mathbf{E}\left[X^2\right] - \mathbf{E}\left[X\right]^2 \ . \tag{4}$$

Ruutjuurt dispersioonist nimetatakse standardhälbeks. Lõpliku populatsiooni korral avalduvad jaotuseülesed valemid (3) ja (4) neile vastavate valimiüleste valemitena (1) ja (2). Jaotuseülesed valemeid on vaja, et uurida juhuslikkust sisaldavate suuruste oodatud käitumist.

2.2 Veahinnangud

Numbriliselt ülesannete lahendamisel võib täpse lahendi leidmine olla aeganõudev ja kulukas. Ülesande ligikaudne lahendamine võib osutuda otstarbekamaks, näiteks populatsiooni keskväärtuse leidmise asemel leida valimikeskmine. Antud näite korral on valimikeskmine ligikaudne hinnang keskväärtusele. Ligikaudsete väärtuste headust on võimlaik mõõta kasutades veahinnanguid. Veahinnangud kirjeldavad mingi täpse arvu ja selle lähendi erinevust. Vastuse lähendamisel praktilises olukorras ei pruugi olla täpset lahendust teada. Seega sobivad veahinnangud lähendamismeetodite teoreetiliseks hindamiseks.

Olgu a ligikaudne väärtus arvust a_0. Ligikaudse arvu a absoluutseks veaks nimetatakse arvu

$$\Delta a = a - a_0 \ .$$

Absoluutse vea õigesti mõistmiseks tuleb arvestada lähendatava arvu skaalaga. Arvude puhul, mis ulatuvad mitmetesse tuhandetesse, tähendab absoluutne viga $0{,}5$ väga täpset lähendit. Nulli ja ühe vaheliste arvude puhul mitte. Seevastu saab kasutusele võtta hinnangu, mis arvestab skaalaga. Ligikaudse arvu a relatiivseks ehk suhteliseks veaks nimetatakse suurust

$$\delta a = \frac{\Delta a}{a_0} = \frac{a - a_0}{a_0} = \frac{a}{a_0} - 1 \ .$$

Mõnikord esitatakse relatiivne viga protsentides.

Ligikaudsete väärtuste puhul on eesmärgiks nende veahinnangute absoluutväärtuselt minimeerimine. Alati ei pruugi see olla võimalik, tuleb leppida mingi veaga. Lisaks võib raske olla täielikult veenduda, et leitud lähendi viga on väiksem kui maksimaalne lubatud viga. See-eest võib võimalik olla uurida kui suure tõenäosusega on saavutatud viga, mis on väiksem lubatud maksimaalsest veast

$$\Pr\left[\mid a - a_0 \mid < \varepsilon\right] = 1 - \alpha \ ,$$

kus ε on absoluutse vea absoluutväärtuse ülemine piir, suurust α nimetatakse olulisusnivooks ning vahet $1 - \alpha$ usaldusnivooks või kindluseks. Tihti valitakse α väärtuseks $0{,}05$, mõnikord ka $0{,}01$ või isegi $0{,}32$.

Kui juhuslik suurus on vaadeldav paljude sõltumatute juhuslike suuruste summana on alust arvata, et summa on ligikaudu normaaljaotusega [3]. Seega kui ligikaudsed väärtused sisaldavad paljude sõltumatute juhuslike veakomponentide summat võib ka veahinnangute puhul eeldada, et need on normaaljaotusega.

Juhuslik suurus X on normaaljaotusega kui tema tihedusfunktsioon on

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \ ,$$

kus jaotuse parameetrid μ ja σ on vastavalt X keskväärtus ja standardhälve. Seda fakti tähistatakse lühidalt $X \sim \mathcal{N}(\mu, \sigma)$. Normaaljaotuse tihedus on sümmeetriline keskväärtuse μ ümber. Veahinnangute puhul on nende oodatud väärtus null, siis sümmeetrilisuse tõttu on negatiivsed ja positiivsed vead sama tõenäolised. Oluline omadus normaaljaotuse puhul on see, kui suure tõenäosuse katavad standardhälbe täisarvkordsed vahemikud keskväärtuse ümbruses. Vahemik $\mu \pm \sigma$ katab tõenäosuse 68%, $\mu \pm 2\sigma$ tõenäosuse 95% ning $\mu \pm 3\sigma$ tõenäosuse 99%.

2.2.1 Summa ja vahe relatiivse vea omadus

Vahel võib uuritav suurus olla arvutatav mitme ligikaudse suuruse kaudu. Ligikaudsete suurustega tehtud arvutused annavad ligikaudseid tulemusi. Seega tasub tulemuse täpsuse hindamiseks uurida kuidas viga arvutustel edasi kandub. Sealjuures tehakse tihti lihtsustav eeldus, et lähendid on normaaljaotusega.

Olgu x ja y ligikaudsed väärtused arvudest x_0 ja y_0 absoluutsete vigadega vastavalt ε_x ning ε_y:

$$\begin{aligned} x &= x_0 + \varepsilon_x, \qquad \varepsilon_x \sim \mathcal{N}(0, \sigma_x) \ , \\ y &= y_0 + \varepsilon_y, \qquad \varepsilon_y \sim \mathcal{N}(0, \sigma_y) \ , \end{aligned}$$

kus vead ε_x ja ε_y on sõltumatud. Ligikaudsete väärtuste x ja y summa relatiivne viga avaldub kujul

$$\delta_+ = \frac{(x+y) - (x_0 + y_0)}{x_0 + y_0} = \frac{x_0 + \varepsilon_x + y_0 + \varepsilon_y - x_0 - y_0}{x_0 + y_0} = \frac{\varepsilon_x + \varepsilon_y}{x_0 + y_0} \ .$$

Kuna ε_x ja ε_y keskväärtused on nullid on ka summa relatiivse vea δ_+ keskväärtus null. Vigade sõltumatuse tõttu on korrutise $\varepsilon_x \cdot \varepsilon_y$ keskväärtus samuti null. Seega on summa relatiivse vea hajuvus leitav kui selle ruudu keskväärtus

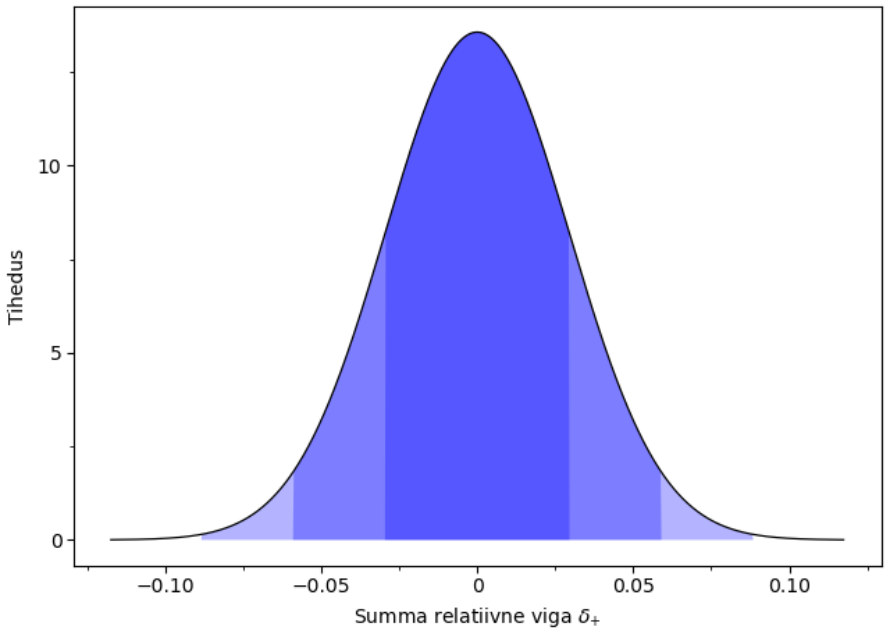
$$\mathbf{D}\left[\delta_+\right] = \mathbf{E}\left[\delta_+^2\right] = \mathbf{E}\left[\frac{\varepsilon_x^2 + 2\varepsilon_x\varepsilon_y + \varepsilon_y^2}{(x_0 + y_0)^2}\right] \ = \mathbf{E}\left[\frac{\varepsilon_x^2 + \varepsilon_y^2}{(x_0 + y_0)^2}\right] = \frac{\mathbf{D}\left[\varepsilon_x\right] + \mathbf{D}\left[\varepsilon_y\right]}{(x_0 + y_0)^2} \ .$$

Ligikaudsete arvude x ja y vahe relatiivne viga avaldub sarnaselt summale

$$\delta_- = \frac{\varepsilon_x - \varepsilon_y}{x_0 - y_0} \ .$$

Vahe relatiivse vea dispersioon on kujult samuti sarnane summale, kuid juhul kui täpsed väärtused on lähedased on selle hajuvus väga suur

$$\mathbf{D}\left[\delta_-\right] = \frac{\mathbf{D}\left[\varepsilon_x\right] + \mathbf{D}\left[\varepsilon_y\right]}{(x_0 - y_0)^2} \ .$$



Joonis 1. Summa relatiivse vea tihedus juhul $x \sim \mathcal{N}(0{,}8; 0{,}04)$, $y \sim \mathcal{N}(0{,}9; 0{,}03)$.

Kuna veakomponentide ε_x ja ε_y puhul on eeldatud normaaljaotust, on võimalik normaaljaotuse omadusi kasutades arvutada, mis tõenäosusega viga δ_+ või δ_- mingisse vahemikku jääb. Näiteks täpsete väärtuste $x_0 = 0{,}8$, $y_0 = 0{,}9$ ning absoluutsete vigage $\varepsilon_x \sim \mathcal{N}(0; 0{,}04)$, $\varepsilon_y \sim \mathcal{N}(0; 0{,}03)$ korral jääb $x + y$ relatiivne viga tõenäosusega ligikaudu 68% vahemikku $\pm 0{,}029$. Joonisel 1 on kujutatud antud näite korral summa $x + y$ relatiivse vea tihedus. Heledusega esiletõstetud alad märgivad standardhälbe täisarvkordseid vahemikke keskväärtuse ümber.

2.2.2 Korrutise relatiivse vea omadus

Mõnikord võib lähendatav suurus olla leitav ligikaudsete arvude korrutisena. Näiteks riskülikukujulise põranda pindala leidmiseks peab korrutama põranda laiuse ja pikkuse, mis mõõtemääramatuse või vigase mõõteriista tõttu ei pruugi olla täpsed.

Olgu x ja y ligikaudsed väärtused arvudest x_0 ja y_0 absoluutsete vigadega vastavalt ε_x ning ε_y:

$$\begin{aligned} x &= x_0 + \varepsilon_x, \qquad \varepsilon_x \sim \mathcal{N}(0, \sigma_x) \ , \\ y &= y_0 + \varepsilon_y, \qquad \varepsilon_y \sim \mathcal{N}(0, \sigma_y) \ , \end{aligned}$$

kus vead ε_x ja ε_y on sõltumatud. Suuruste x ja y korrutise relatiivne viga avaldub kujul

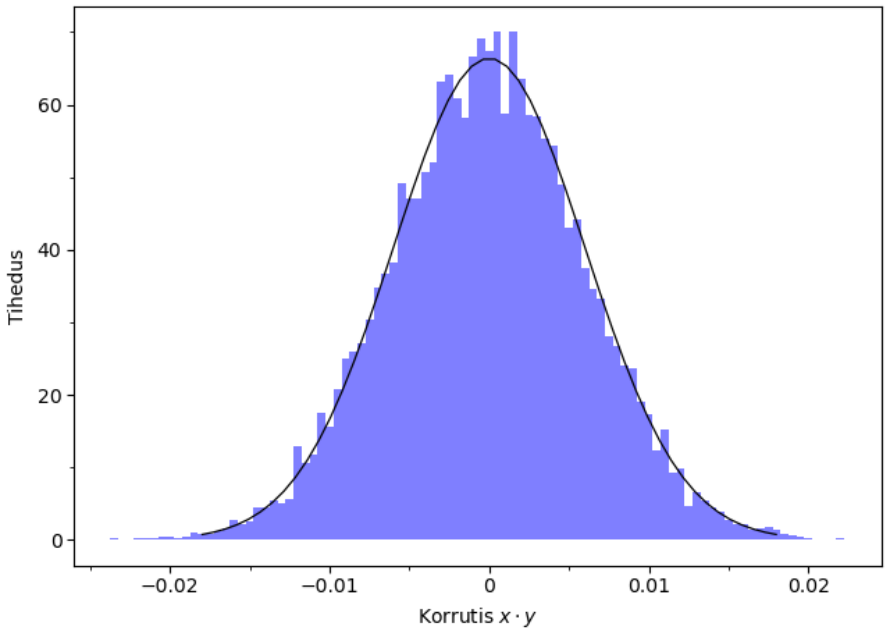
$$\begin{aligned}\delta &= \frac{x \cdot y - x_0 \cdot y_0}{x_0 \cdot y_0} = \frac{(x_0 + \varepsilon_x) \cdot (y_0 + \varepsilon_y) - x_0 \cdot y_0}{x_0 \cdot y_0} \\ &= \frac{x_0 \cdot \varepsilon_y + y_0 \cdot \varepsilon_x + \varepsilon_x \cdot \varepsilon_y}{x_0 \cdot y_0} = \frac{\varepsilon_y}{y_0} + \frac{\varepsilon_x}{x_0} + \frac{\varepsilon_x}{x_0} \cdot \frac{\varepsilon_y}{y_0} \quad .\end{aligned}$$

Vigade ε_x ja ε_y sõltumatuse tõttu on δ keskväärtus null. Arvestades eelnevat saab tuletada korrutise relatiivse vea dispersiooni

$$\begin{aligned}\mathbf{D}[\delta] &= \mathbf{E}\left[\delta^2\right] - \mathbf{E}[\delta]^2 = \mathbf{E}\left[\delta^2\right] - 0 \\ &= \mathbf{E}\left[\left(\frac{\varepsilon_y}{y_0}\right)^2 + \left(\frac{\varepsilon_x}{x_0}\right)^2 + \left(\frac{\varepsilon_x}{x_0} \cdot \frac{\varepsilon_y}{y_0}\right)^2 + 2 \cdot \frac{\varepsilon_x \cdot \varepsilon_y}{x_0 \cdot y_0} + 2 \cdot \frac{\varepsilon_x^2 \cdot \varepsilon_y}{x_0^2 \cdot y_0} + 2 \cdot \frac{\varepsilon_x \cdot \varepsilon_y^2}{x_0 \cdot y_0^2}\right] \\ &= \mathbf{E}\left[\left(\frac{\varepsilon_y}{y_0}\right)^2 + \left(\frac{\varepsilon_x}{x_0}\right)^2 + \left(\frac{\varepsilon_x}{x_0} \cdot \frac{\varepsilon_y}{y_0}\right)^2\right] \\ &= \mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right] + \mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right] + \mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right] \cdot \mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right] \quad .\end{aligned}$$

Kui x ja y relatiivsete vigade dispersioonid on väikesed on nende korrutis veelgi väiksem. Sellisel juhul saab δ dispersiooni hinnata küllaltki täpselt jättes korrutise arvutusest välja

$$\mathbf{D}[\delta] \approx \mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right] + \mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right] \quad . \tag{5}$$



Joonis 2. Korrutise $x \cdot y$ relatiivse vea simuleeritud tihedus histogrammina ja teoreetiline lähendus normaaöjaotsega juhul $x \sim \mathcal{N}(0{,}8; 0{,}004)$, $y \sim \mathcal{N}(0{,}9; 0{,}003)$.

Joonisel 2 on histogrammina esitatud simuleeritud korrutise $x \cdot y$ relatiivne viga kümnetuhande juhusliku x ja y väärtuse puhul, kus $x \sim \mathcal{N}(0{,}8; 0{,}004)$, $y \sim \mathcal{N}(0{,}9; 0{,}003)$. Tumeda joonega on näidatud korrutise relatiivse vea lähendus normaaljaotusega tulemuse (5) põhjal.

2.2.3 Jagatise relatiivse vea omadus

Osutub, et jagatise puhul on vea edasikandumine natukene keerulisem. Mõningatel eeldustel on leitav piisavalt täpne hinnang jagatise relatiivse vea kohta.

Olgu x ja y ligikaudsed väärtused arvudest x_0 ja y_0 absoluutsete vigadega vastavalt ε_x ning ε_y:

$$\begin{aligned} x &= x_0 + \varepsilon_x, \qquad \varepsilon_x \sim \mathcal{N}(0, \sigma_x) \ , \\ y &= y_0 + \varepsilon_y, \qquad \varepsilon_y \sim \mathcal{N}(0, \sigma_y) \ , \end{aligned}$$

kus vead ε_x ja ε_y on sõltumatud. Suuruste x ja y jagatise relatiivne viga avaldub kujul

$$\delta = \left(\frac{x}{y} - \frac{x_0}{y_0} \right) : \frac{x_0}{y_0} = \frac{x \cdot y_0}{y \cdot x_0} - 1 = \frac{y_0}{x_0} \cdot \frac{x_0 + \varepsilon_x}{y_0 + \varepsilon_y} - 1 \ .$$

Jagatise relatiivse vea dispersiooni leidmist saab taandada korrutise relatiivse vea dispersioonile, sest $x/y = x \cdot y^{-1}$. Seega on vaja hinnata y pöördväärtuse relatiivse vea disperisooni

$$\mathbf{D}\left[\left(\frac{1}{y}-\frac{1}{y_0}\right):\frac{1}{y_0}\right]=\mathbf{D}\left[\frac{y_0}{y}\right]=y_0^2\cdot\mathbf{D}\left[\frac{1}{y}\right] \quad . \tag{6}$$

Taylori arenduse põhjal on y^{-1} esitatav ligikaudselt

$$\frac{1}{y}=\frac{1}{y_0+\varepsilon_y}\approx\frac{1}{y_0}-\frac{1}{y_0^2}\cdot\varepsilon_y \quad .$$

Eeldusel, et y relatiivne viga on väike, on ka Taylori arenduse jääkliige väike. Vahetulemuse põhjal

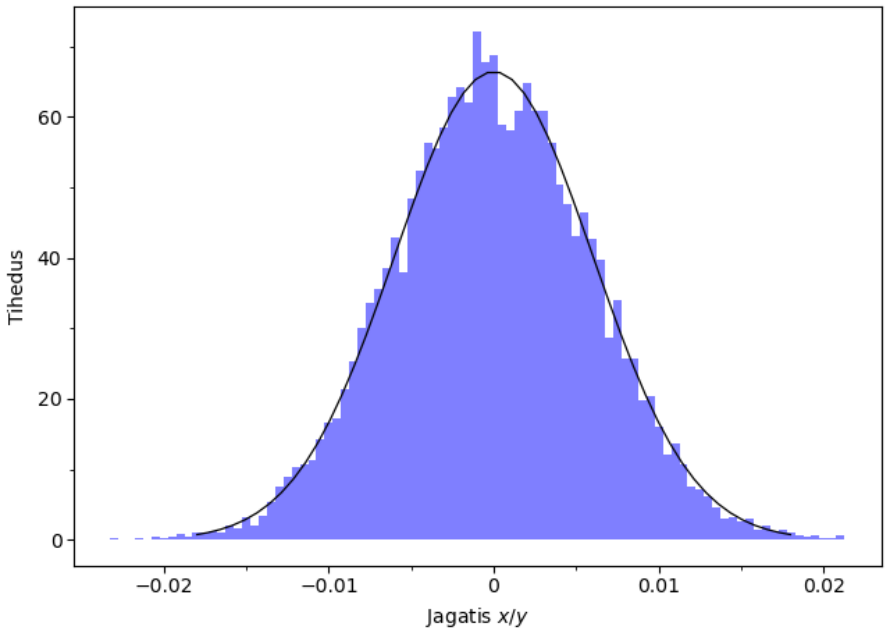
$$\begin{aligned}\mathbf{D}\left[\frac{1}{y}\right]&\approx\mathbf{D}\left[\frac{1}{y_0}-\frac{1}{y_0^2}\cdot\varepsilon_y\right]\\&=\mathbf{E}\left[\left(\frac{1}{y_0}-\frac{1}{y_0^2}\cdot\varepsilon_y\right)^2\right]-\mathbf{E}\left[\frac{1}{y_0}-\frac{1}{y_0^2}\cdot\varepsilon_y\right]^2\\&=\frac{1}{y_0^2}+\frac{\sigma_y^2}{y_0^4}-\frac{1}{y_0^2}=\frac{\sigma_y^2}{y_0^4} \quad .\end{aligned} \tag{7}$$

Tulemuste (6) ja (7) põhjal saab avaldada y^{-1} relatiivse vea ligikaudse dispersiooni

$$\mathbf{D}\left[\left(\frac{1}{y}-\frac{1}{y_0}\right):\frac{1}{y_0}\right]\approx y_0^2\cdot\frac{\sigma_y^2}{y_0^4}=\mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right] \quad .$$

Seega on x ja y jagatise relatiivse vea dispersioon, väikese y relatiivse vea korral, ligikaudu võrdne korrutise omaga

$$\mathbf{D}\left[\delta\right]\approx\mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right]+\mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right]+\mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right]\cdot\mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right]\approx\mathbf{D}\left[\frac{\varepsilon_x}{x_0}\right]+\mathbf{D}\left[\frac{\varepsilon_y}{y_0}\right] \quad . \tag{8}$$



Joonis 3. Jagatise $x : y$ relatiivse vea simuleeritud tihedus histogrammina ja teoreetiline lähendus normaaljaotsega juhul $x \sim \mathcal{N}(0{,}8; 0{,}004)$, $y \sim \mathcal{N}(0{,}9; 0{,}003)$.

Joonisel 3 on histogrammina esitatud simuleeritud jagatise $x : y$ relatiivne viga kümnetuhande juhuslikult genereeritud x ja y väärtuse puhul, kus $x \sim \mathcal{N}(0{,}8; 0{,}004)$, $y \sim \mathcal{N}(0{,}9; 0{,}003)$. Tumeda joonega on näidatud relatiivse vea lähendus normaaljaotusega tulemuse (8) põhjal.

3 Masinõppe meetodite kvaliteedimõõtude veahinnangud

Klassifitseerimismeetodite puhul kolm laialdaselt kasutatud kvaliteedimõõtu on õigsus, täpsus ja saagis. Õigsus näitab, kui sagedasti meetod andmepunkte õigesti klassifitseerib (tähistatakse Acc). Täpsus mõõdab päris positiivsete klassifikatsioonide sagedust ennustatud positiivsete seast (tähistatakse $Prec$). Saagis on päris positiivsete klassifikatsioonide sagedus positiivsete andmepunktide seast (tähistatakse Rec).

Tähistagu i-ndat andmepunkti elementide paar (x_i, y_i), kus x_i tähistab andmepunkti tunnuseid ning y_i andmepunkti klassi. Antud töös vaatleme ainult binaarse klassifitseerimisülesannet, kus märgend y_i võib olla üks kahest võimalikust väärtusest, negatiivne klass tähistusega 0 või positiivne klass tähistusega 1. Tihti vaadatakse meetodi käitumist andmepunktidel. Meetodi A puhul andmepunktile vastav klassifikatsioon on $A(x_i) = a_i$. Juhusliku andmepunkti võib vaadelda juhusliku suurusena (X, Y) ning meetodi klassifikatsiooni selle tunnustel kui juhuslikku suurust A. Seega on võimalik meetodi A kvaliteedimõõdud defineerida läbi ooteväärtuste

$$Acc = \mathbf{E}\left[A = Y\right] \ , \tag{9}$$

$$Prec = \mathbf{E}\left[A = Y | A = 1\right] \ , \tag{10}$$

$$Rec = \mathbf{E}\left[A = Y | Y = 1\right] \ . \tag{11}$$

Klassifitseerimismeetodi kvaliteedimõõtude tegelikke väärtusi on praktikas peaaegu võimatu leida. Sageli leitakse neile lähendid rakendades meetodit lõplikul hulgal märgendatud testandmetel.

3.1 Õigsuse lähend

Õigsuse, täpsuse ja saagise puhul osutub nende kõigi analüüs analoogseks. Samas on õigsuse uurimine tehniliselt kõige lihtsam, sest definitsiooni järgi ei sisalda see tinglikku jaotust. Seetõttu on antud töös õigsust uuritud esimesena.

Meetodi rakendamise tulemust valimi i-ndal andmepunktil saab tõlgendada kui Bernoulli katset ehk Bernoulli jaotusega juhuslikku suurust $Z_i = [a_i = y_i]$, mille võimlalikud väärtused on 1 ja 0 vastavalt õige ning vale klassifikatsiooni korral. Meetodi tegelik õigsus Acc määrab juhusliku valimi korral iga katse õnnestumise tulemuse

$$Acc = \mathbf{E}\left[A = Y\right] = 0 \cdot \Pr\left[A \neq Y\right] + 1 \cdot \Pr\left[A = Y\right] = \Pr\left[a_i = y_i\right] \ .$$

Lähtudes eelnevast on meetodi õigsus (9) lähendatav statistilise tõenäosusena üle N andmepunkti suuruse valimi

$$\widehat{Acc} = \frac{1}{N} \cdot \sum_{i=1}^{N} Z_i = \frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = y_i] \ . \tag{12}$$

Kasutades keskväärtuse, dispersiooni ja Bernoulli jaotuse omadusi saab leida õigsuse lähendi keskväärtuse ning dispersiooni

$$\mathbf{E}\left[\widehat{Acc}\right] = \frac{1}{N} \cdot \mathbf{E}\left[\sum_{i=1}^{N} Z_i\right] = \frac{1}{N} \cdot N \cdot Acc = Acc \ ,$$
$$\mathbf{D}\left[\widehat{Acc}\right] = \frac{1}{N^2} \cdot \mathbf{D}\left[\sum_{i=1}^{N} Z_i\right] = \frac{1}{N} \cdot Acc \cdot (1 - Acc) \ .$$

Kuna suurus $\widehat{Acc}$ on ligikaudne väärtus tegelikust õigsusest, on otstarbekas valida piisavalt suur valim, et lähendi absoluutne viga $\Delta Acc = \widehat{Acc} - Acc$ oleks üle valimi küllaltki suure tõenäosusega absoluutväärtuselt võimalikult väike.

3.2 Täpsuse ja saagise lähendid

Populatsiooni või valimi korral, mille klasside sagedus ei ole tasakaalus, võib õigsus olla eksitav hinnang meetodi kvaliteedi kohta. Valimis, mille andmepunktidest 90% kuuluvad positiivsesse ning 10% negatiivsesse klassi, kõikide andmepunktide positiivseks klassifitseerimine annab õigsuse $Acc = 0{,}9$. Lisaks hindab õigsus kõiki klassifitseerimisel tehtud vigu ühtemoodi. Kui valepositiivne või valenegatiive klassifikatsioon võib põhjustada tõsiseid tagajärgi on oluline tehtud vigu üksteisest eristada.

Täpsus mõõdab päris positiivsete klassifikatsioonide sagedust ennustatud positiivsete seast. Täpsuse hindamiseks läbi statistilise tõenäosuse on kõigepealt vaja leida kuidas see avaldub tõenäosuste kaudu

$$\begin{aligned} Prec &= \mathbf{E}\left[A = Y | A = 1\right] = 0 \cdot \Pr\left[A \neq Y | A = 1\right] + 1 \cdot \Pr\left[A = Y | A = 1\right] \\ &= \Pr\left[Y = 1 | A = 1\right] = \frac{\Pr\left[Y = 1 \wedge A = 1\right]}{\Pr\left[A = 1\right]} \ . \end{aligned} \tag{13}$$

Tulemuse (13) põhjal on võimalik leida hinnang täpsusele (10) üle N andmepunkti suuruse valimi

$$\widehat{Prec} = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = 1] \cdot [y_i = 1]}{\frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = 1]} = \frac{\sum_{i=1}^{N} [a_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N} [a_i = 1]} \ . \tag{14}$$

Saagis on päris positiivsete klassifikatsioonide sagedus positiivsete andmepunktide seast. Sarnaselt täpsuse tõenäosusesitusele (13) saab saagise esitada tõenäosuste kaudu

$$Rec = \mathbf{E}\left[A = Y | Y = 1\right] = \frac{\Pr\left[Y = 1 \wedge A = 1\right]}{\Pr\left[Y = 1\right]} \ ,$$

mille põhjal on saagise (11) lähend üle valimi arvutatav järgnevalt

$$\widehat{Rec} = \frac{\frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = 1] \cdot [y_i = 1]}{\frac{1}{N} \cdot \sum_{i=1}^{N} [y_i = 1]} = \frac{\sum_{i=1}^{N} [a_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N} [y_i = 1]} \ . \tag{15}$$

Sellisel kujul esitatud lähendite tõenäosuslik hindamine võib olla keeruline, sest nii murru lugejas kui ka nimetajas on ligikaudsed suurused. Lihtsam on uurida lähendit üle valimi tinglikust jaotusest. Antud juhul on tingimus kvaliteedimõõdu definitsioonis olev sündmus, näiteks täpsuse puhul $A = 1$. Tinglikule jaotusele vastava valimi leidmiseks saab kasutada valikumeetodit (*rejection sampling*):

1. Võta jaotusest juhuslik andmepunkt.
2. Kontrolli andmepunkti vastavust tingimusele.
3. Kui andmepunkt vastab tingimusele võta see valimisse vastasel juhul mitte.
4. Korda kuni on leitud soovitud koguses tingimusele vastavaid andmepunkte.

Valikumeetodi põhjal on leitav valim A^+, mis koosneb positiivseks klassifitseeritud andmepunktidest, ning valim Y^+, mis koosneb positiivse klassiga andmepunktidest. Kasutades vastavaid valimeid on lähendid (14) ja (15) esitatavad kujul:

$$\widehat{Prec} = \frac{1}{|A^+|} \cdot \sum_{i \in A^+} [y_i = 1] \ , \tag{16}$$

$$\widehat{Rec} = \frac{1}{|Y^+|} \cdot \sum_{i \in Y^+} [a_i = 1] \ . \tag{17}$$

On oluline tähele panna, et lähendid (16) ja (17) on analoogsed õigsuse lähendile (12) ning on seetõttu on nende absoluutsed ja relatiivsed vead samasuguste omadustega.

Lihtsa valikumeetodi puhul võib osutuda probleemseks tingimuse kontrollimine valimi moodustamisel. Täpsuse puhul on tingimuse kontrollimine lihtne, sest piisab vaid meetodi rakendamisest andmepunktile. Saagise puhul ei pruugi meetod väga hästi toimida, sest tingimuse kontrollimiseks peab välja selgitama andmepunkti tegeliku klassi. Tegeliku klassi leidmine on võrreldes klassifikatsiooni leidmisega kulukas. Valikumeetodi rakendamine saagise lähendamiseks on eriti kulukas, kui positiivse klassiga andmepunktid jaotuses on haruldased. Leidub ülesandeid mille puhul võib positiivse juhtumi esinemissagedus olla $1 : 10000$ nagu tekstist faktide eraldamine. Sellisel juhul on 1000 elemendilise valimi leidiseks vaja märgendada 10 miljonit andmepunkti. Isesõitvate autode puhul võivad huvipakkuvad sündmused esineda ühel korral miljonist ning seega

on naiivse lähenemise korral vaja märgendada ligi miljard andmepunkti. Kuid vajaliku töökindluse saavutamiseks peab selliseid sündmusi ikkagi arvestama. See on ka üks põhjus, miks suure töökindlusega praktikas kasutatavate masinõppe algoritmide loomine on keerukas protsess.

3.3 Õigsuse lähendi absoluutne viga

Valimi põhjal leitud lähendid kvaliteedimõõtudele sisaldavad viga. Veendumaks lähendi vastavuses selle täpsele väärtusele tekib küsimus vea suuruse kohta. Vea arvutamiseks peab teadma täpset väärtust, mis ei pruugi olla võimalik. Täpset väärtust teadmata saab viga hinnata tõenäosuslikult. Sageli seatakse eesmärgiks hinnata kui suur on viga 95% juhtudest.

Üks viis vea tõenäosuslikuks hindamiseks on kasutada konsentratsioonivõrratusi, näiteks Höffdingi võrratust. Höffdingi võrratus on suurte arvude seaduse konkreetne erijuht, mis annab üldistest hinnangutest täpsemaid tõkkeid tõenäosustele. Höffdingi võrratus sätib ülemise tõkke tõenäosusele, et tõkestatud paarikaupa sõltumatute juhuslike suuruste summa erineb selle summa keskväärtusest (oodatud väärtusest) vähemalt mingi konstandi võrra [1, 2]. Täpsemalt, kui juhuslikud suurused $Z_1, Z_2, \ldots, Z_N$ on sõltumatud ning leiduvad tõkked

$$a_i \leq Z_i \leq b_i \ ,$$

siis summa $S_N = Z_1 + \cdots + Z_N$ ning iga positiivse c korral

$$\Pr\left[|S_N - \mathbf{E}\left[S_N\right]| \geq c\right] \leq 2\exp\left(-\frac{2c^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ .$$

Kuna Bernoulli jaotusega juhuslike suuruste jaoks leiduvad tõkked $0 \leq Z_i \leq 1$ järeldub Höffdingi võrratusest seos

$$\Pr\left[|\ \widehat{Acc} - Acc\ | \geq \frac{c}{N}\right] \leq 2\exp\left(-\frac{2c^2}{N}\right) \ , \tag{18}$$

mille põhjal saab hinnata lähendi $\widehat{Acc}$ absoluutse vea alumise tõkke tõenäosust. Sättides võrratuse (18) parema poole võrdseks olulisusega α avaldub veahinnangu alumine tõke kujul

$$\varepsilon := \frac{c}{N} = \sqrt{-\frac{1}{2N} \cdot \ln\left(\frac{\alpha}{2}\right)} \ .$$

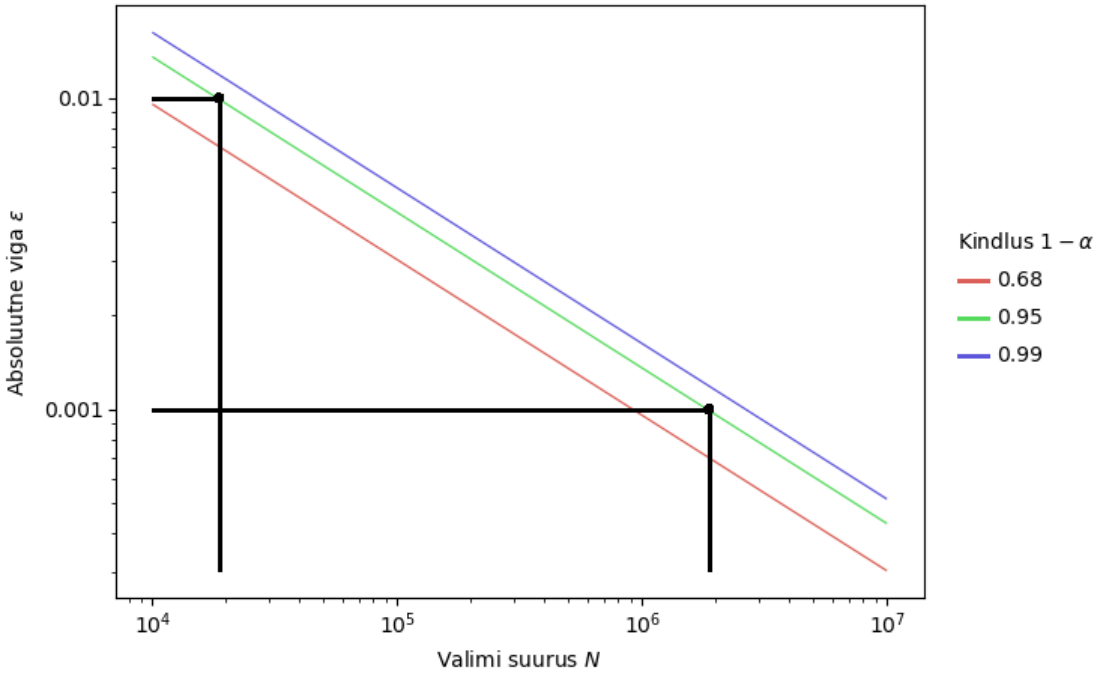
Tulemust kasutades on võimalik leida valimi vajalik suurus

$$N \geq -\frac{1}{2\varepsilon^2} \cdot \ln\left(\frac{\alpha}{2}\right) \ ,$$

et fikseeritud olulisuse korral saavutada soovitud suurusega veahinnang.

Tabel 1. Valimi suurused kindluse ning absoluutse vea suhtes Höffdinig võrratuse põhjal. Kindluse puhul on sulgudes tähistatud, mitu normaaljaotuse standardhälvet keskväärtuse ümbruses see katab.

$1-\alpha$	$\varepsilon = 10\%$	$\varepsilon = 1\%$	$\varepsilon = 0{,}1\%$
99% (3σ)	265	26492	2649159
95% (2σ)	185	18445	1844440
68% (1σ)	92	9163	916291



Joonis 4. Valimi suurused olulisuse ning absoluutse vea suhtes Höffdinig võrratuse põhjal kindlusega $1-\alpha$.

Tabelis 1 on välja toodud vajalik valimi suurus soovitud kindluse ja lähendi absoluutse vea suhtes Höffdingi võrratuse põhjal, kindluse puhul on sulgudes tähistatud, kui mitu normaaljaotuse standardhälvet σ keskväärtuse ümbruses see katab. Näiteks kui on soov olla 95% kindel, et valimi põhjal arvutatud õigsus $\widehat{Acc}$ erineb algoritmi tegelikust õigsusest kuni ühe protsendi võrra, peab õigsust hindama valimil suurusega vähemalt 18504. Joonisel 4 on sama mõttekäik esitatud graafiliselt.

Õigsuse lähendi absoluutse vea hindamiseks saab ka kasutada asjaolu, et suurused Z_i on Bernoulli jaotusega. See tähendab, et summa

$$S_N = \sum_{i=1}^{N} Z_i \ ,$$

on binoomjaotusega. Mingi binoomjaotusega juhusliku suuruse X puhul tähistatakse seda $X \sim \mathcal{B}(n,p)$, kus parameeter n on Bernoulli katsete arv ja p katse õnnestumise tõenäosus. Sellest lähtudes saab väita järgnevat

$$S_N \sim \mathcal{B}(N, Acc) \ .$$

Seega on õigsuse lähend $\widehat{Acc}$ skaleeritud binoomjaotusega juhuslik suurus. Eespool uuritud Höffdingi võrratusel põhinevad tõkked selle omadusega ei arvestanud ning olid binoomjaotuse parameetrist p sõltumatud, tegu oli konservatiivse hinnanguga. Höffidngi võrratus on universaalne üle kõikide binoomjaotuse parameetrite p, millest halvim variant realiseerub juhul $p = 0{,}5$. Binoomjaotusel põhinevad tõkked on täpsed ning annavad aimu Höffidngi võrratuse tulemuste ebatäpsustest. Lisaks näitavad binoomjaotusel põhinevad arvutused kuidas meetodi tegelik õigsus Acc mõjutab tulemusi.

Kasutades teadmist, et uuritav summa on binoomjaotusega saab leida jaotuse parameetritest sõltuva hinnangu. Olulisuse α korral saab absoluutse veahinnangu tõkke ε arvutada lähtudes võrrandist

$$\Pr\left[\mid \widehat{Acc} - Acc \mid \geq \varepsilon \right] = \alpha \ , \tag{19}$$

ning valides protsendipunktid sümmeetriliselt

$$\Pr\left[\widehat{Acc} \leq Acc - \varepsilon\right] = \frac{\alpha}{2} \ ,$$
$$\Pr\left[\widehat{Acc} < Acc + \varepsilon\right] = 1 - \frac{\alpha}{2} \ ,$$

millest binoomjaotusega juhusliku suuruse S_N eraldamisel ühele poole võrratuse märki saab

$$\Pr\left[S_N \leq N \cdot (Acc - \varepsilon)\right] = \frac{\alpha}{2} \ ,$$
$$\Pr\left[S_N < N \cdot (Acc + \varepsilon)\right] = 1 - \frac{\alpha}{2} \ .$$

Paremale poole võrratuse märki tekkinud avaldised on S_N jaotuse vastavalt $\frac{\alpha}{2}$ ja $1 - \frac{\alpha}{2}$ protsendipunktid, ehk väärtused, millest S_N võtab väiksemaid väärtusi protsendipunktile vastava tõenäosusega

$$q_1 = N \cdot (Acc - \varepsilon) \ ,$$
$$q_2 = N \cdot (Acc + \varepsilon) \ .$$

Fikseeritud olulisuse ja binoomjaotuse parameetrite korral on protsendipunkt arvutatav kasutades S_N jaotuse omadusi.

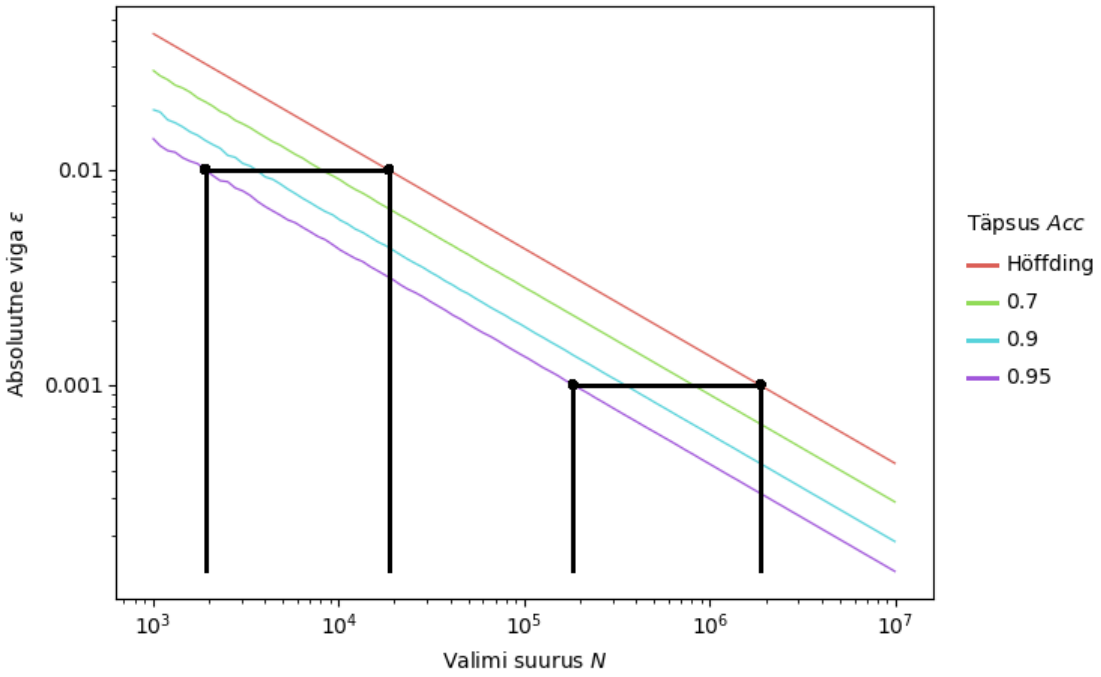
Absoluutse vea tõkkeks ε võrrandist (19) on valitud suurem protsentipunktide põhjal arvutatud veahinnang

$$\varepsilon = \max\left(Acc - \frac{q_1}{N}, \frac{q_2}{N} - Acc\right) \quad . \tag{20}$$

Parameetri N kasvades koondub binoomjaotus normaaljaotuseks [3]. Seega suuremate N väärtuste puhul muutub binoomjaotus sümmeetriliseks, järelikult erinevad q_1 ja q_2 põhjal arvutatud veahinnangute väärtused tegelikkuses vähe. Tulemust (20) kasutades saab arvutada vajaliku valim suuruse soovitud olulisuse suhtes, kuid selle jaoks peab tegema mudeli tegeliku õigsuse Acc kohta oletusi.

Tabel 2. Valimi suurused absoluutse vea ning oletatava õigsuse suhtes binoomjaotuse põhjal kindlusega 95%. Sulgudes on näidatud kui mitu korda Höffdingi võrratusel põhinev valim suurem on.

Acc	$\varepsilon = 10\%$	$\varepsilon = 1\%$	$\varepsilon = 0,1\%$
70%	75 (2,47)	8100 (2,28)	806289 (2,29)
90%	35 (5,29)	3600 (5,12)	347020 (5,32)
95%	20 (9)	1850 (9,97)	183285 (10,06)



Joonis 5. Valimi suurused absoluutse vea ning oletatava õigsuse suhtes binoomjaotuse põhjal kindlusega 95%.

Tabelis 2 ja joonisel 5 on esitatud valimi vajalikud suurused meetodi oletatud õigsuse ja absoluutse veahinnangu soovitud suuruse suhtes binoomjaotuse omaduste põhjal kindlusega 95%. Võrdluseks on välja toodud ka Höffdingi võrratusel põhinevad tulemused sama kindlusega, joonisel on see esitatud eraldi joonega, tabelis on sulgudes kirjas kui mitu korda on Höffdingi võrratuse põhine valim suurem. Eeldusel, et algoritmi tegelik õigsus on 90%, absoluutse vea kuni üks protsent jaoks läheb vaja valimit suurusega 3600, mis on umbes viis korda väiksem kui valim, mida on vaja Höffdingi võrratuse põhjal.

3.4 Õigsuse lähendi relatiivne viga

Lisaks ligikaudse väärtuse absoluutsele veale kasutatakse lähendi headuse mõõtmiseks relatiivset viga. Õigsuse lähendi relatiivse vea keskväärtus ja dispersioon avalduvad

järgnevalt

$$\mathbf{E}\left[\frac{\widehat{Acc}}{Acc}-1\right]=\frac{1}{Acc}\cdot\mathbf{E}\left[\widehat{Acc}\right]-1=1-1=0\ ,$$
$$\mathbf{D}\left[\frac{\widehat{Acc}}{Acc}-1\right]=\frac{1}{Acc^2}\cdot\frac{Acc\cdot(1-Acc)}{N}=\frac{1}{N}\cdot\frac{1-Acc}{Acc}\ .$$

Uurides lähendi relatiivse ja absoluutse vea hajuvuse suhet

$$\mathbf{D}\left[\frac{\widehat{Acc}}{Acc}\right]:\mathbf{D}\left[\widehat{Acc}-Acc\right]=\left(\frac{1}{N}\cdot\frac{1-Acc}{Acc}\right):\left(\frac{1}{N}\cdot Acc\cdot(1-Acc)\right)=\frac{1}{Acc^2}\ ,$$

selgub, et kõrge õigsuse korral on õigsuse lähendi vead ligikaudu sama hajuvusega.

Nagu absoluutse veahinnangu puhul on ka relatiivse veahinnangu puhul eesmärk seda absoluutväärtuselt minimeerida. Seega võib küsida kui suurt valimit läheb vaja, et piisavalt suure kindlusega oleks relatiivne viga võimalikult väike.

Õigsuse lähend sisaldab binoomjaotusega juhuslikku suurust. Järelikult on võimalik relatiivset viga tõenäosuslikult hinnata kasutades binoomjaotuse omadusi. Lähtudes võrrandist

$$\Pr\left[\left|\frac{\widehat{Acc}}{Acc}-1\right|\geq\varepsilon\right]=\alpha\ , \tag{21}$$

millest tõenäosusmärgi aluses võrratuses eraldada binoomjaotusega juhusliku suuruse S_N ühele poole võrratusemärki

$$\Pr\left[S_N\leq N\cdot Acc\cdot(1-\varepsilon)\right]=\frac{\alpha}{2}\ ,$$
$$\Pr\left[S_N<N\cdot Acc\cdot(1+\varepsilon)\right]=1-\frac{\alpha}{2}\ ,$$

avalduvad olulisusele vastavad protsendipunktid kujul

$$q_1=N\cdot Acc\cdot(1-\varepsilon)\ ,$$
$$q_2=N\cdot Acc\cdot(1+\varepsilon)\ .$$

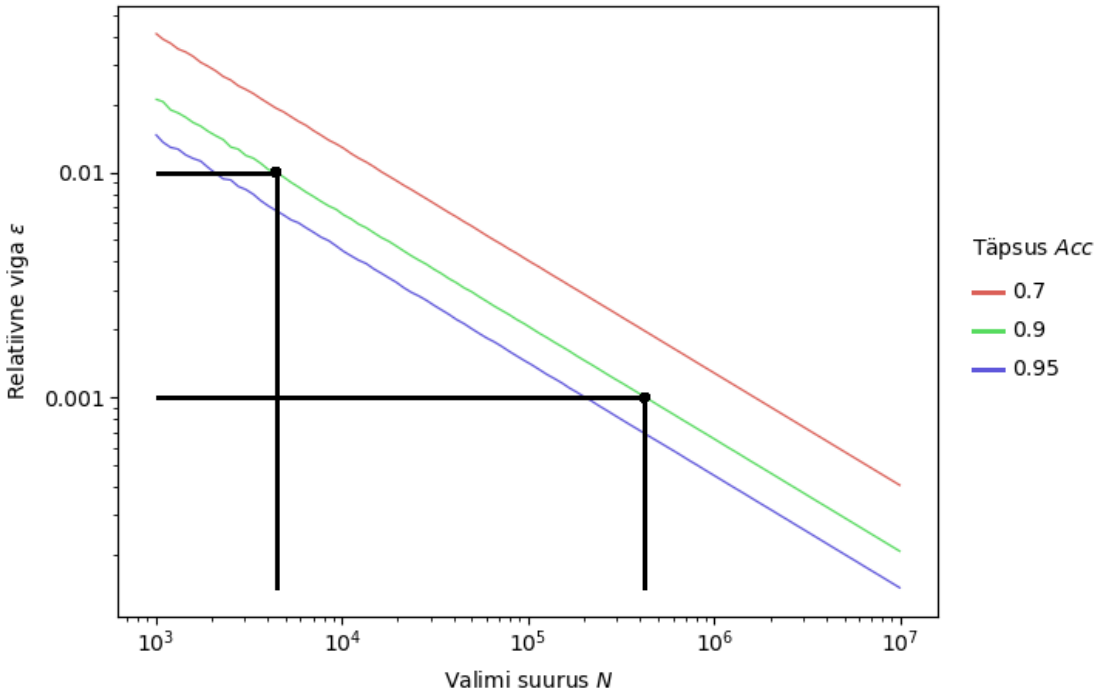
Relatiivse vea tõkkeks ε võrrandist (21) on valitud suurem protsendipunktide põhjal avalduv veahinnang

$$\varepsilon=\max\left(1-\frac{q_1}{N\cdot Acc},\frac{q_2}{N\cdot Acc}-1\right)\ .$$

Tulemuse põhjal saab arvutada valimi vajalikud suurused olulisuse suhtes.

Tabel 3. Valimi suurused relatiivse vea ning eeldatud õigsuse suhtes binoomjaotuse põhjal kindlusega 95%.

Acc	$\varepsilon = 10\%$	$\varepsilon = 1\%$	$\varepsilon = 0{,}1\%$
70%	161	16331	1646727
90%	53	4176	425856
95%	20	1926	202223

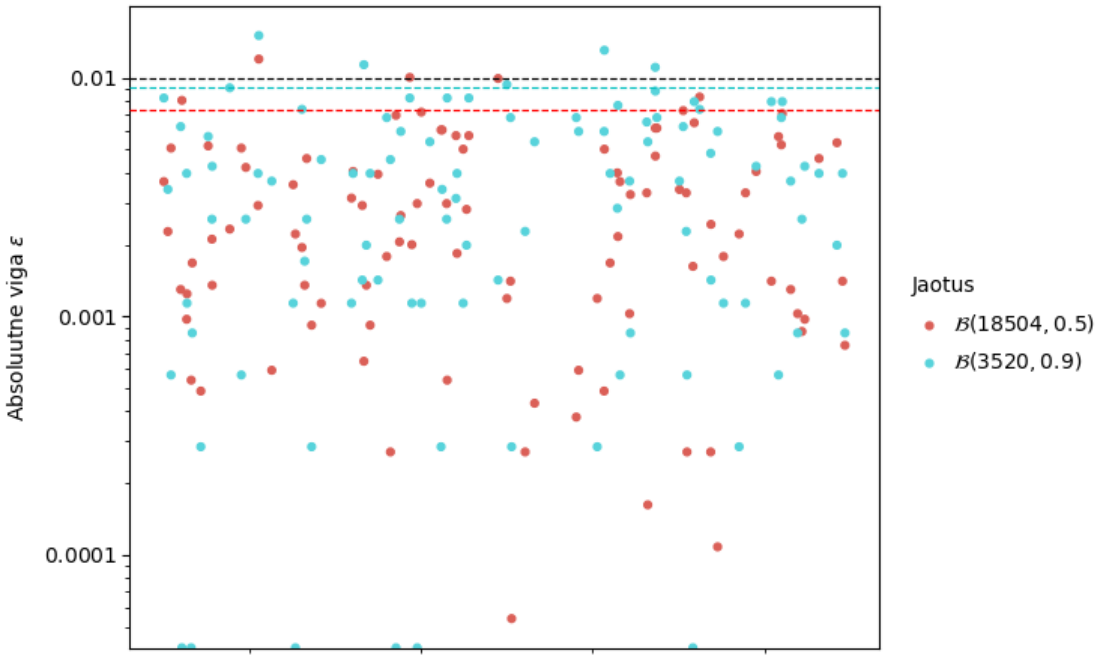


Joonis 6. Valimi suurused relatiivse vea ning eeldatud õigsuse suhtes binoomjaotuse põhjal kindlusega 95%.

Tabelis 3 ja joonisel 6 on esitatud valimi vajalikud suurused meetodi oletatud õigsuse ja relatiivse veahinnangu soovitud suuruse suhtes binoomjaotuse omaduste põhjal kindlusega 95%.

3.5 Tulemuste empiiriline testimine

Teoreetilisi tulemusi on alati hea praktikas kontrollida. Niimoodi on võimalik leide lihtsasti valideerida ning ka avastada enamuse arvutusvigadest.



Joonis 7. Höffdingi võrratusel (punane) ja binoomjaotusel (sinine) põhinevate veahinnangute empiiriline test lähendi veaga kuni 1% saamiseks kindlusega 95%. Binoomjaotusega summad S_N õigsuse hinnangutes on juhuslikult genereeritud vastavatest jaotustest, punktid joonisel tähistavad summale vastava õigsuse hinnangu absoluutse vea absoluutväärtust. Värvilised jooned tähistavad empiiriliselt saadud $0{,}95$ protsendipunkte, must joon nende oodatud asukohta.

Näiteks Höffdingi võrratuse põhjal kindlusega 95% kuni ühe protsendise absoluutse veahinnanguga õigsuse lähendi saavutamiseks on vaja valimit suurusega 18504. Eeldusel, et klassifitseerimismeetodi tegelik õigsus on 90% on binoomjaotuse hinnangu põhjal vaja selleks 3520 andmepunkti. Joonisel 7 on kujutatud antud näite empiiriline test, kus summa

$$S_N = \sum_{i=1}^{N} Z_i \ ,$$

on juhuslikult genereeritud jaotusest $\mathcal{B}(18504; 0{,}5)$ Höffdingi võrratuse puhul (punane) ning $\mathcal{B}(3520; 0{,}9)$ binoomjaotuse hinnangu puhul (sinine). Punktid joonisel tähistavad summal põhineva lähendi absoluutset viga

$$\varepsilon = \left| \frac{S_N}{N} - p \right| = \left| \widehat{Acc} - Acc \right| \ ,$$

must joon veahinnangu absoluutväärtuse ülemist tõket, millest saadud veahinnang ε peaks olema väiksem 95% juhtudest. Punane ja sinine joon on vastava jaotuse puhul tõmmatud läbi empiirilise testimise tulemusel saadud punkti, millest 95% jäävad allapoole. Binoomjaotuse hinnang on täpne ning langes ka antud juhul oodatule lähedale. See-eest langes punane joon oodatust allapoole. Kuna Höffdingi võrratus ülehindab valimi valjalikku suurust on punase joone langemine tema oodatud asukohast allapoole loomulik, sest suurem valim tähendab täpsemat lähendit ja seega väiksemat viga.

4 Masinõppe meetodite kvaliteedimõõtude võrdlus

Uue meetodi kasutusele võtmine tähendab, et on olemas vanem juba kasutuses meetod. Võib osutuda otstarbekaks kasutada vanemat meetodit uue hindamisel. Vana meetodi asendamisel on eelkõige tähtis kas ja kui palju on uus meetod eelmisest parem. Väljaselgitamiseks võib uurida meetodite kvaliteedimõõtude vahesid.

Järgnevas laiendame eelmises peatükis sisse toodud tähistusi. Tähistagu i-ndat andmepunkti elementide paar (x_i, y_i), kus x_i tähistab andmepunkti tunnuseid ning y_i andmepunkti klassi. Olgu A uus ning B varasem klassifitseerimismeetod, siis saame tähistada andmepunktile vastavad klassifikatsioone $A(x_i) = a_i$ ja $B(x_i) = b_i$. Vaadeldes andmepunkti juhusliku suurusena (X, Y) võib ka meetodite klassifikatsioone sellel vaadelda juhuslike suurustena A ja B ning meetodite kvaliteedimõõdud on defineeritud keskväärtuste kaudu läbi valemite (9), (10) ja (11). Näiteks meetodi B õigsus on $Acc_B = \mathbf{E}[B = Y]$. Siit lähtuvalt saab meetodite erinevuse uurimisel vaadata meetodite kvaliteedimõõtude vahesid

$$\Delta Acc = \mathbf{E}[A = Y] - \mathbf{E}[B = Y] \quad , \tag{22}$$

$$\Delta Prec = \mathbf{E}[A = Y | A = 1] - \mathbf{E}[B = Y | B = 1] \quad , \tag{23}$$

$$\Delta Rec = \mathbf{E}[A = Y | Y = 1] - \mathbf{E}[B = Y | Y = 1] \quad . \tag{24}$$

4.1 Kvaliteedimõõtude vahede lähendid

Kuna defineeritud vahede täpseid väärtusi on praktikas keeruline leida hinnatakse neid kasutades valimi keskmisi. Sealjuures on loomulik mõlema meetodi hindamiseks kasutada sama valimit. Ühelt poolt on see standardpraktika - tüüpiliselt hinnatakse masinõppemeetodite edukust fikseeritud testvalimite (*benchmark*) peal. Teisalt tooks kahe valimi kasutamine vajaduse käsitsi märgendada rohkem andmepunkte ning lisab ka lõpptulemusse rohkem juhuslikkust. Kui vastav testvalim sisaldab N andmepunkti, siis saab õigsuste vahe lähendi avaldada järgnevalt

$$\begin{aligned}
\widehat{\Delta Acc} &= \frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = y_i] - \frac{1}{N} \cdot \sum_{i=1}^{N} [b_i = y_i] \\
&= \frac{1}{N} \cdot \sum_{i=1}^{N} [a_i = y_i] - [b_i = y_i] \\
&= \frac{1}{N} \cdot \sum_{a_i \neq b_i} [a_i = y_i] - [b_i = y_i] \quad .
\end{aligned} \tag{25}$$

Saadud võrduses (25) on summa märgi all nullist erinev arv parajasti siis, kui meetodite klassifikatsioonid on erinevad. Sellest järeldub, et meetodite õigsuste vahe hindamiseks

peab märgendama vaid andmepunkte, kus $a_i \neq b_i$. Näiteks kui meetodid on üle 90% õigsustega võivad nende klassifikatsioonid erineda maksimaalselt 20% andmetest. Sellisel juhul peab märgendama vaid iga viienda andmepunkti. Veelgi enam, kuna iga summa liige valemis (25) on ± 1, saab hinnata $\Delta\widehat{Acc}$ ilma ühtegi andmepunkti märgendamata

$$|\Delta\widehat{Acc}| \leq \frac{\#\{i : a_i \neq b_i\}}{N} \ , \tag{26}$$

kus murru lugejas tähistab $\#$ hulga suurust. Seega on võimalik veenduda, kas kaks algoritmi üldse õigsuse poolest erinevad andmepunktide tegelikke klasse teadmata.

Analoogselt õigsusele võib avaldada ka valimipõhiste täpsushinnangute vahe

$$\Delta\widehat{Prec} = \frac{\sum_{i=1}^{N}[a_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N}[a_i = 1]} - \frac{\sum_{i=1}^{N}[b_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N}[b_i = 1]} \ , \tag{27}$$

mille edasise lihtsustamise muudab raskeks erinevus murru nimetajates. Kuna tehniliselt ei pea täpsuse hindamisel kasutama sama testvalimit mõlema algoritmi jaoks, siis võib esialgset testvalmit kitsendada nii, et murru lugejad langevad kokku

$$\sum_{i=1}^{N_A}[a_i = 1] = S = \sum_{i=1}^{N_B}[b_i = 1] \ ,$$

ning $N = \max(N_A, N_B)$. Selline lähenemine on samaväärne valikumeetodi põhjal kahe S andmepunktise valimi moodustamisega, kus vastuvõtutingimused on vastavalt $a_i = 1$ ja $b_i = 1$. Seda arvestades saab valemi (27) viia lihtsustatud kujule

$$\Delta\widehat{Prec} = \frac{1}{S} \cdot \sum_{i=1}^{N_A}[a_i = 1] \cdot [y_i = 1] - \frac{1}{S} \cdot \sum_{i=1}^{N_B}[b_i = 1] \cdot [y_i = 1] \ , \tag{28}$$

kus a_i on määratud vaid esimesel N_A ja b_i on määratud vaid esimesel N_B andmepunktil. Seega on a_i ja b_i väärtus kõigi N punkti seas kas määramata, 0 või 1. Siit lähtuvalt

$$\begin{aligned}
\Delta\widehat{Prec} &= \frac{1}{S} \cdot \left(\sum_{\substack{a_i=1 \\ b_i=1}} [y_i = 1] + \sum_{\substack{a_i=1 \\ b_i\neq 1}} [y_i = 1] - \sum_{\substack{a_i=1 \\ b_i=1}} [y_i = 1] - \sum_{\substack{a_i\neq 1 \\ b_i=1}} [y_i = 1] \right) \\
&= \frac{1}{S} \cdot \left(\sum_{\substack{a_i=1 \\ b_i\neq 1}} [y_i = 1] - \sum_{\substack{a_i\neq 1 \\ b_i=1}} [y_i = 1] \right) \\
&= \frac{1}{S} \cdot \sum_{a_i \neq b_i} [a_i = 1] \cdot [y_i = 1] - [b_i = 1] \cdot [y_i = 1] \ .
\end{aligned} \tag{29}$$

Andmepunktide märgendamise mõttes on täpsuse vahe lähend (29) samaväärne õigsuse valemiga (25), sest hinnangu arvutamiseks peab märgendama vaid andmepunkte, mille puhul $a_i \neq b_i$.

Täpsuse vahehinnangu praktiline arvutamine algab S fikseerimisest. Seejärel tuleb valimile rakendada meetodeid kuni kumbki on S andmepunkti positiivseks klassifitseerinud. Märgendama peab valimi andmepunktid, mille puhul on olemas mõlema meetodi klassifikatsioonid, mis on üksteisest erinevad. Nüüd on võimalik arvuta summa liikmete väärtused ning seejärel hinnang täpsuste vahele. Sellega on oluliselt vähendatud märgendamist vajavate andmete hulka.

Saagise vahe avaldub sarnaselt täpsusele, lähtudes valemist

$$\Delta\widehat{Rec} = \frac{\sum_{i=1}^{N}[a_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N}[y_i = 1]} - \frac{\sum_{i=1}^{N}[b_i = 1] \cdot [y_i = 1]}{\sum_{i=1}^{N}[y_i = 1]} \ ,$$

kus erinevalt täpsusest on murdude nimetajad definitsiooni järgi võrdsed. Tähistagu nimetajas olevat summat

$$T = \sum_{i=1}^{N}[y_i = 1] \ .$$

Korrates täpsuse valemi (29) tuletamisega analoogset mõttekäiku, võib esitada saagise vahe hinnangu kujul

$$\Delta\widehat{Rec} = \frac{1}{T} \cdot \sum_{a_i \neq b_i} [a_i = 1] \cdot [y_i = 1] - [b_i = 1] \cdot [y_i = 1] \ . \tag{30}$$

Märgendamise seisukohalt on T arvutamine kulukas, kuna iga summa liikme puhul peab teadma andmepunkti tegelikku klassi. Kõigi N andmepunkti manuaalne märgendamine on väga ressursimahukas. Alternatiiviks on väiksema valimi põhjal positiivse klassi esinemise sageduse ennustamine. See on võimalik vaid siis kui positiivse klassi esinemise sagedus on piisavalt suur. Kui positiivse klassi esindajad on sagedusega $1 : 1000$, siis on tarvis adekvaatse hinnangu saamiseks läbi vaadata üle kümnetuhande andmepunkti. Seega on valemi (30) praktiline rakendamine raskendatud veelgi madalama esinemissagedusega sündmuste korral.

4.2 Õigsuste vahe lähendamine

Siiani on lähendid kvaliteedimõõtude vahele avaldatud vahetult läbi vahe oodatud väärtuse definitsiooni. Tehes teisendusi vahe definitsioonis antud avaldises on võimalik see eraldada mitmeks komponendiks. Vahe lähendi saamiseks on ka võimalik lähendada iga komponenti eraldi ning nende põhjal arvutada hinnang vahele.

Olgu olemas kaks klassifitseerimismeetodit koos juhusliku andmepunkti Y klassifitseerimisele vastavate juhuslike suurustega A ja B. Siis saab valemi (22) kirjutada lahti vastavalt definitsioonile

$$\Delta Acc = \mathbf{E}\left[A = Y\right] - \mathbf{E}\left[B = Y\right] = \mathbf{E}\left[\left[A = Y\right] - \left[B = Y\right]\right] \ . \tag{31}$$

Tähistagu Z valemis (31) viimase keskväärtusmärgi all olevat juhuslikkus suurust, ehk $Z = [A = Y] - [B = Y]$. Analoogselt võib valemis (25) oleva summa liikmeid tõlgendada juhuslike suurustena $Z_i = [a_i = y_i] - [b_i = y_i]$. Arvestades, et Z_i on sõltumatud ning sama jaotusega kui Z, on lihtne näha, et lähendi (25) keskväärtus on

$$\mathbf{E}\left[\widehat{\Delta Acc}\right] = \frac{1}{N} \cdot \sum_{i=1}^{N} \cdot \mathbf{E}\left[Z_i\right] = \Delta Acc \ . \tag{32}$$

See tähendab, et $\widehat{Acc}$ on nihketa hinnang õigsuste vahele. Kuna meetodite erinevus avaldub sündmustes, kus meetodite klassifikatsioonid erinevad ($Z \neq 0$), võib ΔAcc avaldada selliste sündmuste kaudu

$$\begin{aligned} \Delta Acc &= \Pr\left[Z = 0\right] \cdot \mathbf{E}\left[Z \mid Z = 0\right] + \Pr\left[Z \neq 0\right] \cdot \mathbf{E}\left[Z \mid Z \neq 0\right] \\ &= \Pr\left[Z \neq 0\right] \cdot \mathbf{E}\left[Z \mid Z \neq 0\right] \ . \end{aligned}$$

Edasiste arvutuste selgemaks esitamiseks on otstarbekas kasutusele võtta tähistused

$$\beta = \Pr\left[Z \neq 0\right] \ , \tag{33}$$

$$\gamma = \mathbf{E}\left[Z \mid Z \neq 0\right] \ , \tag{34}$$

$$\kappa = \Pr\left[Z = 1 | Z \neq 0\right] \ . \tag{35}$$

Need kolm suurust pole sõltumatud parameetrid. Keskväärtuse definitsioonist lähtuvalt on γ ja κ omavahel seotud

$$\gamma = \mathbf{E}\left[Z \mid Z \neq 0\right] = 1 \cdot \kappa - 1 \cdot (1 - \kappa) = 2\kappa - 1 \ ,$$

ning me saame esitada õigsuste vahe antud suuruste kaudu

$$\Delta Acc = \beta \cdot \gamma = \beta \cdot (2\kappa - 1) \ . \tag{36}$$

Võrdusest (36) lähtuvalt on võimalik hinnata meetodite õigsuste vahet lähendades suurusi β ja γ. Kuna $\beta = \Pr\left[Z \neq 0\right]$ on tõenäosus saab seda hinnata statistilise tõenäosusena üle N elemendilise valimi

$$\hat{\beta} = \frac{1}{N} \cdot \sum_{i=1}^{N} [Z_i \neq 0] = \frac{1}{N} \cdot \sum_{i=1}^{N} [a_i \neq b_i] \ , \tag{37}$$

[4]. Lähendi leidmisel summa märgi alune juhuslik suurus võtab väärtusi 0 ja 1. See tähendab, et lähendi absoluutse ja relatiivse vea hinnangud on leitavad eelmises peatükis kirjeldatud meetodeid kasutades.

Lähend tinglikule keskväärtusele $\gamma = \mathbf{E}[Z \mid Z \neq 0]$ on leitav sarnaselt, valimikeskimisena, kus iga valimi andmepunkti korral $a_i \neq b_i$. Olgu K sellise valimi suurus, siis saab lähendi esitada kujul

$$\hat{\gamma} = \frac{1}{K} \cdot \sum_{i=1}^{K} [a_i = y_i] - [b_i = y_i] \ . \tag{38}$$

Kuna lähendis $\hat{\gamma}$ on summa liikmete võimalikud väärtused -1 ja 1, ei ole summa liikmed Bernoulli jaotusega. See-eest on ikka tegu binaarse tunnusega, mille põhjal saab defineerida uue Bernoulli jaotusega juhusliku suuruse, mis võtab väärtuse 0 kui summeritav suurus võtab -1 ning muidu 1

$$W = \frac{Z+1}{2} \ ,$$

mille puhul

$$\begin{aligned} \Pr[W = 1] &= \Pr[Z = 1 \mid Z \neq 0] = \kappa \ , \\ \Pr[W = 0] &= \Pr[Z = -1 \mid Z \neq 0] = 1 - \kappa \ . \end{aligned}$$

Sündmusena on uus defineeritud suurus samaväärne vanaga, endiselt saab rakendada binoomjaotusel põhinevaid tulemusi eelmisest peatükist. Lähtudes võrrandist

$$\Pr\left[\left|\frac{\hat{\gamma}}{\gamma} - 1\right| \geq \varepsilon\right] = \alpha \ ,$$

korrates sama mõttekäiku, mis valemi (21) puhul.

Leitud lähendite korrutis annab omakorda lähendi meetodite õigsuste vahele

$$\Delta\widehat{Acc} = \hat{\beta} \cdot \hat{\gamma} \ .$$

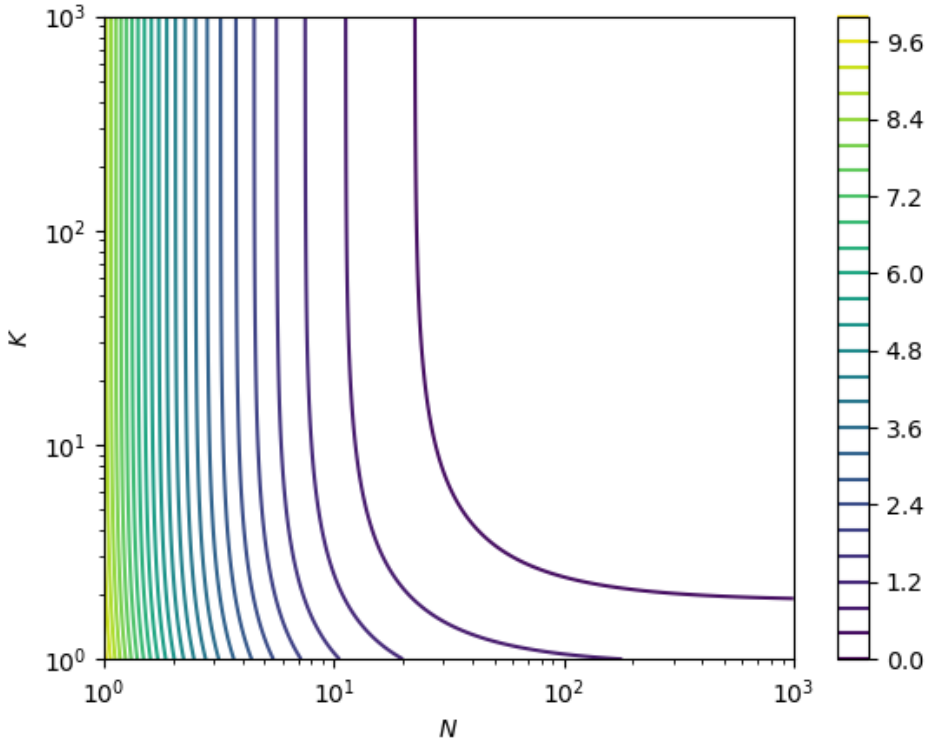
Vahe lähendi relatiivse vea dispersioon avaldub relatiivse vea korrutise omaduse (5) põhjal kujul

$$\mathbf{D}\left[\frac{\hat{\beta} \cdot \hat{\gamma}}{\beta \cdot \gamma}\right] \approx \mathbf{D}\left[\frac{\hat{\beta}}{\beta}\right] + \mathbf{D}\left[\frac{\hat{\gamma}}{\gamma}\right] \ , \tag{39}$$

kus hinnangute $\hat{\beta}$ ja $\hat{\gamma}$ relatiivsete vigade dispersioonid on

$$\mathbf{D}\left[\frac{\hat{\beta}}{\beta}\right] = \frac{1}{N} \cdot \frac{1-\beta}{\beta} \ , \tag{40}$$

$$\mathbf{D}\left[\frac{\hat{\gamma}}{\gamma}\right] = \frac{1}{K} \cdot (\mathbf{E}\left[Z^2 \mid Z \neq 0\right] - \mathbf{E}[Z \mid Z \neq 0]^2) = \frac{1}{K} \cdot (1 - \gamma^2) \ . \tag{41}$$



Joonis 8. Õigsuste vahe lähendi relatiivse vea dispersioon juhusliku valimi suuruse N ning märgendatud tingliku $(a_i \neq b_i)$ valimi suuruse K suhtes. Juhul kui $\Delta Acc = 5\%$ ja $\beta = 10\%$. Heledus tähistab dispersiooni.

Jooniselt 8 ning vahetulemuste (40) ja (41) kaudu on näha, et $\Delta\widehat{Acc}$ relatiivse vea dispersioon kahaneb valimisuuruste kasvades. Telgedel olevatest suurustest tähistab N valimi suurust, mida ei pea märgendama. Kuna sellise valimi leidmine ei ole keeruline võib eeldada, et N on fikseeritud ja küllaltki suur. Oluline on aga suurus K, mis tähistab märgendamist vajavate andmepunktide arvu. Märgendatud valimi suuruse prognoosimiseks fikseeritud N põhjal võib uurida olukorda, kus saavutatakse võrdsed lähendite $\hat{\beta}$ ja $\hat{\gamma}$ dispersioonid

$$\frac{1}{N} \cdot \frac{1-\beta}{\beta} = \frac{1}{K} \cdot (1-\gamma^2) \ . \tag{42}$$

Kuna $\Delta Acc = \beta \cdot \gamma$, on valem (42) esitatav kujul

$$K = N \cdot \frac{\beta}{1-\beta} \cdot \left(1 - \frac{(\Delta Acc)^2}{\beta^2}\right) \ , \tag{43}$$

mille põhjal on võimalik hinnata märgendatud valimi suurust.

4.3 Saagiste vahe lähendamine

Saagise puhul on hinnangu (30) analüüsimine keerulisem, kuna saagise definitsioonis on tinglik keskväärtus. Valemile (36) analoogi tuletamiseks peab kõigepealt eraldama saagise definitsioonist tingliku jaotuse

$$\begin{aligned}\mathbf{E}\left[A = Y \mid Y = 1\right] &= \mathbf{E}\left[A = 1 \mid Y = 1\right] = \Pr\left[A = 1 \mid Y = 1\right] \ , \\ \mathbf{E}\left[B = Y \mid Y = 1\right] &= \mathbf{E}\left[B = 1 \mid Y = 1\right] = \Pr\left[B = 1 \mid Y = 1\right] \ ,\end{aligned}$$

millest saab tuletada

$$\mathbf{E}\left[A = Y \mid Y = 1\right] = \frac{\Pr\left[A = 1 \wedge Y = 1\right]}{\Pr\left[Y = 1\right]} = \frac{\mathbf{E}\left[[A = 1] \cdot [Y = 1]\right]}{\mathbf{E}\left[Y = 1\right]} \ .$$

Analoogselt meetodi B puhul

$$\mathbf{E}\left[A = Y \mid Y = 1\right] = \frac{\Pr\left[B = 1 \wedge Y = 1\right]}{\Pr\left[Y = 1\right]} = \frac{\mathbf{E}\left[[B = 1] \cdot [Y = 1]\right]}{\mathbf{E}\left[Y = 1\right]} \ .$$

Nende kahe võrduse põhjal võib saagiste vahe esitada kujul

$$\begin{aligned}\Delta Rec &= \mathbf{E}\left[A = Y \mid Y = 1\right] - \mathbf{E}\left[B = Y \mid Y = 1\right] \\ &= \frac{\mathbf{E}\left[[A = 1] \cdot [Y = 1] - [B = 1] \cdot [Y = 1]\right]}{\mathbf{E}\left[Y = 1\right]} \ .\end{aligned}$$

Sellest saab järeldada algse saagise lähendi valemi (15) põhjal, et ka saagiste vahe lähend (30) on nihketa. Kuid sellisel kujul esitatud vahele ΔRec ei ole lähendi leidmine märgendamise seisukohalt veel kõige parem.

Kehtivad seosed

$$\begin{aligned}\mathbf{E}\left[[A = 1][Y = 1]\right] &= \mathbf{E}\left[[A = 1][Y = 1][B = 1]\right] + \mathbf{E}\left[[A = 1][Y = 1][B = 0]\right] \ , \\ \mathbf{E}\left[[B = 1][Y = 1]\right] &= \mathbf{E}\left[[B = 1][Y = 1][A = 1]\right] + \mathbf{E}\left[[B = 1][Y = 1][A = 0]\right] \ ,\end{aligned}$$

nende seoste ja keskväärtuse definitsiooni põhjal on saagise vahe lugeja esitatav korrutisena tõenäosusest ja tinglikust keskväärtusest

$$\begin{aligned}&\mathbf{E}\left[[A = 1][Y = 1] - [B = 1][Y = 1]\right] \\ &= \mathbf{E}\left[[A = 1][Y = 1][A \neq B] - [B = 1][Y = 1][A \neq B]\right] \\ &= \Pr\left[A \neq B\right] \cdot \mathbf{E}\left[[A = 1][Y = 1] - [B = 1][Y = 1] \mid [A \neq B]\right] \ .\end{aligned}$$

Edasiste arvutuste selgemaks esitamiseks on jällegi hea kasutusele võtta tähistused

$$\begin{aligned}\beta &= \Pr\left[A \neq B\right] \ , \\ \nu &= \Pr\left[Y = 1\right] \ , \\ \eta &= \mathbf{E}\left[[A = 1][Y = 1] - [B = 1][Y = 1] \mid [A \neq B]\right] \ ,\end{aligned}$$

mille kaudu saab esitada saagiste vahe

$$\Delta Rec = \frac{\beta \cdot \eta}{\nu} \ . \tag{44}$$

Tulemuse (44) põhjal on võimalik meetodite saagiste vahet hinnata lähendades suurusi β, ν ja η. Tõenäosuse β puhul on hinnang leitav samamoodi nagu (37) eelmises alampeatükis. Ning ν lähend sarnaselt üle N suuruse juhusliku valimi

$$\hat{\beta} = \sum_{i=1}^{N} [a_i \neq b_i] \ ,$$
$$\hat{\nu} = \sum_{i=1}^{N} [y_i = 1] \ .$$

Tinglikku keskväärtust η on võimalik lähendada valimikeskmisena üle tingliku valimi, kus $a_i \neq b_i$. Olgu selle valimi suurus K, siis

$$\hat{\eta} = \frac{1}{K} \cdot \sum_{i=1}^{K} [a_i = 1][y_i = 1] - [b_i = 1][y_i = 1] \ .$$

Kuna hinnangute $\hat{\beta}$ ja $\hat{\nu}$ puhul on summa liikmete võimalikud väärtused 0 ja 1, on lähendite absoluutse ja relatiivse vea hinnangud leitavad meetoditega eelmisest peatükist. Kuna lähendi $\hat{\eta}$ puhul on summa liikmete võimalikud väärtused -1, 0, ja 1, ei saa rakendada binoomjaotusel põhinevaid tulemusi. See-eest on endiselt võimalik kasutada Höffdinig võrratust lähendi absoluutse vea jaoks, seekord tõketega -1 ja 1, mille põhjal avaldub valimi vajalik suurus

$$K \geq -\frac{2}{\varepsilon^2} \cdot \ln\left(\frac{\alpha}{2}\right) \ ,$$

kus ε tähistab absoluutse vea absoluutväärtuse maksimaalset lubatud suurust ning α olulisust.

Relatiivse vea korrutise (5) ja jagatise (8) dispersiooni omaduste põhjal võib saagise vahe hinnangu relatiivse vea dispersiooni esitada summana

$$\mathbf{D}\left[\frac{\hat{\beta} \cdot \hat{\eta}}{\hat{\nu}} : \frac{\beta \cdot \eta}{\nu}\right] \approx \mathbf{D}\left[\frac{\hat{\beta}}{\beta}\right] + \mathbf{D}\left[\frac{\hat{\nu}}{\nu}\right] + \mathbf{D}\left[\frac{\hat{\eta}}{\eta}\right] \ ,$$

millest

$$\mathbf{D}\left[\frac{\hat{\beta}}{\beta}\right] = \frac{1}{N} \cdot \frac{1-\beta}{\beta}, \qquad \mathbf{D}\left[\frac{\hat{\nu}}{\nu}\right] = \frac{1}{N} \cdot \frac{1-\nu}{\nu} \ .$$

4.4 Täpsuste vahe lähendamine

Täpsuse jaoks on võrrandi (44) analoogi tuletamine keeruline, sest täpsuste vahe lähendis on summa liikmed mittesõltumatud. Üheks võimalikuks lahenduseks täpsuse hindamisel on kahesammuline lähenemine, kus hinnatakse esmalt lähendi (27) dispersiooni funktsioonina valikumeetodi abil saadud S elemendilisest valimist (kahe keskmise vahe) ning seejärel lähendatakse selle optimeeritud esitust (29) veelgi väiksema K-elemendilise juhuvalimiga. Kuna saadud lahendus ei ole elegantne ning kõik sammud iseseisvalt on analoogsed peatükis 3 saadud tulemustega, siis ei ole seda antud töös pikemalt käsitletud.

5 Kokkuvõte

Töös leiti vajalikud valimi suurused kindluse, veahinnangu lubatud maksimaalse suuruse ja mõnel juhul oletatud tegeliku kvaliteedimõõdu suhtes.Höffdingi võrratuse põhjal leitud tulemused ülehindasid vajaliku valimi suurust võrreldes binoomjaotuse omaduste põhjal arvutatud tulemustega, halvimal juhul isegi suurusjärgu võrra. Tehes arvutusi binoomjaotuse omaduste põhjal, pidi oletama kvaliteedimõõdu tegelikku väärtust. Seeeest andis Höffdingi võrratus universaalse hinnangu, mis oli sõltumatu meetodi tegelikust kvaliteedimõõdust (binoomjaotuse parameetrist p).

Tehtud töös uuriti ka tehnikaid kahe masinõppe meetodi võrdlemiseks kvaliteedimõõtude põhjal. Uue mudeli kasutusele võtmisel oli küsimuseks kui palju see vanast parem on. Vastust küsimusele otsiti meetodite kvaliteedimõõtude vahede kaudu. Õigsuse ja täpsuse puhul oli võimalik vahele hinnangut leida niimoodi, et manuaalselt peab märgendama vaid andepunkte, mille puhul meetodite klassifikatsioonid erinevad. Selle tulemusel oli võimalik vähendada märgendamist vajavate andmepunktide arvu. Saagise puhul on osa lähendist avaldatav sarnaselt õigsusele ja täpsusele. Kuid saagise arvutamiseks peab ikkagi teadma kõikide andmepunktide tegelikke klassimärgendeid, seega on saagise hindamine märgendamise mõttes ikka raske.

Lisaks uuriti viise kuidas lähendada kvaliteedimõõtude vahesid komponentide kaupa. Definitsioonis kirjeldatud vahe eraldati komponentideks, millest iga komponenti oli võimalik lähendada erinevate valimitega kasutades tulemusi eelnevatest peatükkidest. Tulemusena oli osa vahe hinnangust leitav täiesti märgendamata valimil. Saagise olemuse tõttu oli saagise hindamine märgendamise seisukohalt ikka kulukas.

Viidatud kirjandus

[1] W. Hoeffding. “Probability inequalities for sums of bounded random variables”. *Journal of the American Statistical Association* 58 (1963), lk. 13–30.

[2] J. Lember. *Tõenäosusteooria 2 loengukonspekt*. 2022.

[3] K. Pärna. *Tõenäosusteooria algkursus*. Tartu: Tartu Ülikooli Kirjastus, 2013.

[4] E.-M. Tiit ja M. Möls. *Rakendusstatistika algkursus*. Tartu: Tartu Ülikooli Kirjastus, 1997.

Lisad

I. Lähtekood

Töös esitatud graafikute lähtekood on kättesaadaval aadressil `https://github.com/mart-mihkel/mudelite_hindamine`.

II. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, **Mart-Mihkel Aun**,

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

   Masinõppe mudelite hindamine väheste märgenditega andmetel

   mille juhendaja on Sven Laur,

   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.

Mart-Mihkel Aun
09.05.2023