

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:

Klassifitseerimisülesandeid lahendavate masinõppe mudelite hindamiseks kasutatakse kvaliteedimõõte nagu õigsus täpsus ja saagis. Need suurused või nende hinnangud avalduvad andmepunktide tegelike klassimärgendite ja meetodi klassifikatsioonide kaudu. Tegelike klassimärgendite leidmiseks peame need manuaalselt üle vaatama. Sageli hinnatakse kvaliteedimõõte üle lõpliku valimi, leitud hinnangud sisaldavad vigu. Antud töö käigus avaldame vajaliku valimi suuruse, et mingi kindlusega ei ületaks hinnangu viga selle lubatud piiri. Lisaks peame valimi puhul õigsuse, täpsuse või saagise hindamiseks definitsiooni põhjal leidma kõikide valimi andmepunktide märgendid. Kui lisaks hinnatavale klassifitseerimismeetodile on olemas teine meetod, saame seda kasutada esimese hindamiseks. Seejuures on võimalik märgendamiseks vajalikku manuaalset tööd vähendada, kui uurime uue meetodi kvaliteedimõõdu arvutamise asemel kui palju on uus meetod vanast parem. Töös uurime tehnikaid, mis aitavad vähendada märgendamist vajavate andmepunktide 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 using the class labels of data points and the classifications made by the method on those data points. To find the actual class labels, we must manually review the datapoints. Quality measures are often evaluated using a finite sample, and the obtained estimates contain errors. In this thesis we determine the necessary sample size for the estimates not to exceed a given limit of their error with a certain confidence level. The definition-based method for determining accuracy, precision, or recall over a given sample requires all the data points' labels to be known. If we have a second method alongside the one being evaluated, we can use it to assess the first. In this case, we can reduce the amount of manual work required for labeling by estimating how much better the new method is compared to the old one. In this thesis we explore techniques which 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, accuracy, precision, recall

CERCS:P176 Artificial intelligence

Sisukord

1	Sissejuhatus	4
2	Taust	5
2.1	Lõplikud ja lõpmatud andmekogumid	5
2.2	Veahinnangud	7
2.2.1	Summa ja vahe relatiivse vea omadus	8
2.2.2	Korrutise relatiivse vea omadus	9
2.2.3	Jagatise relatiivse vea omadus	11
3	Masinõppe meetodite kvaliteedimõõtude veahinnangud	14
3.1	Õigsuse lähend	14
3.2	Täpsuse ja saagise lähendid	15
3.3	Õigsuse lähendi absoluutne viga	17
3.4	Õigsuse lähendi relatiivne viga	21
3.5	Tulemuste empiiriline testimine	23
4	Masinõppe meetodite kvaliteedimõõtude võrdlus	26
4.1	Kvaliteedimõõtude vahede lähendid	26
4.2	Õigsuste vahe lähendamine	28
4.3	Saagiste vahe lähendamine	32
4.4	Täpsuste vahe lähendamine	34
5	Kokkuvõte	35

1 Sissejuhatus

Masinõppe mudelite hindamiseks kasutatakse erinevaid viise. Klassifitseerimisülesannete puhul saab ülesannet lahendava masinõppe meetodi headust hinnata kvaliteedimõõnudega. 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. Need suurused või nende hinnangud avalduvad mingi hulga andmepunktide tegelike klassimärgendite ja meetodi poolt klassifitseeritud klassimärgendite kaudu. Tegelike klassimärgendite leidmiseks peame need manuaalselt üle vaatama, mis on suurte valimite puhul kulukas. Kvaliteedimõõnude täpsete väärtuste arvutamiseks peaksime teadma kogu populatsiooni andmepunktide märgendeid. Kuna kogu populatsiooni uurimine ei pruugi praktikas olla võimalik, leitakse sageli kvaliteedimõõnudele hinnangud üle lõpliku valimi.

Üle valimi arvutatud suuruste hinnangud ei ühti tavaliselt nende suuruste oodatud väärtusega, üle populatsiooni arvutatud suurustega. Leitud hinnangud sisaldavad valimi juhuslikkuse tõttu vigu. Seega võime küsida, kui suur on saadud viga ning kui kindlad saame olla, et viga ei ületa lubatud piiri. Selle töö käigus leiame kui suurt valimit vajame (või kui väike võib valim olla), et mingi kindlusega ei ületaks hinnangu viga selle lubatud piiri.

Valimi puhul peame õigsuse, täpsuse või saagise hindamiseks definitsiooni põhjal leidma kõikide valimi andmepunktide märgendid. Ka minimaalse võimaliku valimi puhul võib see olla probleemne. Juhul kui asendame vanat mudelit uuega, ei pea leidma uue meetodi õigsust, vaid võime uurida kas ja kui palju on uue meetodi õigsus suurem vana omast. Selles töös uurime tehnikaid, mis aitavad vähendada märgendamist vajavate andmepunktide arvu kahe klassifitseerimismeetodi kvaliteedimõõnude vahede hindamiseks.

Esimese osas toome välja töös kasutatud mõisted ja definitsioonid ning kirjeldame töö mõistmist aitavaid taustteadmisi. Töö teises osas defineerime klassifitseerimismeetodi kvaliteedimõõnude nende oodatud väärtuste kaudu, kirjeldame viise nende hindamiseks ning uurime kui suurt valimit vajame, et küllaltki suure kindlusega, saad piisavalt täpne hinnang. Kolmandas osas leiame kuidas hinnata uut klassifitseerimismeetodit võrreldes seda juba olemasolevaga ning selle juures vältida valimi 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. Tehatud 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 saame võtta kuitahes palju sõltumatuid andmepunkte. Antud töös vaatame protsesse, millel on lõplik arv väljundväärtusi, see tähendab, et jaotus on diskreetne. Sel juhul fikseerib jaotus iga konkreetse andmepunkti jaoks selle esinemistõenäosuse, millest võime mõelda kui andmepunkti oodatavat sagedusest valimis, millesse on võetud piisavalt palju andmepunkte jaotusest.

Lõpliku valimi korral saame 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 saame moodustada võttes populatsioonist andmepunkte üksteisest sõltumatult ja juhuslikult. Võib juhtuda, et samad populatsiooni andmepunktid satuvad valimisse mitmekordselt. Saadud valimi puhul eeldame, et kõik selle andmepunktid on sama jaotusega. Tegelikuses 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

kõikide andmepunktide väärtuste aritmeetiline keskmine

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

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 hälvetate ruutude aritmeetiline keskmine

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

Dispersioon näitab andmete hajuvust. Valimidispersiooni kaudu saame 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 võime 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 defineeritakse valimiüleste valemite (1) ja (2) analoogid, milleks need lõpmata suure valimi korral koonduvad.

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

$$\mathbf{E}[X] = \sum_i x_i p_i , \quad (3)$$

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

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

mitmemõõtmelise juhusliku suuruse korral on selle keskvväärtus leitav analoogiliselt, mitmemõõtmelise integraaliga.

Valimidispersiooni analoog dispersioon on juhusliku suuruse hälve ruudu keskvväärtus

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

Ruutjuurt dispersioonist nimetatakse standardhälbeks. Lõpliku populatsiooni korral avalduvad jaotuseülesed valemid (3) ja (4) neile vastavate valimiüleste valemitega (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 võime leida valimikeskmise. Antud näite korral on valimikeskmise ligikaudne hinnang keskväärtusele. Ligikaudsete väärtuste headust saame mõõta kasutades veahinnanguid. Veahinnangud kirjeldavad mingi täpse arvu ja selle arvu lähendi erinevust. Vastuse lähendamisel praktilises olukorras ei pruugi me teade täpset lahendit. See-eest sobivad veahinnangud lähendamismeetodite teoreetiliseks hindamiseks.

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

$$\Delta a = a - a_0 \text{ .}$$

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 saame kasutada hinnangut, mis arvestab arvude skaalaga. Ligikaudse arvu a relatiivseks ehk suhteliseks veaks nimetatakse suurust

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

Mõnikord esitatakse relatiivne viga protsentides.

Ligikaudsete lähendite puhul taheme nende veahinnanguid absoluutväärtuselt minimeerimine. Alati ei pruugi see olla võimalik ning peame leppima mingi veaga. Lisaks võib raske olla täielikult veenduda, et leitud lähendi viga on väiksem kui maksimaalne viga, mida võime lubada. See-eest võime uurida kui suure tõenäosusega on saavutatud viga, mis on väiksem lubatud maksimaalsest veast

$$\Pr [|a - a_0| < \varepsilon] = 1 - \alpha \text{ ,}$$

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 võime ka vastavate veahinnangute puhul eeldada, et need on normaaljaotusega.

Juhuslik suurus X on normaaljaotusega kui tema tihedusfunktsioon on

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

kus jaotuse parameetrid μ ja σ on vastavalt X keskvärtus ja standardhälve. Seda fakti tähistatakse lühidalt $X \sim \mathcal{N}(\mu, \sigma)$. Normaalkaotuse tihedus on sümmeetriline keskvaartuse μ ümber. Veahinnangute puhul on nende oodatud väärtus null, siis sümmeetrisuse tõttu on negatiivsed ja positiivsed vead sama tõenäolised. Oluline omadus normaalkaotuse puhul on see, kui suure tõenäosuse katavad standardhälbe täisarvkorreds vahemikud keskvaartuse 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 suuruste vead kanduvad arvutusi tehes tulemustele edasi. Seega uurime tulemuse täpsuse hindamiseks kuidas viga arvutustel edasi kandub. Sealjuures teeme lihtsustava eelduse, et lähendid on normaalkaotusega.

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

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

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 keskvaartused on nullid on ka summa relatiivse vea δ_+ keskvaartus null. Vigade sõltumatuse tõttu on korrutise $\varepsilon_x \cdot \varepsilon_y$ keskvaartus samuti null. Seega on summa relatiivse vea hajuvus leitav kui selle ruudu keskvaartus

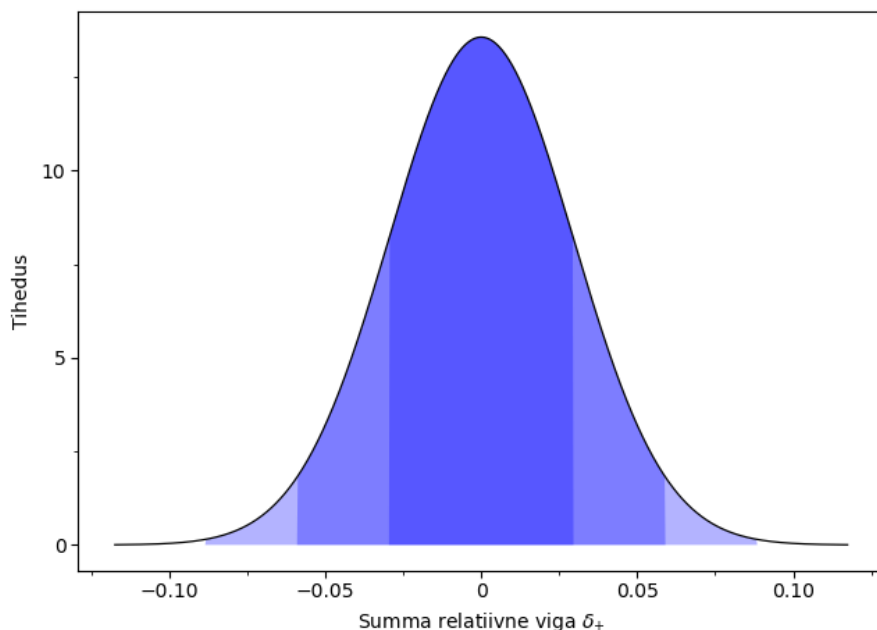
$$\mathbf{D}[\delta_+] = \mathbf{E}[\delta_+^2] = \mathbf{E}\left[\frac{\varepsilon_x^2 + 2\varepsilon_x\varepsilon_y + \varepsilon_y^2}{(x_0 + y_0)^2}\right] = \mathbf{E}\left[\frac{\varepsilon_x^2 + \varepsilon_y^2}{(x_0 + y_0)^2}\right] = \frac{\mathbf{D}[\varepsilon_x] + \mathbf{D}[\varepsilon_y]}{(x_0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

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

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



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 eeldame normaaljaotust, saame 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 antud näite korral summa $x + y$ relatiivse vea tihedus. Helesiniselt märgitud alad tähistavad standardhälbe täisarvkordseid vahemikke keskväärtuse ümber.

2.2.2 Korrutise relatiivse vea omadus

MMõnikord võib lähendatav suurus olla leitav ligikaudsete arvude korrutisena. Näiteks riskülikukujulise põranda pindala leidmiseks peame 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 :

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

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

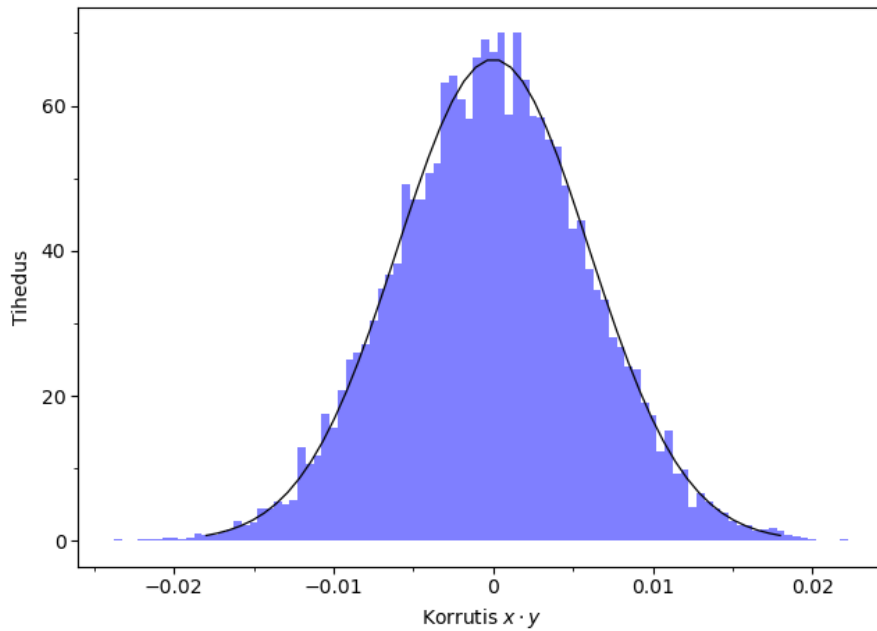
$$\begin{aligned}\delta &= \frac{x \cdot y - x_0 \cdot y_0}{x_0 \cdot y_0} = \frac{(x_0 + \varepsilon_x) \cdot (y_0 + \varepsilon_y) - x_0 \cdot y_0}{x_0 \cdot y_0} \\ &= \frac{x_0 \cdot \varepsilon_y + y_0 \cdot \varepsilon_x + \varepsilon_x \cdot \varepsilon_y}{x_0 \cdot y_0} = \frac{\varepsilon_y}{y_0}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

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

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

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

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



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

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

2.2.3 Jagatise relatiivse vea omadus

Ligikaudsete väärtuste jagatise puhul on vea edasikandumine keerulisem. Tehese mõned lihtsustavad eeldused saame leida piisavalt täpse hinnangu jagatise relatiivsele veale.

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

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

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

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

Jagatise relatiivse vea dispersiooni leidmist saame taandada korrutise relatiivse vea dispersioonile, sest $x \div y = x \cdot y^{-1}$. Seega peame hindama y pöördväärtuse relatiivse vea

disperisooni

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

Taylori arenduse põhjal saame esitada y^{-1} ligikaudselt

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

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

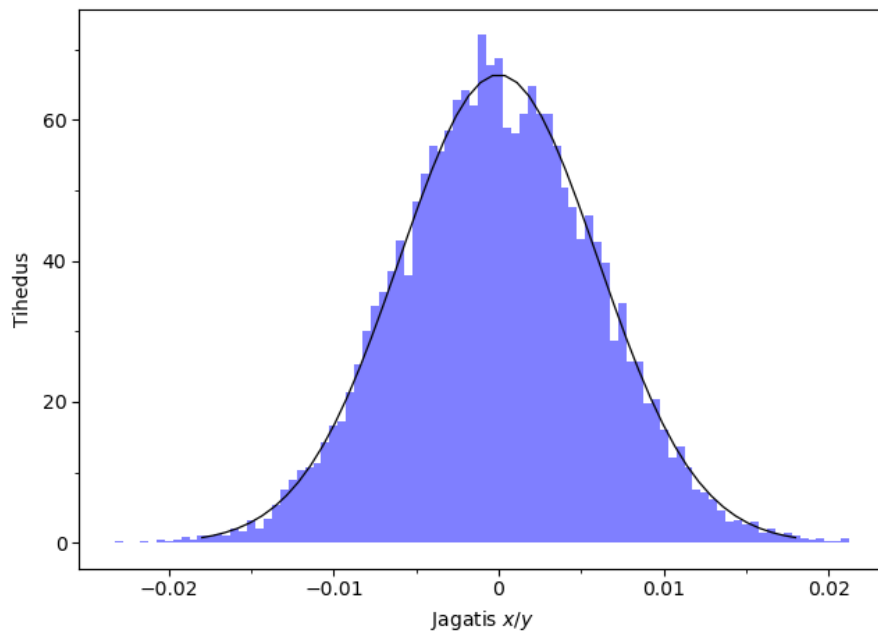
$$\begin{aligned} \mathbf{D} \left[\frac{1}{y} \right] &\approx \mathbf{D} \left[\frac{1}{y_0} - \frac{1}{y_0^2} \cdot \varepsilon_y \right] \\ &= \mathbf{E} \left[\left(\frac{1}{y_0} - \frac{1}{y_0^2} \cdot \varepsilon_y \right)^2 \right] - \mathbf{E} \left[\frac{1}{y_0} - \frac{1}{y_0^2} \cdot \varepsilon_y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

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

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

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

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



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

Joonisel 3 on histogrammina simuleeritud jagatise $x \div y$ relatiivne viga kümnetuhande juhuslikult genereeritud x ja y väärtuse puhul, kus $x \sim \mathcal{N}(0,8; 0,004)$ ning $y \sim \mathcal{N}(0,9; 0,003)$. Tume joon on relatiivse vea lähendus normaaljaotusega tulemusse (8) põhjal.

3 Masinõppe meetodite kvaliteedimõõtude veahinnangud

Kolm sageli kasutatud kvaliteedimõõtu klassifitseerimismeetodite puhul on õigsus, täpsus ja saagis. Õigsus näitab, kui sagedasti meetod andmepunkte õigesti klassifitseerib (tähistame Acc). Täpsus mõõdab päris positiivsete klassifikatsioonide sagedust ennustatud positiivsete seast (tähistame $Prec$). Saagis on päris positiivsete klassifikatsioonide sagedus positiivsete andmepunktide seast (tähistame 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 klassifitseerimise ülesannet, kus märgend y_i võib olla üks kahest võimalikust väärtusest, negatiivne klass tähistusega 0 või positiivne klass tähistusega 1. Vaatleme meetodi käitumist andmepunktidel, meetodi A puhul andmepunktile vastav klassifikatsioon on $A(x_i) = a_i$. Juhuslikku andmepunkti võime vaadelda juhusliku suurusena (X, Y) ning meetodi klassifikatsiooni selle tunnustel kui juhuslikku suurusena A . Selle põhjal defineerime meetodi A kvaliteedimõõdud läbi ooteväärtuste

$$Acc = \mathbf{E}[A = Y] \quad , \quad (9)$$

$$Prec = \mathbf{E}[A = Y | A = 1] \quad , \quad (10)$$

$$Rec = \mathbf{E}[A = Y | Y = 1] \quad . \quad (11)$$

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 on nende kõigi analüüs analoogseks. Samas on õigsuse uurimine tehniliselt kõige lihtsam, sest definitsiooni järgi ei sisalda see tinglikku jaotust. Sellel põhjusel uurime õigsust esimesena.

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

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

Lähtudes eelnevast saame meetodi õigsust (9) lähendada statistilise tõenäosusena üle N andmepunkti suuruse valimi

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

Kasutades keskväärtuse, dispersiooni ja Bernoulli jaotuse omadusi avaldame õigsuse lähendi keskväärtuse ning dispersiooni

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

Kuna suurus \widehat{Acc} on ligikaudne väärtus tegelikule õigsusele, tahame märgendada piisavalt suure valimi, 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 meetodile õigsuse $Acc = 0,9$. Lisaks hindab õigsus kõiki klassifitseerimisel tehtud vigu ühtemoodi. Kui valepositiivne või valenegatiivne klassifikatsioon võib põhjustada palju kahju on oluline tehtud vigu üksteisest eristada.

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

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

Tulemuse (13) põhjal saame hinnata täpsust (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 esitusele tõenäosuste kaudu (13) esitame ka saagise tõenäosuste kaudu

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

mille põhjal saame saagise (11) lähendi üle valimi arvutada järgnevalt

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

Sellisel kujul 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 tingimuseks kvaliteedimõõdu definitsioonis olev sündmus, näiteks täpsuse puhul $A = 1$. Tinglikule jaotusele vastava valimi leidmiseks kasutame valikumeetodit (*rejection sampling*):

1. Võtame jaotusest juhusliku andmepunkti.
2. Kontrollime andmepunkti vastavust tingimusele.
3. Kui andmepunkt vastab tingimusele võtame selle valimisse, vastasel juhul mitte.
4. Kordame kuni on leitud soovitud koguses tingimusele vastavaid andmepunkte.

Valikumeetodi põhjal leiame valimi A^+ , mis koosneb positiivseks klassifitseeritud andmepunktidest, ning valimi Y^+ , mis koosneb positiivse klassiga andmepunktidest. Kasutades vastavaid valimeid avaldame lähendid (14) ja (15) õigsusele (12) sarnasel kujul

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

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

On oluline märgata, et lähendid (16) ja (17) on analoogsed õigsuse lähendile (12) ning seelle tõttu on nende absoluutsed ja relatiivsed vead samasuguste omadustega.

Lihtsa valikumeetodi puhul on probleemiks 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 peame 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 $1 \div 10000$, näiteks tekstist faktide eraldamine. Sellisel juhul peame 1000 elemendilise valimi saamiseks märgendama 10 miljonit andmepunkti. Isesõitvate

autode puhul võivad huvipakkuvad sündmused esineda ühel korral miljonist ning seega peaksime naiivse lähenemise korral vaja märgendada ligi miljard andmepunkti. Kuid vajaliku töökindluse saavutamiseks peame selliste sündmustega 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. Kui tahame veenduda lähendi vastavuses selle täpsele väärtusele tekib küsimus vea suuruse kohta. Vea arvutamiseks peame teadma täpset väärtust, mis ei pruugi olla võimalik. Täpset väärtust teadmata saame viga hinnata tõenäosuslikult. Standardselt 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 annab 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, \dots, Z_N on sõltumatud ning leiduvad tõkked

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

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

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

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

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

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

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

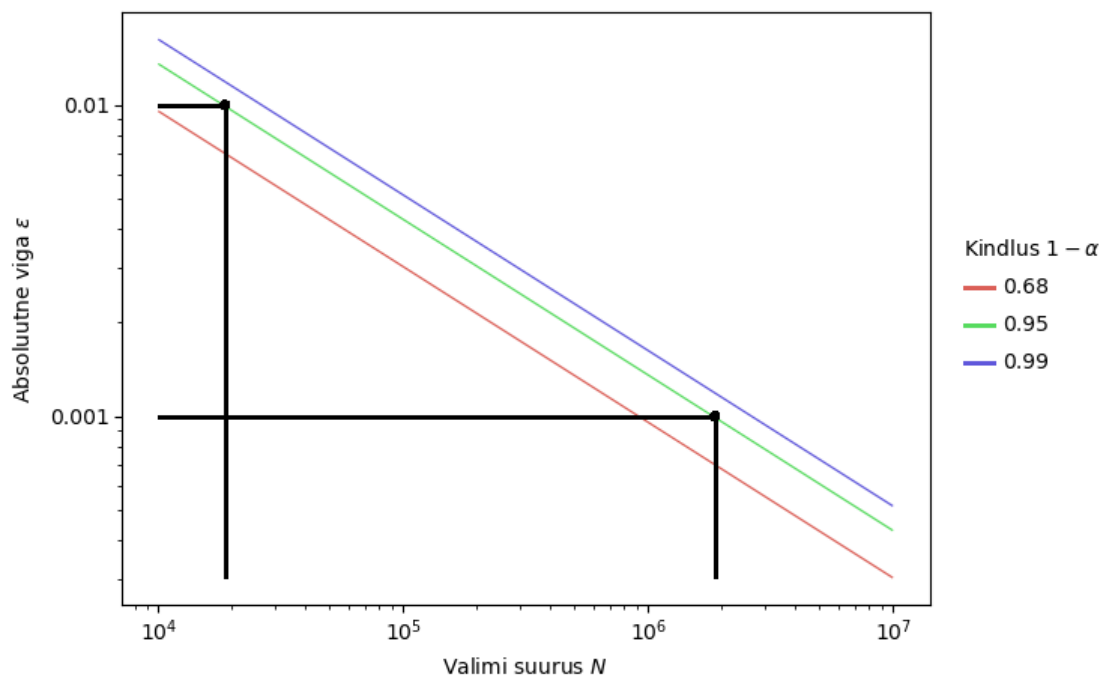
Tulemuse põhjal saame arvutada valimi vajaliku suuruse

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

et fikseeritud olulisuse korral saavutada soovitud suurusega veahinnang.

Tabel 1. Valimi suurused kindluse ning absoluutse vea suhtes Höffdinig võrratuse põhjal. Kindluse puhul on sulgudes, 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 vajalik valimi suurus soovitud kindluse ja lähendi absoluutse vea suhtes Höffdingi võrratuse põhjal, kindluse puhul on sulgudes, kui mitu normaaljaotuse standardhälvet σ keskväärtuse ümbruses see katab. Näiteks kui tahame olla 95% kindlad, et valimi põhjal arvatud õigsus \widehat{Acc} erineb algoritmi tegelikust õigsusest kuni ühe protsendi võrra, peame õigsust hindama valimil suurusega vähemalt 18445. Joonisel 4 on sama mõttekäik visuaalselt.

Õigsuse lähendi absoluutse vea hindamiseks saame 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 väidame järgnevat

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

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

Kasutades teadmist, et uuritav summa on binoomjaotusega leiame jaotuse parameetritest sõltuva hinnangu. Olulisuse α korral leiame absoluutse veahinnangu tõkke ε läheduses võrrandist

$$\Pr \left[|\widehat{Acc} - Acc| \geq \varepsilon \right] = \alpha , \quad (19)$$

valime protsendipunktid, arvud millest S_N võtab väiksemaid väärtusi protsendipunktile vastava tõenäosusega, sümmeetriliselt

$$\begin{aligned} \Pr \left[\widehat{Acc} \leq Acc - \varepsilon \right] &= \frac{\alpha}{2} , \\ \Pr \left[\widehat{Acc} > Acc + \varepsilon \right] &= 1 - \frac{\alpha}{2} , \end{aligned}$$

kui eraldame binoomjaotusega juhusliku suuruse S_N ühele poole võrratuse märki jõuame tulemuseni

$$\begin{aligned} \Pr [S_N \leq N \cdot (Acc - \varepsilon)] &= \frac{\alpha}{2} , \\ \Pr [S_N > N \cdot (Acc + \varepsilon)] &= 1 - \frac{\alpha}{2} . \end{aligned}$$

Paremale poole võrratuse märki tekkinud avaldised on S_N jaotuse vastavalt $\frac{\alpha}{2}$ ja $1 - \frac{\alpha}{2}$ protsendipunktid,

$$\begin{aligned} q_1 &= N \cdot (Acc - \varepsilon) , \\ q_2 &= N \cdot (Acc + \varepsilon) . \end{aligned}$$

Fikseeritud olulisuse ja binoomjaotuse parameetrite korral võime protsendipunktid arvutada kasutades S_N jaotuse omadusi.

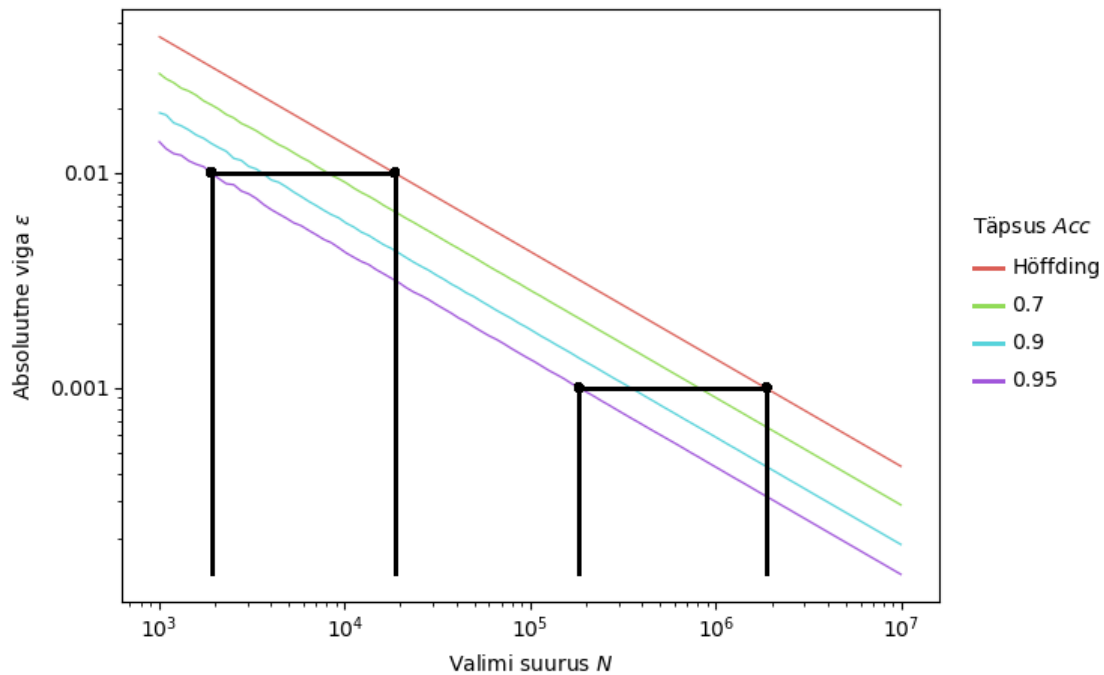
Absoluutse vea tõkkeks ε võrrandist (19) valime suurema protsentipunktide põhjal arvutatud veahinnangutest

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

Parameetri N kasvades koondub binoomjaotus normaaljaotuseks [3]. Seega suuremate N väärtuste puhul on binoomjaotus sümmeetriline, järelikult erinevad q_1 ja q_2 põhjal arvutatud veahinnangute väärtused tegelikkuses vähe. Tulemust (20) kasutades saame arvutada vajaliku valim suuruse soovitud olulisuse suhtes, kuid selle jaoks peame oletama mudeli tegelikku õigsust Acc .

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

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

3.4 Õigsuse lähendi relatiivne viga

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

järgnevalt

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

Kui avaldame lähendi relatiivse ja absoluutse vea hajuvuse suhte, leiame et kõrge õigsuse korral on õigsuse lähendi veahinnangud ligikaudu sama hajuvusega

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

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

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

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

millest eraldame tõenäosusmärgi aluses võrratuses binoomjaotusega juhusliku suuruse S_N ühele poole võrratusemärgi

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

oleme avaldanud olulisusele vastavad protsendipunktid

$$\begin{aligned}q_1 &= N \cdot Acc \cdot (1 - \varepsilon) , \\ q_2 &= N \cdot Acc \cdot (1 + \varepsilon) .\end{aligned}$$

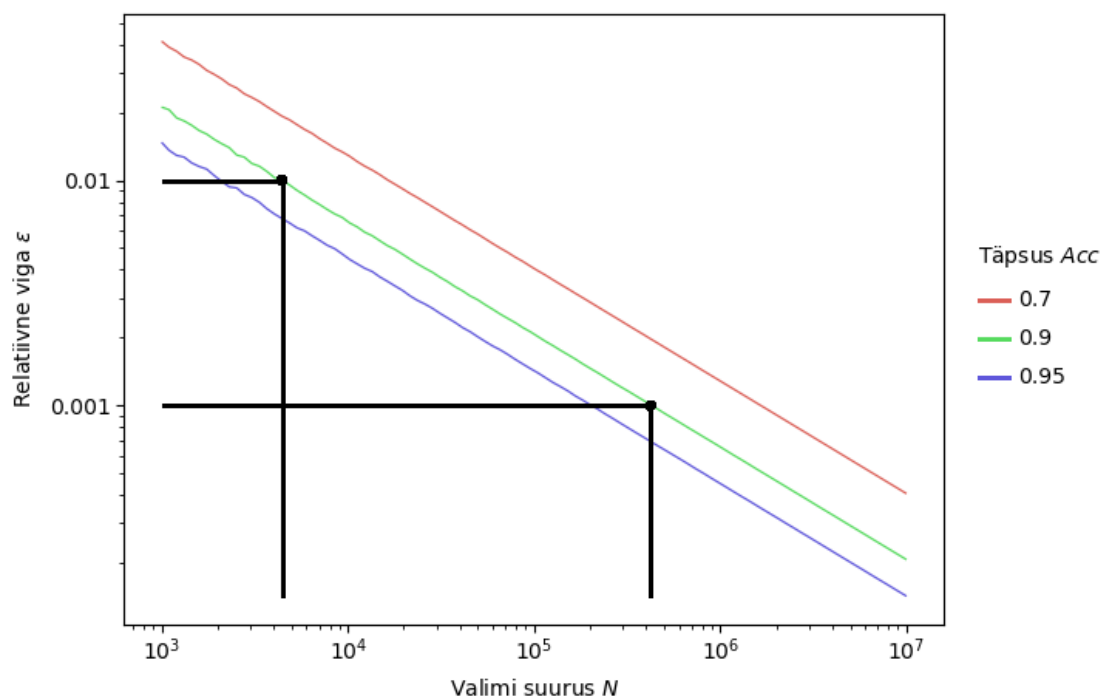
Relatiivse vea tõkkeks ε võrrandist (21) valime suurema protsendipunktide põhjal avalduva veahinnangu

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

Tulemuse põhjal saame leida 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

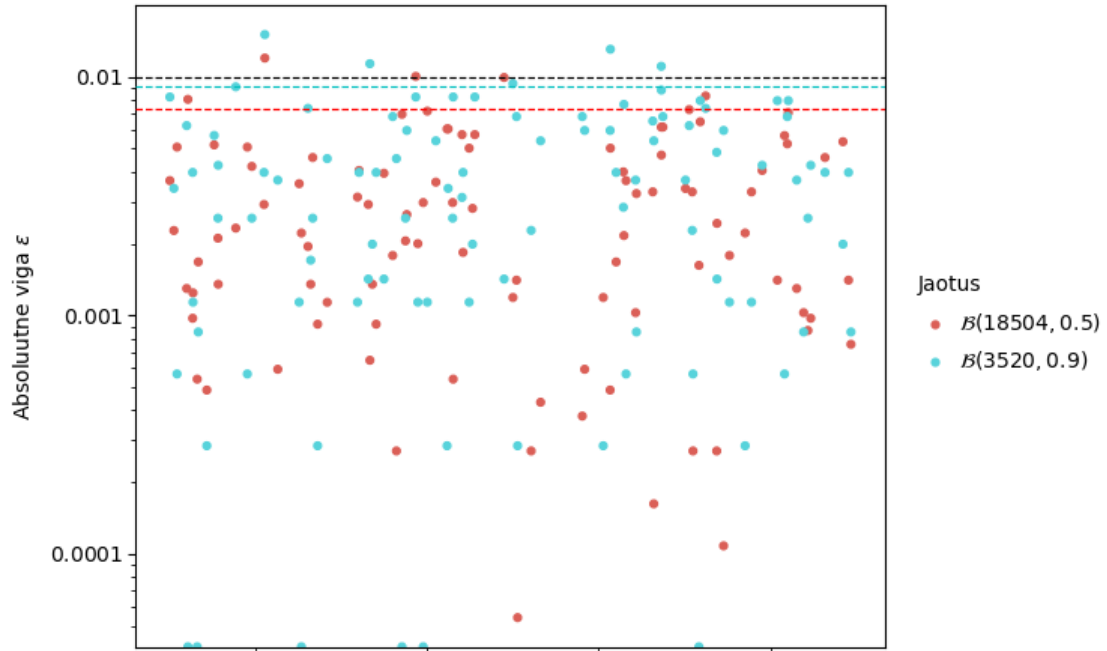


Joonis 6. Valimi suurused relatiivse vea ning eeldatud õigsuse suhtes binoomjaotuse põhjal kindlusega 95%.

Tabelis 3 ja joonisel 6 on 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 praktiliselt kontrollida. Nii saame leide lihtsasti valideerida ning lihtsasti avastada paljud arvutusvead.



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 märgivad empiiriliselt saadud 0,95 protsendipunkte, must joon nende oodatud asukohta.

Näiteks Höffdingi võrratuse põhjal kindlusega 95% kuni ühe protsendise absoluutse veahinnanguga õigsuse lähendi saavutamiseks vajame valimit suurusega vähemalt 18445. Oletades, et klassifitseerimismeetodi tegelik õigsus on 90% vajame binoomjaotuse hinnangu põhjal selleks vähemalt 3600 andmepunkti. Joonisel 7 on antud näite empiiriline test, kus summa

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

on juhuslikult genereeritud jaotusest $\mathcal{B}(18445; 0,5)$ Höffdingi võrratuse puhul (punane) ning $\mathcal{B}(3600; 0,9)$ binoomjaotuse hinnangu puhul (sinine). Punktid joonisel tähistavad genereeritud summal põhineva lähendi absoluutset viga

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

ning must joon veahinnangu absoluutväärtuse ülemist tõket, millest 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. Võime kasutada eelnevad meetodit uue hindamiseks. Vana meetodi asendamisel on eelkõige tähtis kas ja kui palju on uus meetod sellest parem. Väljaselgitamiseks uurime 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 vastavaid klassifikatsioone $A(x_i) = a_i$ ja $B(x_i) = b_i$. Vaadeldes andmepunkti juhusliku suurusena (X, Y) võime meetodite klassifikatsioone sellel vaadelda juhuslike suurusena A ja B . Meetodite kvaliteedimõõdud defineerime keskväärtuste kaudu läbi valemite (9), (10) ja (11). Näiteks meetodi B õigsus on $Acc_B = \mathbf{E}[B = Y]$. Eelnevast lähtuvalt saame meetodite erinevuse uurimisel vaadata meetodite kvaliteedimõõtude vahesid

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

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

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

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 masinõppe meetodite 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, avaldame õigsuste vahe lähendi järgnevalt

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

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

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

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

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

Analoogselt õigsusele avaldame ka valimipõhiste täpsushinnangute vahe

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

mille edasise lihtsustamise teeb raskeks erinevus murru nimetajates. Kuna tehniliselt ei pea täpsuse hindamisel kasutama sama testvalimit mõlema algoritmi jaoks, võime 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 andmepunktilise valimi moodustamisega, kus vastuvõtutingimused on vastavalt $a_i = 1$ ja $b_i = 1$. Seda arvestades saame valemi (27) viia lihtsustatud kujule

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

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

$$\begin{aligned} \widehat{\Delta Prec} &= \frac{1}{S} \cdot \left(\sum_{\substack{a_i=1 \\ b_i=1}} [y_i = 1] + \sum_{\substack{a_i=1 \\ b_i \neq 1}} [y_i = 1] - \sum_{\substack{a_i \neq 1 \\ b_i=1}} [y_i = 1] - \sum_{\substack{a_i \neq 1 \\ b_i \neq 1}} [y_i = 1] \right) \\ &= \frac{1}{S} \cdot \left(\sum_{\substack{a_i=1 \\ b_i \neq 1}} [y_i = 1] - \sum_{\substack{a_i \neq 1 \\ b_i=1}} [y_i = 1] \right) \\ &= \frac{1}{S} \cdot \sum_{a_i \neq b_i} [a_i = 1] \cdot [y_i = 1] - [b_i = 1] \cdot [y_i = 1] . \end{aligned} \quad (28)$$

Andmepunktide märgendamise mõttes on täpsuse vahe lähend (28) samaväärne õigsuse valemiga (25), sest hinnangu arvutamiseks peame märgendama vaid andmepunkte, mille puhul $a_i \neq b_i$.

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

Saagise vahe avaldame 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 (28) tuletamisega analoogset mõttekäiku, võime esitada saagise vahe hinnangu kujul

$$\Delta \widehat{Rec} = \frac{1}{T} \cdot \sum_{a_i \neq b_i} [a_i = 1] \cdot [y_i = 1] - [b_i = 1] \cdot [y_i = 1] . \quad (29)$$

Märgendamise seisukohalt on T arvutamine kulukas, kuna iga summa liikme puhul on vajalik teada andmepunkti tegelikku klassi. Kõigi N andmepunkti manuaalne märgendamine on väga ressursimahukas. Alternatiivina saame väiksema valimi põhjal positiivse klassi esinemise sagedust ennustada. See on võimalik vaid siis kui positiivse klassi esinemise sagedus on piisavalt suur. Kui positiivse klassi esindajad on sagedusega 1 : 1000, peaksime adekvaatse hinnangu saamiseks läbi vaatama üle kümnetuhande andmepunkti. Seega on valemi (29) praktiline rakendamine raskendatud veelgi madalama esinemissagedusega sündmuste korral.

4.2 Õigsuste vahe lähendamine

Siiani oleme lähendeid kvaliteedimõõtude vahele avaldatud vahetult läbi vahe oodatud väärtuse definitsiooni. Tehes teisendusi vahe definitsioonis antud avaldises võime selle eraldada mitmeks komponendiks. Lõpliku lähendi leidmiseks peame lähendama iga komponenti eraldi ning nende põhjal arvutama vahele hinnangu.

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

$$\Delta Acc = \mathbf{E}[A = Y] - \mathbf{E}[B = Y] = \mathbf{E}[[A = Y] - [B = Y]] \quad . \quad (30)$$

Tähistagu Z valemis (30) viimase keskväärtusmärgi all olevat juhuslikkust suurust, $Z = [A = Y] - [B = Y]$. Analoogselt võime valemis (25) oleva summa liikmeid tõlgendada juhuslike suurustena $Z_i = [a_i = y_i] - [b_i = y_i]$. Kui arvestame, et juhuslikud suurused Z_i on sõltumatud ning sama jaotusega kui Z , näeme et lähendi (25) keskväärtus on

$$\mathbf{E}[\widehat{\Delta Acc}] = \frac{1}{N} \cdot \sum_{i=1}^N \cdot \mathbf{E}[Z_i] = \Delta Acc \quad . \quad (31)$$

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

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

Edasiste arvutuste selgemaks esitamiseks võtame kasutusele tähistused

$$\beta = \Pr[Z \neq 0] \quad , \quad (32)$$

$$\gamma = \mathbf{E}[Z | Z \neq 0] \quad , \quad (33)$$

$$\kappa = \Pr[Z = 1 | Z \neq 0] \quad . \quad (34)$$

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

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

ning me saame esitada õigsuste vahe antud suuruste kaudu

$$\Delta Acc = \beta \cdot \gamma = \beta \cdot (2\kappa - 1) \quad . \quad (35)$$

Võrdusest (35) lähtuvalt hindame meetodite õigsuste vahet lähendades suurusi β ja γ . Kuna $\beta = \Pr[Z \neq 0]$ on tõenäosus, saame seda hinnata statistilise tõenäosusena üle N elemendilise valimi

$$\hat{\beta} = \frac{1}{N} \cdot \sum_{i=1}^N [Z_i \neq 0] = \frac{1}{N} \cdot \sum_{i=1}^N [a_i \neq b_i] \quad , \quad (36)$$

[4]. Summa märgi alune juhuslik suurus võtab väärtusi 0 ja 1. See tähendab, et lähendi absoluutse ja relatiivse vea hinnangud saame leida eelmises peatükis kirjeldatud meetodeid kasutades.

Lähendi tinglikule keskväärtusele $\gamma = \mathbf{E}[Z | Z \neq 0]$ on leiame sarnaselt, valimi-keskimisena kus iga valimi andmepunkti korral $a_i \neq b_i$. Olgu K sellise valimi suurus, esitame lähendi kujul

$$\hat{\gamma} = \frac{1}{K} \cdot \sum_{i=1}^K [a_i = y_i] - [b_i = y_i] . \quad (37)$$

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 defineerime uue Bernoulli jaotusega juhusliku suuruse W , mis võtab väärtuse 0 kui summeritav suurus võtab -1 ning muidu 1

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

mille puhul

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

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

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

ning kordama sama mõttekäiku, mis valemi (21) puhul.

Leitud lähendite korrutisest leiame omakorda lähendi meetodite õigsuste vahele

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

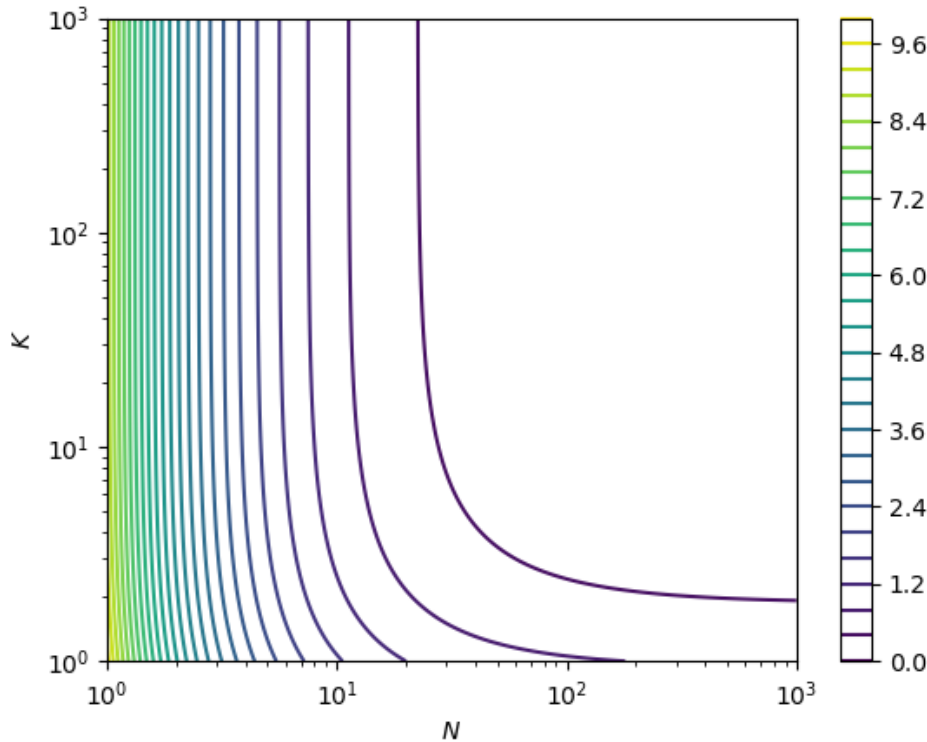
Vahe lähendi relatiivse vea dispersioon avaldame relatiivse vea korrutise omaduse (5) põhjal

$$\mathbf{D} \left[\frac{\hat{\beta} \cdot \hat{\gamma}}{\beta \cdot \gamma} \right] \approx \mathbf{D} \left[\frac{\hat{\beta}}{\beta} \right] + \mathbf{D} \left[\frac{\hat{\gamma}}{\gamma} \right] , \quad (38)$$

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

$$\mathbf{D} \left[\frac{\hat{\beta}}{\beta} \right] = \frac{1}{N} \cdot \frac{1 - \beta}{\beta} , \quad (39)$$

$$\mathbf{D} \left[\frac{\hat{\gamma}}{\gamma} \right] = \frac{1}{K} \cdot (\mathbf{E}[Z^2 | Z \neq 0] - \mathbf{E}[Z | Z \neq 0]^2) = \frac{1}{K} \cdot (1 - \gamma^2) . \quad (40)$$



Joonis 8. Õigsuste vahe lähendi relativse 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.

Joonise 8 ning vahetulemuste (39) ja (40) kaudu näeme, et $\widehat{\Delta Acc}$ relativse vea dispersioon kahaneb valimisuuruste kasvades. Telgedel olevatest suurustest tähistab N valimi suurust, mida ei pea märgendama. Kuna sellise valimi leidmine ei ole keeruline võime 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 uurime olukorda, kus saavutame võrdsed lähendite $\hat{\beta}$ ja $\hat{\gamma}$ dispersioonid

$$\frac{1}{N} \cdot \frac{1 - \beta}{\beta} = \frac{1}{K} \cdot (1 - \gamma^2) . \quad (41)$$

Kuna $\Delta Acc = \beta \cdot \gamma$, saame valemi (41) esitada kujul

$$K = N \cdot \frac{\beta}{1 - \beta} \cdot \left(1 - \frac{(\Delta Acc)^2}{\beta^2} \right) , \quad (42)$$

mille põhjal saame hinnata märgendamist vajava valimi suurust.

4.3 Saagiste vahe lähendamine

Saagise puhul on hinnangu (29) analüüsimine keerulisem, kuna saagise definitsioonis on tinglik keskvärtus. Valemile (35) analoogi tuletamiseks peame kõigepealt eraldama saagise definitsioonist tingliku jaotuse

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

millest saame tuletada

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

Analoogselt meetodi B puhul

$$\mathbf{E}[B = Y|Y = 1] = \frac{\Pr[B = 1 \wedge Y = 1]}{\Pr[Y = 1]} = \frac{\mathbf{E}[[B = 1] \cdot [Y = 1]]}{\mathbf{E}[Y = 1]} .$$

Nende kahe võrduse põhjal esitame saagiste vahe kujul

$$\begin{aligned}\Delta_{Rec} &= \mathbf{E}[A = Y|Y = 1] - \mathbf{E}[B = Y|Y = 1] \\ &= \frac{\mathbf{E}[[A = 1] \cdot [Y = 1] - [B = 1] \cdot [Y = 1]]}{\mathbf{E}[Y = 1]} .\end{aligned}$$

Sellest saame järeldada algse saagise lähendi valemi (15) põhjal, et ka saagiste vahe lähend (29) on nihketa. Kuid sellisel kujul esitatud vahele Δ_{Rec} ei ole lähendi leidmine märgendamise seisukohalt veel kõige parem.

Kehtivad seosed

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

nende seoste ja keskvärtuse definitsiooni põhjal esitame saagise vahe lugeja korruutisena tõenäosusest ja tinglikust keskvärtusest

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

Edasiste arvutuste selgemaks esitamiseks võtame jällegi kasutusele abistavad tähistused

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

mille kaudu saame esitada saagiste vahe

$$\Delta Rec = \frac{\beta \cdot \eta}{\nu} . \quad (43)$$

Tulemuse (43) põhjal on võimalik meetodite saagiste vahet hinnata kui lähendame suursi β , ν ja η . Tõenäosuse β puhul saame hinnang leida samamoodi nagu (36) eelmises alampeatükis. Ning ν lähendi 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 lähendame 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 saame endiselt kasutada Höffdinig võrratust lähendi absoluutse vea hindamiseks, 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 on absoluutse vea absoluutväärtuse maksimaalset lubatud suurust ning α olulisus.

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

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

millest

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

4.4 Täpsuste vahe lähendamine

Täpsuse jaoks on võrrandi (43) analoogi tuletamine keeruline, sest täpsuste vahe lähendis on summa liikmed mittesõltumatud. Üheks võimalikuks lahenduseks täpsuse hindamisel on kahesammuline lähenemine, kus hindame esmalt lähendi (27) dispersiooni funktsioonina valikumeetodi abil saadud S elemendilisest valimist (kahe keskmise vahe) ning seejärel lähendame selle optimeeritud esitust (28) veelgi väiksema K -elemendilise juhuvalimiga. Kuna saadud lahendus ei ole elegantne ning kõik sammud iseseisvalt on analoogsed peatükis 3 saadud tulemustega, siis ei me seda antud töös pikemalt ei käsitle.

5 Kokkuvõte

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

Uurisime ka tehnikaid kahe masinõppe meetodi võrdlemiseks kvaliteedimõõdude põhjal. Uue mudeli kasutusele võtmisel oli küsimuseks kui palju see vanast parem on. Vastust küsimusele otsisime meetodite kvaliteedimõõdude vahede kaudu. Õigsuse ja täpsuse puhul oli võimalik vahele hinnangut leida märgendades manuaalselt vaid andepunkte, mille puhul meetodite klassifikatsioonid erinesid. Selle tulemusel oli võimalik vähendada märgendamist vajavate andmepunktide arvu. Saagise puhul oli osa lähendist avaldatav sarnaselt õigsusele ja täpsusele. Kuid saagise arvutamiseks on vajalik ikkagi teada kõikide andmepunktide tegelikke klassimärgendeid, seega on saagise hindamine märgendamise mõttes ikka raske.

Lisaks uurisime viise kuidas lähendada kvaliteedimõõdude vahesid komponentide kaupa. Definitsioonis kirjeldatud vahe eraldasime 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 Commons'i 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