TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
MATEMAATILISE STATISTIKA INSTITUUT

Reigo Hendrikson

Nominaalsete sisendtunnuste vaheliste seoste kasutamine lähinaabrite meetodi korral

Bakalaureusetöö

Juhendaja: prof Kalev Pärna

Tartu 2013

Sisukord

Sissejuhatus

Lähinaabrite meetod on mitteparameetrilise regressiooni tehnika, mis kasutab uuritava tunnuse hindamiseks mingil objektil vaid neid treeningandmestikku kuuluvad objekte, mis on lähedal uuritavale objektile. Meetod tugineb eeldusele, et uuritava tunnuse väärtus sarnaneb väärtustega, mis vastavad pigem objektile lähemal kui kaugemal paiknevatele objektidele.

Käesolev bakalaureusetöö keskendub lähinaabrite meetodi rakendamisele nominaalsetel tunnustel. Nominaaltunnuste korral traditsiooniliselt kasutatav kaugus on nn Hamming'i (0-1)-kaugus, mis on aga liiga kohmakas analüüsivahend. Samuti ei tundu olevat õige kasutada objektidevahelise kauguse mõõduna üksiktunnuste järgi võetud erinevuste summat, kuna see ei võta arvesse nominaaltunnuste vahelisi seoseid. Nominaaltunnuste vaheliste seoste arvesse võtmiseks pakume välja moodustada kõigist nominaaltunnustest liittunnus, mille väärtusteks on lähtetunnuste väärtuste kombinatsioonid. Liittunnuse probleemiks on aga väärtuskombinatsioonide paljusus ja erinevaid tunnuste kombinatsioone esindavate vaatluste vähesus. Selle probleemi lahendamiseks grupeerime tunnuste kombinatsioonid sarnasuse põhjal. Selleks defineerime uue kaugusfunktsiooni, mis erineb traditsioonilisest väärtustega 0 ja 1 kaugusfunktsioonist. Seega pakume antud bakalaureusetöös välja ühe võimaliku viisi nominaalsete tunnuste vaheliste seoste arvestamiseks lähinaabrite meetodi korral.

Bakalaureusetöö on jagatud viieks osaks. Esimeses peatükis anname lühikese ülevaate lähinaabrite meetodist ja defineerime uue kaugusfunktsiooni. Teises osas kirjeldame töös kasutatavat meetodit uuritava tunnuse väärtuste prognoosimiseks. Töö kolmandas osas kirjeldame k-keskmise meetodit ja Lloyd'i iteratiivset algoritmi. Neljandas osas keskendume optimaalse klasside arvu määramisele ning viiendas osas rakendame teises peatükis kirjeldatud mee-

todit reaalsetel andmetel.

Töös esitatud joonised ja andmed nende moodustamiseks on saadud programmi R abil. Töö on kirjutatud tekstitöötlusprogrammis MiKTeX.

Autor tänab professor Kalev Pärnat, kes juhtis tähelepanu olulistele uurimuspunktidele ja andis nõu nende käsitlemises ning aitas tööd formuleerida.

1 Lähinaabrite meetod

Lähinaabrite meetod on mitteparameetrilise regressiooni tehnika, mis kasutab väljundi y hindamiseks mingi sisendi $\boldsymbol{x}$ korral vaid sellele sisendile lähedal olevaid vaatlusi. Meetod tugineb eeldusele, et väljundi y väärtus uuritava sisendi $\boldsymbol{x}$ korral sarnaneb y väärtustega, mis vastavad pigem $\boldsymbol{x}$-le lähemal kui kaugemal paiknevatele vaatlustele.

1.1 k-lähinaabri hinnang

Olgu meil antud n vaatlust (x_i, y_i), kus iga $i = 1, \ldots, n$ korral $x_i = (x_{i1}, \ldots, x_{il})$ on sisend ja y_i on sisendile x_i vastav väljund[1]. Edaspidi nimetame n vaatlust (x_i, y_i) treeningandmestikuks[2]. Meie soov on hinnata uuele sisendi väärtusele $\boldsymbol{x}$ vastava väljundi y väärtust. Tähistame selle hinnangu $\hat{y} = \hat{y}(\boldsymbol{x})$. Lähinaabrite meetod kasutab $\hat{y}$ määramiseks neid treeningandmestikus olevaid vaatlusi (x_i, y_i), mis on teatud mõttes lähimad $\boldsymbol{x}$-le. Täpsemalt, k-lähima naabri korral on $\hat{y}$ defineeritud järgnevalt ([1], lk 14):

$$\hat{y}(\boldsymbol{x}) = \frac{1}{k} \sum_{x_i \in N_k(\boldsymbol{x})} y_i, \tag{1}$$

kus $N_k(\boldsymbol{x})$ on sisendi $\boldsymbol{x}$ ümbrus (naabruskond), kuhu kuuluvad $\boldsymbol{x}$-le k lähimat treeningandmestikus olevat sisendit x_i. Ümbruse $N_k(\boldsymbol{x})$ määramiseks on meil vaja defineerida kahe sisendi vaheline kaugus.

1.2 Kahe sisendi vaheline kaugus

Vaatleme treeningandmestiku sisendeid $x_1, x_2, \ldots, x_n$, kus $x_i = (x_{i1}, \ldots, x_{il})$. Suurusega x_{ij} on märgitud i-nda sisendi j-nda tunnuse väärtus.

[1] Me järgime tehisõppe-alast terminoloogiat, kus argumenttunnuseid nimetatakse *sisendiks* ja funktsioontunnust nimetatakse *väljundiks*.

[2] Nimetus tuleneb sellest, et siin on sisendile vastavad väljundid teada.

Mõõdetud tunnused võivad olla nii arvulised (pidevad, diskreetsed) kui ka mittearvulised (nominaalsed, järjestustunnused). Kauguse kahe sisendi x_i ja $x_{i'}$ vahel defineerime seosega:

$$d(x_i, x_{i'}) = \sum_{j=1}^{l} d_j(x_{ij}, x_{i'j}), \tag{2}$$

kus $d_j(x_{ij}, x_{i'j})$ on sisendite x_i ja $x_{i'}$ vaheline kaugus (erinevus) tunnuse j lõikes.

Olgu a ja b tunnuse j kaks mingit väärtust. Kui j on arvuline tunnus (k.a arvuliseks kodeeritud järjestustunnus), siis saab kaugusfunktsioonina kasutada näiteks

$$d_j(a, b) = (a - b)^2.$$

Kui j on nominaaltunnus, sobib kaugusfunktsiooniks näiteks

$$d_j(a, b) = \begin{cases} 0, \text{ kui } a = b \\ 1, \text{ kui } a \neq b, \end{cases}$$

mille alternatiivne kuju on $d_j(a, b) = 1_{a \neq b}$.

Seosega (2) defineeritud kahe sisendi x_i ja $x_{i'}$ vaheline kaugus avaldub seega sisendeid iseloomustavate üksikute tunnuste vaheliste kauguste kogusummana. Kui tunnused on mõõdetud erinevatel skaaladel, tekib olukord, kus sisendite vaheline kaugus on suuresti mõjutatud neist tunnustest, mis on mõõdetud laiemal skaalal ning seega ei ole saadav tulemus adekvaatne. Antud probleemi lahendab kauguste normeerimine.

Lähinaabrite meetodi teiste probleemidega ja nende võimalike lahendustega tutvumiseks soovitame töid [2, 3].

1.3 Nominaaltunnuste probleem

Edaspidi vaatleme olukorda, kus kõik sisendil $x_i = (x_{i1}, \ldots, x_{il})$ mõõdetud l tunnust on nominaalsed ning vastav väljund y_i on kvantitatiivne. Töös esi-

nevates näidetes kasutame tabelis 1 olevat fiktiivset andmestikku.

Tabel 1: Fiktiivne andmestik.

Jrk	Sugu	Mark	Kahju
1	N	Opel	500
2	M	BMW	700
3	M	BMW	900
4	N	Volvo	650
5	N	Volvo	750
6	N	Volvo	700

Märkus. Tunnus *mark* näitab auto marki. *Sugu* ja *mark* on vaadeldavad kui sisendtunnused ja tunnus *kahju* kui väljund. Tunnus *kahju* näitab kindlustuskahju summat.

Selguse mõttes eristame edaspidi tunnuse väärtust ja tunnuse taset.

Definitsioon 1.1 *Tunnuse iga unikaalset väärtust nimetatakse selle tunnuse tasemeks.*

Näide 1.1 Tabelis 1 on tunnuse *sugu* väärtusteks: N, M, M, N, N, N. Samas tunnuse *sugu* unikaalsed väärtused ehk tasemed on N ja M. Märgime, et tunnuse kõik tasemed ei pruugi andmestikus esineda.

Järgnevas arutelus käsitleme l nominaalset tunnust ühe *liittunnusena*, mille tasemeteks on lähtetunnuste tasemete kombinatsioonid.

Näide 1.2 Tabelis 1 on kaks nominaalset tunnust (l=2): *sugu* ja *mark*. Moodustame neist liittunnuse *sugu-mark*. Et tunnuse *sugu* tasemed on N ja M ning tunnuse *mark* treeningandmestikus esinevad tasemed on *Opel, BMW, Volvo*, siis liittunnuse *sugu-mark* tasemed on *N-Opel, M-BMW* ja *N-Volvo*.

Seega võime üldistust kitsendamata eeldada, et meil on tegemist ühe nominaalse tunnusega $\boldsymbol{T}$, mis on oma sisult l nominaalsest tunnusest moodustatud

liittunnus.

Olgu nominaalse tunnuse $\boldsymbol{T}$ taseme arv m ning treeningandmestiku suurus n. Eeldusel, et treeningandmestiku vaatlused (x_i, y_i), kus sisendist $x_i = (x_{i1}, \ldots, x_{il})$ on moodustatud tunnuse $\boldsymbol{T}$ i-s väärtus (sisend), jagunevad võrdselt m taseme vahel, saame keskmiselt $\frac{n}{m}$ vaatlust igale tunnuse $\boldsymbol{T}$ tasemele. Kui nüüd m on liiga suur või n liiga väike, tekib olukord, kus tunnuse $\boldsymbol{T}$ üksikute tasemete kohta on väga vähe vaatlusi.

Meie soov on hinnata valemiga (1) uuele sisendi väärtusele $\boldsymbol{x}$ vastavat väljundi y väärtust. Selleks peame leidma sisendi $\boldsymbol{x}$ ümbruse $N_k(\boldsymbol{x})$. Kuna tegemist on nominaalse tunnusega, on esimene mõte kasutada kahe sisendi x_i ja $x_{i'}$ vahelise kauguse (erinevuse) määramiseks kaugusfunktsiooni $d(x_i, x_{i'}) = 1_{x_i \neq x_{i'}}$, kus x_i ja $x_{i'}$ on tunnuse $\boldsymbol{T}$ kaks võimalikku väärtust. Sel juhul on $N_k(\boldsymbol{x})$ jaoks ainult kaks võimaluset: $N_k(\boldsymbol{x})$ sisaldab ainult neid sisendeid x_i, mis ühtivad sisendiga $\boldsymbol{x}$, $N_k(\boldsymbol{x}) = \{x_i : x_i = \boldsymbol{x}\}$ või $N_k(\boldsymbol{x})$ sisaldab kogu treeningandmestikku. Esimesel juhul on uue sisendi $\boldsymbol{x}$ naabruskonda $N_k(\boldsymbol{x})$ kuuluvate vaatluste arv tüüpiliselt väga väike ning kuna väärtuse $\hat{y}(\boldsymbol{x})$ leidmiseks kasutame ainult naabruskonda $N_k(\boldsymbol{x})$ kuuluvaid vaatlusi, siis ei ole leitav prognoos usaldusväärne. Teisel juhul aga prognoos $\hat{y}(\boldsymbol{x})$ on ühesugune kõikide sisendite $\boldsymbol{x}$ korral ning on seetõttu väheefektiivne.

Järgmises punktis pakume välja alternatiivse viisi nominaaltunnuse tasemete vaheliste kauguste määramiseks.

1.4 Nominaaltunnuse tasemete vahelise kauguse uus mõõt

Olgu meil n vaatlust (x_i, y_i), kus x_i tähistab nominaaltunnuse $\boldsymbol{T}$ sisendi väärtust i-ndal vaatlusel ja sisendile x_i vastavat väljundi väärtust tähistab y_i. Olgu tunnuse $\boldsymbol{T}$ tasemed $z_1, z_2, \ldots, z_m$. Siiani oleme kasutanud nominaaltunnuse tasemete vahelise kauguse määramiseks kaugusfunktsiooni $d(a, b) = 1_{a \neq b}$,

kus a ja b on tunnuse $\boldsymbol{T}$ kaks võimalikku väärtust.

Käesolevas töös pakume välja uue meetodi nominaaltunnuse tasemete vahelise kauguse määramiseks. Meetod seisneb selles, et igale nominaaltunnuse tasemele z_j omistatakse uus arvuline väärtus z'_j, mis leitakse väljundtunnuse y tingliku keskmisena:

- Leiame vaatluste (x_i,y_i) hulgast kõik vaatlused, mille korral $x_i = z_j$. Tähistagu n_j leitud vaatluste arvu.
- Leitud vaatluste y keskmisest väärtusest saab taseme z_j uus arvuline väärtus
$$z'_j = \frac{1}{n_j} \sum_{x_i = z_j} y_i. \tag{3}$$
- Kui tunnuse $\boldsymbol{T}$ tase z_j ei esine treeningandmestikus kordagi ehk $n_j = 0$, siis tema arvuliseks väärtuseks z'_j saab tunnuse y keskmine väärtus üle kõigi treeningandmestiku vaatluste:
$$z'_j = \frac{1}{n} \sum_{i=1}^{n} y_i. \tag{4}$$

Sellisel moel oleme andnud nominaaltunnuse $\boldsymbol{T}$ igale tasemele z_j arvulise väärtuse z'_j ehk teisiti öeldes, oleme muutnud nominaaltunnuse arvuliseks tunnuseks. Tunnuse $\boldsymbol{T}$ tasemete vaheliste kauguste määramiseks saame nüüd kasutada kaugusfunktsiooni $d(a, b) = (a-b)^2$. Sellega oleme tekitanud olukorra, kus nominaaltunnuse taseme naabruskond ei ole rangelt piiratud ainult sama taseme vaatlustega, vaid sõltub tasemetele omistatud arvulistest väärtustest z'_j.

Näide 1.3 Vaatame tabelis 1 olevat andmestikku. Tunnuse $\boldsymbol{T}$ rollis on liittunnus *sugu-mark*. Tunnuse y rollis on *kahju*. Kohandatud andmestik on toodud tabelis 2.

Tabel 2: Kohandatud andmestik.

Taseme z_j nr	Sugu-mark	Kahju
1	N-Opel	500
2	M-BMW	700
2	M-BMW	900
3	N-Volvo	650
3	N-Volvo	750
3	N-Volvo	700

Märkus. Tunnustest *sugu* ja *mark* on moodustatud liittunnus *sugu-mark*.

Tunnuse *sugu-mark* andmestikus esinevad tasemed on *N-Opel* (tase 1), *M-BMW* (tase 2) ja *N-Volvo* (tase 3). Leiame neile tasemetele uue arvulise väärtuse, milleks on *keskmine kahju* samale tasemele kuuluvate vaatluste seas. Lisaks leiame uue arvulise väärtuse tasemele *M-Audi*. Tabelis 3 on toodud tunnuse *sugu-mark* tase ja taseme uus arvuline väärtus - *keskmine kahju*. Taseme uus arvuline väärtus kolme esimese taseme jaoks on arvutatud valemiga (3), neljanda taseme ehk *M-Audi* jaoks valemiga (4).

Tabel 3: Uus arvuline väärtus.

Taseme z_j nr	Tase z_j	Taseme keskmine kahju z'_j
1	N-Opel	$500 : 1 = 500$
2	M-BMW	$(700 + 900) : 2 = 800$
3	N-Volvo	$(650 + 750 + 700) : 3 = 700$
4	M-Audi	$(500 + 700 + 900 + 650 + 750 + 700) : 6 = 700$

Märkus. Taseme *keskmine kahju* z'_j näitab tasemele z_j omistatud uut arvulist väärtust.

Järgnevas peatükis vaatleme nominaaltunnuse $\boldsymbol{T}$ arvuliste väärtuste z'_j ka-

sutamist väljundtunnuse y prognoosimiseks. Peamine idee seisneb selles, et me ühendame lähedaste z'_j väärtustega tasemed suurematekss gruppideks, mis on piisavalt suured selleks, et nende baasil saadud väljundi y hinnang $\hat{y}$ on rahuldava täpsusega.

2 Uue väljundi prognoosimine

Olgu meil taaskord n vaatlust (x_i, y_i), kus x_i tähistab nominaaltunnuse $\boldsymbol{T}$ sisendi väärtust i-ndal vaatlusel ja sisendile x_i vastavat väljundi väärtust tähistab y_i. Olgu tunnuse $\boldsymbol{T}$ tasemed $z_1, z_2, \ldots, z_m$. Meie soov on hinnata uuele sisendi väärtusele $\boldsymbol{x}$ vastavat väljundi y väärtust. Tähistame selle hinnangu $\hat{y} = \hat{y}(\boldsymbol{x})$. Käesolevas töös läbime väärtuse $\hat{y}(\boldsymbol{x})$ leidmiseks 6 etappi, mida järgnevalt kirjeldame.

2.1 Nominaaltunnuse tasemetele arvuliste väärtuste omistamine

Omistame nominaaltunnuse $\boldsymbol{T}$ igale tasemele z_j arvulise väärtuse z'_j, kasutades valemeid (3) ja (4). Lähemalt on arvulise väärtuse omistamist kirjeldatud punktis 1.4.

Enne teise etapi juurde minemist tutvume mõistega *klass*.

Definitsioon 2.1 *Klassiks nimetatakse hulka, kuhu kuulub üks või rohkem nominaaltunnuse taset.*

Näide 2.1 Vaatame tabelis 2 olevat andmestikku. Tunnuse *sugu-mark* andmestikus esindatud tasemd on *N-Opel*, *M-BMW* ja *N-Volvo*. Nendest tasemetest on võimalik moodustada üks, kaks ja kolm klassi (vt tabel 4, 5, 6).

Märgime, et tasemeid võib kahte klassi jaotada ka muul viisil.

Tabel 4: Üks klass.

Klass	Klassi liikmed
1	N-Opel
	M-BMW
	N-Volvo

Tabel 5: Kaks klassi.

Klass	Klassi liikmed
1	N-Opel
2	M-BMW
	N-Volvo

Tabel 6: Kolm klassi.

Klass	Klassi liikmed
1	N-Opel
2	M-BMW
3	N-Volvo

2.2 Klasside moodustamine

Moodustame tunnuse $\boldsymbol{T}$ tasemetest $z_1, z_2, \ldots, z_m$ klassid. Olgu klasside arv p ja tähistagu $K_1, K_2, \ldots, K_p$, kus $p \in \{1, \ldots, m\}$, vastavaid klasse. Klasside moodustamisel lähtume ideest, et ühte klassi peaksid kuuluma sarnased tasemed. Antud töös määrab tasemete sarnasuse vastavate arvuliste väärtuste z'_j lähedus: mida väiksem on suurus $|(z'_{j1} - z'_{j2})|$, seda sarnasemaks peame tasemeid z_{j1} ja z_{j2}.

Vaatleme arvväärtuste järkstatistikuid $z'_{(1)} \leq z'_{(2)} \leq \cdots \leq z'_{(m)}$, kus $z'_{(1)}$ tähistab väikseimat tunnuse $\boldsymbol{T}$ tasemele omistatud arvväärtust ja $z'_{(m)}$ tähistab suurimat tunnuse $\boldsymbol{T}$ tasemele omistatud arvväärtust.

Tasemetest klasside moodustamisel lähtume sellest, et kui $i < j < k$ ja $z_{(i)}, z_{(k)} \in K_h$, $h \in \{1, , \ldots, p\}$ siis ka $z_{(j)} \in K_h$. Sisuliselt tähendab see reaalteljel $p-1$ tükelduspunkti ehk klassipiiri leidmist ning kahe järjestikuse tükelduspunkti vahele jäävatele arvväärtustele vastavate tasemete ühendamist ühte klassi.

Näide 2.2 Vaatame tabelis 3 olevat esimest kolme taset (veerg tase z_j) ja neile tasemetele omistatud arvulisi väärtusi (veerg *keskmine kahju* z'_j) ning

järjestame tabelis olevad vaatlused veeru *keskmine kahju* z'_j järgi.

Tabel 7: Tabeli 3 väljavõte.

Taseme z_j nr	Tase z_j	Keskmine kahju z'_j
1	N-Opel	500
3	N-Volvo	700
2	M-BMW	800

Olgu meie poolt soovitud klasside arv $p = 2$. Tabelis 8 ja 9 on võimalikud klasside jaotused klasside arvu 2 korral. Märkame, et tasemeid on võimalik jagada p klassi mitmel erineval moel. Kui $p = 2$ on võimalikke klassipiire (koht, kus lõpeb üks klass ja algab teine klass) 3 taseme puhul 2. Kui tasemeid on 4, on võimalike poolituskohtade arv kahte klassi jaotamisel 3 jne.

Tabel 8: Jaotus 1.

Klass	Klassi liikmed	z'_j
1	N-Opel	500
	N-Volvo	700
2	M-BMW	800

Tabel 9: Jaotus 2.

Klass	Klassi liikmed	z'_j
1	N-Opel	500
2	N-Volvo	700
	M-BMW	800

Nägime, et tunnuse $\boldsymbol{T}$ tasemeid $z_1, z_2, \ldots, z_m$ on võimalik p klassi jagada erinevatel viisidel. Meie soov on jagada tasemed klassidesse selliselt, et moodustatud klasside klassisisene varieeruvus väljundi y mõttest oleks minimaalne ehk teisisõnu ühte klassi kuuluvad tasemed oleksid y mõttes võimalikult sarnased.[3]

[3] Tõenäosusteooria keeles oleks selle ülesande formuleering järgmine: leida parim lähend (prognoos) väljundile y mis oleks 1) tunnuse $\boldsymbol{T}$ suhtes mõõtuv ning 2) omaks ülimalt p mõõtuvat väärtust.

2.3 Optimaalsete klassipiiride määramine

Leiame nüüd punktis 2.2 kirjeldatud parimad klassipiirid ehk tükelduspunktid, mis jagavad arvväärtused $z'_1, \ldots, z'_m$ homogeenseteks klassideks.

Klassisisese varieeruvuse minimeerimiseks ja optimaalsete klassipiiride väljaselgitamiseks kasutame k-keskmise meetodit, täpsemalt Lloyd'i iteratiivset algoritmi. Lloyd'i algoritm ja k-keskmise meetod on põhjalikumalt käsitletud antud töö 3. peatükis.

Olgu D_hY tunnuse Y varieeruvus klassis K_h ning olgu pr_h klassi K_h kuulumise tõenäosus (suhteline sagedus treeningandmestikus). Klassisisene varieeruvus W avaldub klasside arvu p korral siis järgnevalt:

$$W = \sum_{h=1}^{p} pr_h D_h Y. \tag{5}$$

Raskus Lloyd'i meetodi rakendamisel seisneb selles, et reeglina ta koondub lokaalselt optimaalseks lahendiks ning lahend sõltub lähteklassipiiridest. Seetõttu testime läbi mingi küllalt suure arvu g võimalikke lähteklassipiire ja valime saadud lahenditest parima.

2.4 Klasside kaalutud keskmiste leidmine

Olgu meil teada tunnuse $\boldsymbol{T}$ tasemete $z_1, \ldots, z_m$ optimaalne klassijaotus klasside arvu p korral. Tähistagu $K_1, \ldots, K_p$ moodustatud klasse. Leiame nende klasside kaalutud keskmised tunnuse y järgi. Klassi K_h kaalutud keskmise $\overline{y}_h$ defineerime järgnevalt:

$$\overline{y}_h = \frac{1}{n_h} \sum_{x_i \in K_h} y_i, \tag{6}$$

kus n_h on klassi K_h kuuluvate vaatluste arv treeningandmestikus.

Klassi K_h keskmine on avaldatav ka arvväärtuste z'_j kaalutud keskmisena:

$$\overline{y}_h = \frac{\sum_{z_j \in K_h} n_j z'_j}{\sum_{z_j \in K_h} n_j}, \tag{7}$$

kus n_j on vaatluste arv tasemel z_j.

Näide 2.3 Vaatame tabelis 10 olevat andmestikku ning olgu parim klassijaotus toodud tabelis 11. Tähistame tabelis 11 olevad klassid vastavalt K_1 ja K_2. Klassi K_1 kuulub üks tase *N-Opel*, milles on üks vaatlus. Klassi K_1

Tabel 10: Andmestik.

Jrk	Sugu-mark	Kahju
1	N-Opel	500
2	M-BMW	700
3	M-BMW	900
4	N-Volvo	650
5	N-Volvo	750
6	N-Volvo	700

Tabel 11: Klassijaotus.

Klass	Klassi liikmed
1	N-Opel
2	M-BMW
	N-Volvo

kaalutud keskmine $\overline{y}_1$ on sel juhul ainsa vaatluse y tunnuse (*kahju*) väärtus ehk 500. Klassi K_2 kuuluvad tasemed *M-BMW* ja *N-Volvo*, milledel on vastavalt kaks ja kolm vaatlust. Seega on klassi K_2 kuuluvaid vaatlusi $2 + 3 = 5$. Klassi K_2 kuuluvad vaatlused 2-6. Vaatluste 2-6 y tunnuse summa on

$$700 + 900 + 650 + 750 + 700 = 3700.$$

Klassi K_2 kaalutud keskmine $\overline{y}_2$ on seega

$$\overline{y}_2 = \frac{1}{5}3700 = 740.$$

2.5 Väljundi prognoos

Eelmisel sammul leitud klassikeskmisi $\overline{y}_1, \ldots, \overline{y}_p$ kasutame väljundi y võimalike prognoosi väärtustena. Täpsemalt, olgu uue sisendi väärtus $\boldsymbol{x} = z_j$. Tähistame klassi, kuhu kuulub tase z_j sümboliga $K_{[j]}$ ning vastava klassi keskmise sümboliga $\overline{y}_{[j]}$. Nüüd loeme sisendile $\boldsymbol{x} = z_j$ vastava väljundi prognoosiks arvu

$$\hat{y}(\boldsymbol{x}) = \overline{y}_{[j]}. \tag{8}$$

Näide 2.4 Kasutame näites 2.3 saadud tulemusi. Tabelis 11 on kolm taset, mis jagunevad kahte klassi K_1 ja K_2. Tähistagu z_1 taset *N-Opel*, z_2 taset *M-BMW* ja z_3 taset *N-Volvo*. Kuna $K_{[1]} = K_1$ on uue sisendi $\boldsymbol{x}$ prognoosiks $\boldsymbol{x} = z_1$ korral $\hat{y}(\boldsymbol{x}) = \overline{y}_{[1]} = \overline{y}_1 = 500$. Tasemete z_2 ja z_3 korral $K_{[2]} = K_{[3]} = K_2$ ning sisendi $\boldsymbol{x} = z_2$ või $\boldsymbol{x} = z_3$ prognoosiks on $\hat{y}(\boldsymbol{x}) = \overline{y}_2 = 740$. Tabelis 12 on tunnuse tase ja uuele sisendile vastav väljundi prognoos kahe klassi korral.

Tabel 12: Väljundi prognoos kahe klassi korral.

Tase z_j	Prognoos
N-Opel	500
M-BMW	740
N-Volvo	740

Märkus. Prognoos näitab väljundi prognoosi uue sisendi $\boldsymbol{x} = z_j$ korral.

2.6 Optimaalne klasside arv

Leiame lõpuks optimaalse klasside arvu $p^* \in \{1, 2, \ldots, m\}$. Optimaalse klasside arvu p^* leidmiseks läbime peatükkides 2.1-2.5 kirjeldatud etapid iga $p = 1, 2, \ldots, m$ korral. Sellega oleme leidnud parima tasemete klassijaotuse ja vastavate klasside keskmised iga $p = 1, 2, \ldots, m$ korral. Parameetri p optimaalsuse kriteeriumiks seejuures on keskmine ruutviga (edaspidi MSE).

Kõigepealt kasutame punkti (8) leidmaks igale treeningandmestiku vaatlusele (x_i, y_i) vastava prognoosi $\hat{y}_i = \hat{y}(x_i)$, $i = 1, 2, \ldots, n$. Siis prognoosi keskmine ruutviga MSE avaldub järgnevalt:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{9}$$

Selliselt leiame MSE iga klasside arvu $p = 1, 2, \ldots, m$ korral. Tähistame leitud MSE-d vastavalt $mse_1, mse_2, \ldots, mse_m$. Optimaalne klasside arv p^* avaldub siis järgnevalt:

$$p^* = \arg\min_p (mse_p). \tag{10}$$

Näide 2.5 Vaatame tabelis 10 olevat andmestikku, mis sisaldab endas tunnuse *sugu-mark* väärtuste näol sisendeid ja tunnuse *kahju* näol teadaolevaid väljundeid. Parima klasside arvu p^* välja selgitamiseks leiame igale vaatlusele vastava prognoosi, mida kasutame MSE arvutamiseks.

Klasside arvu $p = 1$ korral on kõigi vaatluste prognoosiks uuritava tunnuse keskmine üle kõigi vaatluste, milleks on näites 1.3 leitud 700. Olgu MSE ühe klassi korral mse_1. Kasutades valemit (9) saame

$$mse_1 = \frac{1}{6} \left((500 - 700)^2 + \cdots + (700 - 740)^2 \right) = 14166.67.$$

Olgu parim klassijaotus $p = 2$ klassi korral toodud tablis 11. Vastavate tasemete prognoosid leidsime näites 2.5 ja on esitatud tabelis 12. Olgu MSE $p = 2$ klassi korral mse_2. Kasutades valemit (9) saame

$$mse_2 = \frac{1}{6} \left((500 - 500)^2 + \cdots + (700 - 740)^2 \right) = 6166.67.$$

Klasside arvu $p = 3$ korral on prognoosiks vaatlusele vastava taseme keskmine (sest andmestikus on täpselt kolm taset). Vastavad keskmised on leitud näites 2.3 ja toodud tabelis 7. Olgu MSE kolme klassi korral mse_3. Kasutades valemit (9) saame

$$mse_3 = \frac{1}{6} \left((500 - 500)^2 + (700 - 800)^2 \cdots + (700 - 700)^2 \right) = 4166.67.$$

Parima klasside arvu leidmiseks kasutame seost (10). Meie näites $p = 1, 2, 3$ ning väikseim MSE tuli klasside arvu 3 korral. Seega parim klasside arv $p^* = 3$.

3 k-keskmise meetod

Meie eesmärk on jagada nominaaltunnuse $\boldsymbol{T}$ tasemed $z_1, z_2, \ldots, z_m$ klassidesse, kasutades sealjuures vastavaid arvväärtusi $z'_1, z'_2, \ldots, z'_m$. Klasside moodustamiseks kasutame k-keskmise meetodit (*k-means clustering*[4]).

Meetodi eesmärk on minimiseerida klassisisene varieeruvus W, mis on defineeritud valemiga (5). k-keskmise meetod jagab tasemed $z_1, z_2, \ldots, z_m$ p klassi selliselt, et iga tase z_j kuulub mingisse klassi K_h, mille keskmine $\overline{y}_h$ on lähim antud taseme arvulisele väärtusele z'_j. Meetodit rakendatakse tavaliselt nn Lloyd'i iteratiivse algoritmi abil.([1], lk 460)

3.1 Lloyd'i algoritm

Olgu meil jätkuvalt teada tunnuse $\boldsymbol{T}$ tasemed $z_1, z_2, \ldots, z_m$ ja tasemetele omistatud arvulised väärtused $z'_1, z'_2, \ldots, z'_m$. Meie soov on jagada tunnuse $\boldsymbol{T}$ tasemed p klassi $K_1, K_2, \ldots, K_p$.

Lloyd'i algoritmi puhul on tegemist 2-faasilise iteratiivse protsessiga. Andes ette esialgsed p keskmist $\overline{y}_1^{(1)}, \overline{y}_2^{(1)}, \ldots, \overline{y}_m^{(1)}$ (need keskmised on tõlgendatavad kui esialgsed klasside keskmised), on algoritmi kaks faasi järgnevad [4]:

Esimene faas: klasside moodustamine. Ühe klassi moodustavad tasemed, millede arvulised väärtused on lähimad samale keskmisele.

$$K_h^{(t)} = \left\{ z_j : |z'_j - \overline{y}_h^{(t)}| \leq |z'_j - \overline{y}_{h'}^{(t)}|, \forall h' = 1, \ldots, p \right\},$$

[4]Käesolevas töös on mõisted *klaster* ja *klass* samaväärsed.

kus t tähistab iteratsiooni sammu. Iga tase z_j kuulub samaaegselt ühte klassi $K_h^{(t)}$.

Teine faas: uute keskmiste arvutamine. Uuteks keskmisteks saavad esimeses faasis moodustatud klasside kaalutud keskmised. Klasside kaalutud keskmiste leidmist käsitlesime lähemalt peatükis 2.4. Kaalutud keskmiste leidmiseks kasutame valemit (7):

$$\overline{y}_h^{(t+1)} \overset{(7)}{=} \frac{\sum_{z_j \in K_h} n_j z'_j}{\sum_{z_j \in K_h} n_j}.$$

Algoritm lõpetab töö, kui etteantud $\epsilon > 0$ korral $\left|\overline{y}_h^{(t+1)} - \overline{y}_h^{(t)}\right| < \epsilon \ \forall h = 1, \ldots, p$ korral. Algoritm on lõplikult koondunud, kui esimeses faasis moodustatud klassid enam ei muutu ehk $K_h^{(t+1)} = K_h^{(t)} \ \forall h = 1, \ldots, p$.

Algoritmi töö tulemused on (vähesel määral) sõltuvad esialgsetest keskmistest $\overline{y}_1^{(1)}, \overline{y}_2^{(1)}, \ldots, \overline{y}_m^{(1)}$. Parima klassijaotuse leidmiseks kordame kogu protseduuri mingi külllat suure arvu g erinevate algkeskmistega ning valime kõikidest saadud tulemustest parima (klassisisese varieeruvuse mõttes).

4 Optimaalsest klasside arvust p^*

Punktis 2.6 kirjeldasime, kuidas empiiriliselt (proovimise teel) leida optimaalne klasside arv p^*, mis minimiseerib ruutkeskmise prognoosivea. Antud peatükis käsitleme optimaalse p küsimust teoreetiliselt.

Järgnevas vaatleme lähemalt kaht vastandlikku olukorda, alustades juhust, kus sisend ja väljund on sõltumatud, seejärel käsitledes sisendi ja väljundi vahelise tugeva seose juhtu.

4.1 Sisendi ja väljundi sõltumatuse juht

Olgu meil $n + n_t$ vaatlust (x_i, y_i), kus x_i tähistab nominaaltunnuse $\boldsymbol{T}$ ehk sisendi väärtust i-ndal vaatlusel ja sisendile x_i vastavat väljundi väärtust tähistab y_i. Olgu tunnuse $\boldsymbol{T}$ tasemed $z_1, z_2, \ldots, z_m$. Uuele sisendile $\boldsymbol{x}$ vastava väljundi prognoosi tähistame $\hat{y} = \hat{y}(\boldsymbol{x})$. Sisendile vastava väljundi prognoosi leidmisel on olulisel kohal optimaalsete klassipiiride ja klasside arvu leidmine. Selles peatükis vaatame, milline on teoreetiliselt parim klasside arv siis, kui sisendtunnus $\boldsymbol{T}$ ja väljund y on sõltumatud ehk sõltumatute sama jaotusega juhuslike suuruste $y_1, y_2, \ldots, y_{n+n_t}$ korral. Edaspidi käsitleme olemasolevatest vaatlustest n esimest treeningandmestikuna ja n_t viimast testandmestikuna.

Olgu seega $y_1, y_2, \ldots, y_{n+n_t}$ sõltumatud sama jaotusega juhuslikud suurused. Eeldame seejuures, et $y_i \sim F$, $Ey_i = \mu$ ja $Dy_i = \sigma^2$. Kuulugu vaatluse (x_i, y_i) sisend x_i klassi K_h. Siis vastav väljundprognoos $\hat{y}_i(\boldsymbol{x})$ on klassi K_h keskmine $\overline{y}_h$. Järgnevalt analüüsime sellise prognoosi ruutkeskmist viga

$$E(MSE) = E\left(\frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \hat{y}_i)^2\right).$$

Lahutades ja liites μ saame

$$\begin{aligned} E(MSE) &= E\left(\frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \mu + \mu - \hat{y}_i)^2\right) = \\ &= E\left(\frac{1}{n_t}\sum_{i=1}^{n_t}(y_i - \mu)^2 + 2(y_i - \mu)(\mu - \hat{y}_i) + (\mu - \hat{y}_i)^2\right). \end{aligned}$$

Läheme keskväärtusega summa märgi alla

$$E(MSE)\frac{1}{n_t}\sum_{i=1}^{n_t}\left[E\left((y_i - \mu)^2\right) + E\left(2(y_i - \mu)(\mu - \hat{y}_i)\right) + E\left((\mu - \hat{y}_i)^2\right)\right].$$

Olgu n_h klassi K_h kuuluvate vaatluste arv treeningandmestikus ning olgu $y_{1'}, \ldots, y_{n'_h}$ klassi K_h kuuluvate vaatluste väljundid. Prognoosi keskväärtus

$E\hat{y}_i$ avaldub järgnevalt:

$$E\hat{y}_i = E\overline{y}_h = E\left(\frac{y_{1'} + \cdots + y_{n'_h}}{n_h}\right) =$$
$$= \frac{1}{n_h}(Ey_{1'} + \cdots + Ey_{n'_h}) = \frac{1}{n_h}(\mu + \cdots + \mu) = \mu.$$

Kuna $Ey_i = \mu$ ja $E\hat{y}_i = \mu$ siis vastavad dispersioonid on $Dy_i = E\left(y_i - \mu\right)^2$ ja $D\hat{y}_i = E\left(\hat{y}_i - \mu\right)^2$, mistõttu

$$E(MSE) = \frac{1}{n_t}\sum_{i=1}^{n_t}\left[Dy_i + E\left(2(y_i - \mu)(\mu - \hat{y}_i)\right) + D\hat{y}_i\right].$$

Analüüsime keskmist liidetavat:

$$E\left(2(y_i - \mu)(\mu - \hat{y}_i)\right) = 2\left[E(y_i - \mu)E(\mu - \hat{y}_i)\right] = 2\left[(Ey_i - E\mu)E(\mu - \hat{y}_i)\right] =$$
$$= 2\left[(\mu - \mu)E(\mu - \hat{y}_i)\right] = 0.$$

Seega oleme saanud, et

$$E(MSE) = \frac{1}{n_t}\sum_{i=1}^{n_t}(Dy_i + D\hat{y}_i). \tag{11}$$

Järgnevalt selgitame välja, milline tuleb keskmine ruutkeskmine viga klasside arvu 1 ja p korral.

Ühe klassi ($p = 1$) korral moodustub klassi K_1 keskmine treeningandmestiku kõigist väljunditest. Seega

$$\hat{y}_i = \frac{y_1 + \cdots + y_n}{n},$$

kus $\hat{y}_i$ on testandmestikku kuuluva i-nda vaatluse väljundi prognoos ja $y_1, \ldots, y_n$ on treeningandmestikku kuuluvate vaatluste väljundid.

Seega võime sõltumatust arvestades kirjutada, et

$$E(MSE_1) = \frac{1}{n_t}\sum_{i=1}^{n_t}(Dy_i + D\hat{y}_i) = \frac{1}{n_t}\sum_{i=1}^{n_t}\left[\sigma^2 + D\left(\frac{y_1 + \cdots + y_n}{n}\right)\right] =$$
$$= \frac{1}{n_t}\sum_{i=1}^{n_t}\left(\sigma^2 + \frac{1}{n}\sigma^2\right) = \sigma^2 + \frac{1}{n}\sigma^2.$$

Olgu meil nüüd $p > 1$ klassi $K_1, K_2, \ldots, K_p$. Olgu n_h klassi K_h kuuluvate vaatluste arv treeningandmestikus ning olgu $y_{1'}, \ldots, y_{n'_h}$ klassi K_h kuuluvate treeningandmestiku vaatluste väljundid. Klassi K_h kuuluvale sisendile vastava väljundi prognoos $\overline{y}_h$ on järgnev:

$$\overline{y}_h = \frac{y_{1'} + \cdots + y_{n'_h}}{n_h}$$

Leiame prognoosi $\overline{y}_h$ dispersiooni

$$\begin{aligned} D\left(\overline{y}_h\right) &= D\left(\frac{y_{1'} + \cdots + y_{n'_h}}{n_h}\right) = \\ &= \frac{1}{(n_h)^2}\left(Dy_{1'} + \cdots + Dy_{n'_h}\right) = \frac{1}{n_h}\sigma^2 \end{aligned} \tag{12}$$

Kuna klasside $h = 1, 2, \ldots, p$ sagedused $n_h < n$, siis saame võrratuse

$$D\overline{y}_h > \frac{\sigma^2}{n}. \tag{13}$$

Keskmine ruutkeskmine viga avaldub nüüd valemi (11) tõttu järgnevalt:

$$E(MSE_p) = \frac{1}{n_t}\sum_{i=1}^{n_t}(Dy_i + D\hat{y}_i) = \frac{1}{n_t}\sum_{i=1}^{n_t}(\sigma^2 + D\hat{y}_i) = \sigma^2 + \frac{1}{n_t}\sum_{i=1}^{n_t} D\hat{y}_i.$$

Vaatame eraldi liidetavat $\frac{1}{n_t}\sum_{i=1}^{n_t} D\hat{y}_i$. Meil on n_t liidetavat, mille iga väärtus on määratud seosega (12) ja mis rahuldavad võrratust (13). Sellega oleme näidanud, et

$$E(MSE_1) = \sigma^2 + \frac{1}{n}\sigma^2 < \sigma^2 + \frac{1}{n_t}\left(\frac{1}{n_h^{(1)}} + \cdots + \frac{1}{n_h^{(n_t)}}\right)\sigma^2 = E(MSE_p).$$

Seega keskmiselt on parim moodustada sõltumatutest sama jaotusega juhuslikest suurustest üks klass.

Saadud tulemus on huvitav seetõttu, et ta hoiatab sisendi-väljundi sõltumatuse ja samuti nõrga seose korral treeningandmestiku tükeldamise eest, lootuses, et saadavad klassikeskmised on paremad prognoosid kui üldkeskmine.

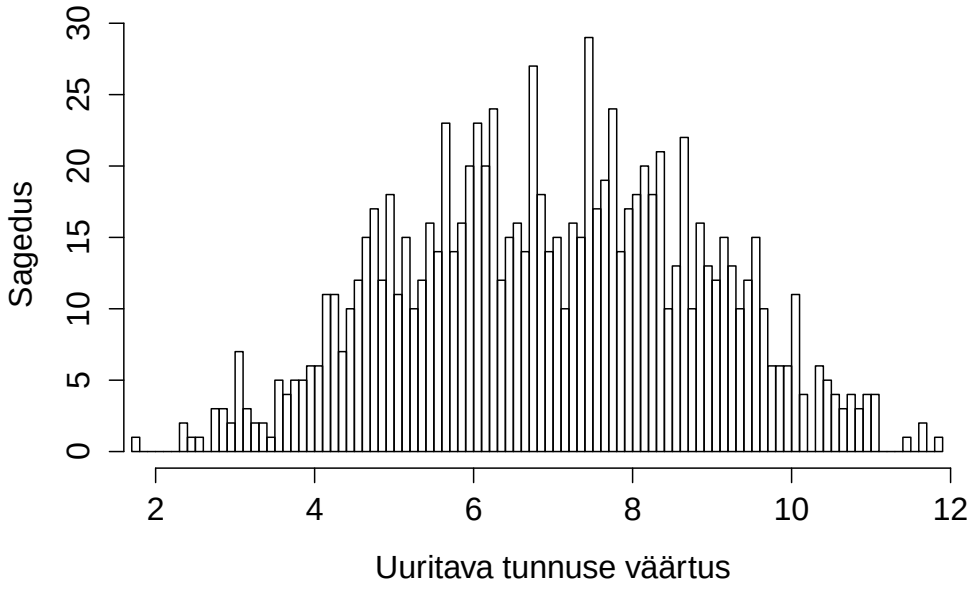
Samas on intuitiivselt arusaadav, et sisendi x ja väljundi y tugeva seose korral on andmestiku tükeldamine kasulik. Järgnev simulatsioonieksperiment ongi läbi viidud selle demostreerimiseks.

4.2 Tugevalt seotud sisend ja väljund

Eelnevalt oleme andnud teoreetilise ülevaate antud töös uuele sisendile vastava väljundi hindamiseks kasutatavast meetodist. Selles punktis testime meetodit genereeritud andmetel. Punktis 4.1 saadud tulemuse õigsuse kinnitamiseks genereerime andmeid erinevatest jaotustest ning leiame parimaid tulemusi andva klasside arvu.

4.2.1 Andmete kirjeldus

Illustreerimaks meetodi kasutamist nominaaltunnuste vaheliste seoste uurimisel, genereerime andmed kolme tunnuse A, B, C jaoks, milledel on vastavalt 3, 4, 4 taset. Neist kolmest tunnusest moodustatud liittunnusel T on seega 48 taset. Andmed genereerime selliselt, et iga liittunnuse T väärtus x_i saab talle vastava uuritava tunnuse väärtuse y_i normaaljaotusest keskväärtusega $(k+u+v)$, kus k, u ja v tähistavad tunnuse A, B ja C vastavate tasemete järjekorranumbreid, ja standardhälbega 0.5. Lühidalt, kui $x_i = z_{(kuv)}$, siis $y_i \sim N((k+u+v), 0.5)$. Näiteks kui liittunnuse tase on moodustatud tunnuse A esimesest, tunnuse B kolmandast ja tunnuse C teisest tasemest, on vastav liittunnuse tase z_{132}, mis saab endale uuritava tunnuse väärtuse jaotusest $N(6, 0.5)$. Märgime, et suurus $(k+u+v) \in \{3, 4, \ldots, 11\}$. Joonisel 1 on histogramm tuhandest sellisel viisil genereeritud uuritava tunnuse Y väärtusest. Kuna keskväärtus $(k+u+v) \in \{3, 4, \ldots, 11\}$, oleme genereerinud uuritava tunnuse Y väärtusi üheksast erinevast jaotusest. Kogu andmestiku moodustavad 1000 vaatlust (x_i, y_i), kus x_i on liittunnuse T väärtus i-ndal vaatlusel ja y_i on i-nda vaatluse uuritava tunnuse väärtus.



Joonis 1: Histogramm tuhandest uuritava tunnuse Y väärtusest.

4.2.2 Ülesande püstitus

Genereerime eelpool kirjeldatud viisil 50 andmestikku $D_1, D_2, \ldots, D_{50}$. Jagame andmestikud kahte ossa, milles mõlemas on 500 vaatlust. Andmestiku esimest osa kasutame treeningandmestikuna, millel leiame tunnuse tasemete prognoosid. Andmestiku teist osa kasutame testandmestikuna, mille peal testime antud prognooside täpsust. Valemiga (9) leiame testandmestikul MSE. Tähistagu p klasside arvu. Keskmise ruutvea leiame iga andmestiku iga klasside arvu $p = 1, 2, \ldots, 48$ korral.

Olgu $mse^{(l)} = \left(mse_1^{(l)}, \ldots, mse_{48}^{(l)}\right)$ vektor, mille iga liige $mse_p^{(l)}$ on l-ndas andmestikus leitud keskmine ruutviga klasside arvu p korral. Meie ülesanne on välja selgitada parim klasside arv MSE mõttes. Parima klasside arvu välja selgitamiseks järjestame vektoris $mse^{(l)}$ olevad väärtused (alates vähimast) ning omistame igale klasside arvule $p = 1, 2, \ldots, 48$ astaku, mis on võrdne

antud klasside arvul leitud keskmise ruutvea postsiooniga saadud paremusjärjestuses.

Sellisel viisil leiame klasside arvule $p = 1, 2 \ldots, 48$ vastava koha (astaku) iga andmestiku $D_1, D_2, \ldots, D_{50}$ korral. Parimaks klasside arvuks p^* loeme keskmiselt parima koha saanud klasside arvu.

4.2.3 Tulemused

Osas 4.1 näitasime, et sõltumatuid sama jaotusega juhuslikke suurusi sisaldavatest tasemetest on keskmiselt parim moodustada üks klass. Praegu on meil segamini 9 sellist juhtu. Kuna me genereerisime andmeid üheksast selgelt eristuvast jaotusest, siis keskmiselt parim klasside arv peaks olema 9. Tabelis 13 on toodud eelmises punktis kirjeldatud meetodit kasutades leitud klassiarvude keskmised kohad erinevate klassiarvude paremusjärjestuses (MSE järgi).

Tabel 13: Klasside karakteristikud.

Jrk	Kl. arv	kesk.koht	std.koht	kesk.MSE	std.MSE
1	9	5.76	10.81	127.70	7.73
2	10	6.12	11.40	129.58	7.85
3	11	6.46	10.50	131.06	8.13
4	12	8.92	10.93	132.61	8.26
5	14	11.26	10.06	133.74	7.49
35	48	27.70	7.83	136.03	7.50
48	1	48.00	0.00	1723.04	94.60

Märkus. Jrk näitab keskmise koha paiknemist paremusjärjestuses, kl.arv näitab klasside arvu, kesk.koht näitab mitmenda koha vaadeldav klasside arv keskmiselt sai, std.koht näitab koha standardhälvet, kesk.MSE näitab keskmist MSE-d antud klasside arvu korral, std.MSE näitab MSE standardhälvet.

Kokkuvõttes näeme, et meie meetod jõudis tulemuseni, mis on kooskõlas genereeritud andmete struktuuriga: mõlemal juhul on tegemist 9 oluliselt erineva vaatluste klassiga.

5 Meetodi rakendamine reaalsetel andmetel

5.1 Andmete kirjeldus

Tegemist on kaskokindlustuse andmestikuga, mis koosneb 15732 vaatlusest. Andmestik sisaldab informatsiooni kindlustatava isiku ja sõiduki omaduste kohta, kindlustaja poolt välja makstud kahjusumma ja maksmata prognoositava kahju ehk reservi suurust ning poliisi kestvust päevades. Järgnevas on läbi viidud 2 analüüsi, mis erinevad sisendtunnuse valiku poolest. Esimesel juhul võtsime sisendtunnuseks tunnuse *mark*, teisel juhul liittunnuse *sugu-mark*, kus *sugu* on kindlustatava isiku sugu ja *mark* on kindlustatava sõiduki mark. Väljundina kasutasime makstud kahjusumma ja reservi kokkuliitmisel saadav summat, mille jagasime poliisi kestvusega (kahjusumma ühe päeva kohta), edaspidi *kahju*. Analüüsisis kasutasime vaid neid vaatlusi, mis sisaldasid informatsiooni vastava sisendtunnuse ja väljundi kohta - esimeses analüüsis tunnuste *mark* ja *kahju* kohta ja teises analüüsis mõlema tunnuse - *sugu-mark* ja *kahju* kohta. Vaatluste arv esimeses analüüsis oli 15732. Andmestikus on infot 50 erineva automargi kohta. Vähem kui 20 vaatlust on 17-ne automargi kohta. Vaatluste arv teises analüüsis 7566. Tunnuse *sugu-mark* erinevaid tasemeid on andmestikus 78, milledest 29-l on vähem kui 20 vaatlust.

5.2 Analüüsi käik

Kaskokindlustuse andmed olid esialgu järjestatud poliiside kestvuse alusel (kauem kestnud poliisid eespool). Objektiivsema hinnagu andmiseks randomiseerisime andmestiku järjekorra mõttes. Seejärel jagasime andmestiku kaheks. Esimest osa kasutasime treeningandmestikuna ja teist testandmestikuna. Märgime, et tunnuse *mark* 50-st tasemest jäi treeningandmestikku vaatlusi 46 taseme ja tunnuse *sugu-mark* 78-st tasemest vaatlusi 61 taseme kohta. Seejärel jagasime sisendtunnuse tasemed klassidesse ning leidsime klasside keskmised, milledest said klassi kuuluvate tasemete *kahju* prognoosid. Leitud prognoose kasutasime testandmestiku *kahjude* hindamiseks ning arvutasime MSE. Läbisime kirjeldatud protsessi kõikide võimalike klasside arvu, esimesel juhul $p = 1, \ldots, 46$ ja teisel juhul $p = 1, \ldots, 61$, korral. Parimaks klasside arvuks lugesime vähima MSE andnud klasside arvu.

5.3 Tulemused

5.3.1 Analüüs I

Viie parima klasside arvu p tulemused on toodud tabelis 14.

Tabel 14: Kahjuprognooside MSE erinevate klasside arvu korral.

Paremus jrk	Klasside arv	MSE
1	5	63171.77
2	4	63177.73
3	2	63178.68
4	6	63181.33
5	10	63181.84

Antud andmetel on parim klasside arv $p^* = 5$. Meie poolt pakutud mee-

tod tuvastas seega väljundtunnuse *kahju* sõltuvuse sisendtunnusest *mark*, sealjuures leides optimaalse sisendtunnuse tasemete klassijaotuse.

5.3.2 Analüüs II

Viie parima klasside arvu p tulemused on toodud tabelis 15.

Tabel 15: Kahjuprognooside MSE erinevate klasside arvu korral.

Paremus jrk	Klasside arv	MSE
1	1	30486.29
2	2	31409.40
3	3	31375.28
4	4	31429.55
5	5	31498.59

Antud andmetel on parim klasside arv $p^* = 1$. Meie poolt pakutud meetod ei tuvastanud antud andmetel väljundtunnuse *kahju* olulist sõltuvust sisendtunnuse rollis olevast liittunnusest *sugu-mark*. Järeldus põhineb sellel, et klasside arvu 1 korral on iga uue sisendi prognoosiks treeningandmestiku üldkeskmine.

Saadud tulemuste üks võimalik põhjus on sõltuvuse tegelik puudumine. Teine võimalik põhjus on sisend- ja väljundtunnuse keskmine või nõrk seos, mida käesoleva bakalaureusetöö raames ei käsitletud.

Töös väljapakutud meetod andis siiski huvipakkuvaid tulemusi, kasutades nominaalsete sisendtunnuste vahelisi seoseid väljundtunnuse väärtuste hindamiseks. Meetodi efektiivsuse väljaselgitamiseks tuleks seda mitmekülgsemalt testida, mis loob võimaluse töö edasiarendamiseks.

Kokkuvõte

Lähinaabrite meetod on mitteparameetriline tehnika, mis kasutab väljundtunnuse prognoosimiseks vastava sisendi lähiümbrust. Meetod toimib põhimõttel, et uue sisendi kohta annavad rohkem informatsiooni talle lähedal olevad vaatlused.

Antud töös keskendusime nominaalsetele sisendtunnustele ning nendevaheliste seoste uurimisele ning pakkusime selleks välja ühe võimaliku meetodi. Töös omistasime igale nominaaltunnuse tasemele arvulise väärtuse, mis on võrdne uuritava väljundunnuse keskmisega antud tasemel. Omistatud väärtusi kasutasime nominaaltunnuse tasemete grupeerimiseks (klassidesse jagamiseks), milles lähtusime ideest, et ühe klassi peaksid moodustama sarnase arvväärtusega tasemed. Moodustatud klasside keskmisi väljundväärtusi kasutasime uutele sisenditele vastavate väljundite prognoosidena. Meetod leiab ühtlasi ka optimaalse klasside arvu, kusjuures kriteeriumiks on võetud keskmine prognoosi ruutviga (MSE). Teoreetilise analüüsi abil näitasime, et sõltumatutest sama jaotusega juhuslikest suurustest on parim moodustada üks klass.

Kirjeldatud meetodit testisime nii genereeritud kui ka reaalsetel empiirilistel andmetel. Genereeritud andmete puhul oli tegemist tugevalt seotud sisend- ja väljundtunnusega. Empiiriliste andmete näol oli tegemist kaskokindlustuse andmetega, kus hindasime kahel juhul kindlustusandjale tekkiva kahju suuruse seost kindlustusvõtjat iseloomustavate tunnustega. Parima klasside arvu väljaselgitamiseks arvutasime kahjuprognoosi keskmise ruutvea.

Töös veendusime, et väljapakutud meetod saab hästi hakkama tugevalt seotud sisend- ja väljundtunnuste korral. Meetodi võime tuvastada nõrka või keskmist väljundtunnuse sõltuvust sisendtunnusest vajaks aga edasist uurimist.

Nearest neighbours prediction with categorical variables Bachelor thesis

Reigo Hendrikson

Summary

The purpose of this thesis is to show how the k-nearest neighbours method can be used with categorical inputs. It concentrates on estimations based on categorical features. In chapter 1 an alternative function for distance measure among categorical features is described. This new distance measure is used to divide categories into homogeneous classes. Chapter 2 gives an overview of the method used to estimate output value for new input. The thesis also gives a brief overview of Lloyd's algorithm which is used to find optimal class boundaries and is described in chapter 3. Chapter 4 consists of discussion about optimal number of classes including some theoretical results. Chapter 5 focuses on testing this technique on real empirical data.

Kasutatud kirjandus

[1] Hastie, T., Tibshirani, R. ja Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

[2] Pärna, K., R. Kangro, A. Kaasik, M. Möls. 2012. *K-Nearest Neighbors as Pricing Tool in Insurance: a Comparative Study.*

[3] Lepik, K. 2012. *Lähinaabrite meetod ja selle rakendamine.*

[4] MacKay, D. 2003 *Information Theory, Inference and Learning algorithms.* Cambridge University Press, lk 284-292

Lisa A. Reaalsete andmete analüüsis II kasutatud R-kood

```
# sisendtunnus  ← sugumark
# väljundtunnus ← kahju

library(plyr)
z_kesk←ddply(treening,~sugumark,summarise,mean=mean(kahju))
# Meil on andmestik, milles on tunnused sugumark ja kahju
# Valmistame andmestiku ette
kahjuprognoos_andmestik←function(treening, test,z_kesk,klassidearv){
  treening_kesk←mean(treening$kahju)
  # Valmistame andmestiku ette funktsiooni "kmeans" kasutamiseks
  pr←rep(NA,length(treening$kahju))
  for (i in 1:length(treening$kahju)){
    pr[i]←z_kesk[z_kesk$sugumark==treening$sugumark[i],2]
  }
  treening←data.frame(treening,pr)
  cells ← treening$pr
  rnames ← treening$sugumark
  cnames ← c("pr")
  x ← matrix(cells, nrow=length(treening$sugumark), ncol=1, byrow=TRUE,
  dimnames=list(rnames, cnames))
  mse←rep(NA,length(unique(treening$sugumark)))
  klassi_nr←rep(NA,length(unique(treening$sugumark)))
  km ← kmeans(x,klassidearv, nstart=500,algorithm = "Lloyd",iter.max=25)
  klass←as.vector(km$cluster)
  klasskesk←(km$centers)
  kl←c(1:klassidearv)
  klasskesk1←data.frame(kl,klasskesk)
  treening←data.frame(treening,klass)
  prognoos←rep(NA,length(treening$kahju))
  for (j in 1:length(treening$kahju)){
    prognoos[j]←klasskesk1[klasskesk1$kl==treening$klass[j],2]
  }
  treening←data.frame(treening,prognoos)
  # Saame andmestiku, kus on tunnuse tase ja taseme prognoos.
  z_prognoos←ddply(treening,~sugumark,summarise,mean=mean(prognoos))
  # Kasutame saadud prognoose testandmestiku väljudnite prognoosimiseks
  test_prognoos←rep(NA,length(test$kahju))
  for (k in 1:(length(test$kahju))){
    if (test$sugumark[k] %in% z_prognoos$sugumark){
      test_prognoos[k]←z_prognoos[z_prognoos$sugumark==test$sugumark[k],2]
    }
    else{
      test_prognoos[k]←treening_kesk
    }
  }
  test←data.frame(test,test_prognoos)
  return(test)
}
```

```
# Funktsioon MSE-de leidmiseks
MSE<-function(treening,test,z_kesk){
  maks_klasse<-length(unique(z_kesk$mean))
  mse<-rep(NA,maks_klasse)
  klassi_nr<-rep(NA,maks_klasse)
  for (i in 1:length(unique(z_kesk$mean))){
    mse_an<-kahjuprognoos_andmestik(treening, test,z_kesk,i)
    klassi_nr[i]<-i
    mse[i]<-(1/(length(test$kahju)))*(sum((mse_an$kahju-mse_an$test_prognoos)^2))
  }
  tulemus<-data.frame(klassi_nr,mse)
  return(tulemus)
}
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Reigo Hendrikson (sünnikuupäev: 11.12.1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose Nominaalsete sisendtunnuste vaheliste seose kasutamine lähinaabrite meetodil, mille juhendaja on proffessor Kalev pärna

   (a) reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   (b) üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus 06.05.2013