

Kandidaatin tutkielma
Rikoksenuusinnan ennustaminen kausaalipäättelyllä

Riku Laine
Valtiotieteellinen tiedekunta, Helsingin yliopisto

25. maaliskuuta 2019

Sisältö

1	Kiitokset – Acknowledgements	3
2	Tiivistelmä – Kypsyysnäyte?	4
3	Johdanto	5
3.1	Takuukäsittely prosessina	5
3.2	Yhteiskunnallinen merkitys ja kritiikki	6
3.3	”Kausaalipäätely uutena paradigmana”	6
3.4	Valikoitumisharha	7
4	Data	8
4.1	COMPAS	8
4.2	Synteettinen	8
5	Metodit	10
5.1	Aiemmat tutkimukset	10
5.2	Validointimetodit	10
5.3	Verkkoteoria	10
5.4	Kausaalipäätely	11
5.4.1	Johdanto?	11
5.4.2	Merkinnät	11
5.4.3	Määritelmät	12
5.4.4	Malli	12
6	Tulokset	14
6.1	Synteettinen	14
6.2	Compas	14
7	Diskussio	15

Kirjallisuutta	16
Liitteet	17
A Abstract in English?	17

Luku 1

Kiitokset – Acknowledgements

Tämän tutkielman aikana on tullut esiin takuujärjestelmään liittyvät ongelmat ja sovel-
lusalueen yhteiskunnallinen merkitys. Tutkielman teko on ollut minulle erityisen miele-
kässtä antoisan aiheen ja mieleisten yhteistyökumppanien vuoksi. Olen kirjoittanut tämän
kandidaatintutkielman yhteistyössä Helsingin yliopiston tietojenkäsittelytieteen osaston
apulaisprofessorin Michael Mathioudakis ja tohtoritutkijan Antti Hyttisen kanssa. He
tarjosivat minulle aiheen ja merkittävää tukea sekä tärkeitä kommentteja tämän tutkiel-
man kirjoittamisen aikana.

Tämän tutkielman on tarkastanut XYZ. Haluan kiittää kaikkia edellä mainittuja hen-
kilöitä sekä ystäviäni ja perhettäni, jotka tukivat minua tämän tutkielman tekemisessä.

Helsingissä XX.XX.2019

Riku Laine

I would like to wholeheartedly thank assistant professor Michael Mathioudakis from
University of Helsinki's Department of Computer Science for numerous things. He pro-
vided me this extremely interesting thesis topic and provided insightful and encouraging
comments throughout the process. Antti Hytiinen from the same department also gave
important insight in the causal modelling.

Luku 2

Tiivistelmä – Kypsyysnäyte?

Johdanto-luvussa esittelen ongelman asettelun ja tilanteen yleisen viitekehyksen. Keskustelemme rikoksenuusinnan ennustamisesta yhdysvaltalaisessa oikeusjärjestelmässä. Esitän kappaleessa yleisen kuvauksen takuukäsittelyn etenemisestä oikeusprosessina, jonka jälkeen pohdin hieman takuukäsittelyn yhteiskunnallista merkitystä ja motivaatiota hyvään ennusteeseen. Kappaleen lopussa kirjoitan hieman kausaalipäätelystä uutena tilastotieteellisenä paradigmana [6].

Kappaleessa 4 esittelen käyttämäni datalähteet ja niiden ominaispiirteet. Esitän COMPAS-tietojen ominaispiirteet ja *jotain muuta*. Esitän myös kuinka olen luonut analyyseissä myöhemmin käytettävän datasetin mukaillen Lakkarajun vuoden 2017 konferenssijulkaisua [3].

Metodit-kappaleessa esitän käyttämäni mallit ja metodit. Teen lyhyen katsauksen aikaisempaan kirjallisuuteen ja tutkimuksiin tällä sovellusalalla. Käyn lisäksi läpi tässä tutkielmassa myöhemmin käytettäviä matemaattisia sekä verkkoteoreettisia merkintöjä ja määritelmiä. Teen joitakin osoituksia ja osoitan kuinka mallimme ei riipu havaitsemattomista (unobservables) muuttujista.

Luvussa 6 esitän algoritmillani saavuttamani tulokset ja vertailen niitä Lakkarajun [3] saavuttamiin. Olen eritellyt erillisiin alalukuihin synteettisellä ja COMPAS-dataseilla saavutetut tulokset.

Viimeisessä kappaleessa *Diskussio* esitän mallien ja tutkielmani virhelähteet ja muut ongelmat sekä keskustelen tulosten mahdollisesta vaikutuksesta, sikäli niitä sovellettaisiin sikäläisen oikeuslaitoksen toimintaan.

Luku 3

Johdanto

Tässä kappaleessa esittelen tutkielman taustaa ja yleisellä tasolla yhdysvaltalaisen oikeuslaitoksen takuukäsittelyprosessin. Sen jälkeen paneudun hieman vangitsemispäätöksen yhteiskunnalliseen merkitykseen: minkä takia ihmisiä vangitaan ja mitä perusteita on vangitsemattajättämispäätökselle. Pyrin luvun aikana myös hieman selvittämään takuujärjestelmän käyttöä Suomessa ja kappaleen lopussa pohdin hieman kausaalipäyttelyä paradigman muutoksena tilastotieteen kentällä. Jätän kuitenkin tarvittavien merkintöjen esittämisen kappaleeseen *Merkinnät* ja mallin esittelyn *Malli*-lukuun.

3.1 Takuukäsittely prosessina

Yhdysvalloissa, kuten monissa muissa anglosaksisissa maissa, on käytössä järjestelmä, jota nimitetään takuu- tai vakuusjärjestelmäksi. Takuujärjestelmä on epäillyn vaihtoehto tutkintavankeudelle hänen odottaessaan oikeudenkäyntiä ja Yhdysvalloissa oikeus takuuseen periytyy maan perustamisen ajalta [1, 7]. Suomen oikeus- ja sisäasiainministeriön alaisen esitutkinta- ja pakkokeino- toimikunnan mukaan takuujärjestelmiä on kolmenlaisia: kahdessa niistä epäilty maksaa itse käteisellä vakuuden tai asettaa omaisuuttaan vakuudeksi ja kolmannessa jokin ulkopuolinen taho ”menee takuuseen epäillyn velvollisuuksien täyttämiseksi” [1].

Yhdysvalloissa epäillyn pidätyksen jälkeen hänet viedään paikallisen oikeusviranomaisen järjestämään takuukuulemiseen (bail hearing) [7]. Kuulemisessa päätetään takuun myöntämisestä, eli voidaanko epäilty vapauttaa, vai halutaanko hänet asettaa vankeuteen ennen oikeudenkäyntiä. Kuulemisessa päätetään myös mahdollisen takuun määrästä sekä vapauttamisen ehdoista [7]. Takuu voidaan suorittaa taattuna tai takaamattomana maksusitoumuksena tai maksaa suoraan (cash) – erityistapauksissa epäilty voidaan vapauttaa myös pelkällä kirjallisella sitoumuksella (release on personal recognizance (ROR)) [7].

3.2 Yhteiskunnallinen merkitys ja kritiikki

Zaniewski toteaa lyhyessä kirjallisuuskatsauksessaan, että takuujärjestelmän vuoden 1982 uusitus ei onnistunut laskemaan tarpeettomia vangitsemisia – päinvastoin niiden suhteellinen määrä kaksinkertaistui 22%:sta 49%:iin vuodesta 1984 vuoteen 2007. Nykyisellään sikäläinen oikeusjärjestelmä suosii suoraan rahalla maksettavia tai taatuilla maksusitoumuksilla hoidettuja takuita, mikä asettaa huonossa taloustilanteessa olevat epäilyt eri tilanteeseen. [7]

Suomessa vakuusjärjestelmää ei ole käytetty, vaikka yllä mainittu toimikunta toteaa-kin sen sisältyvän tullilain 44 §:ään. Kyseisessä pykälässä ”- - säädetään mahdollisuudesta asettaa pidätetyn tai vangitun vapaaksi päästämi[s]en ehdoksi, että hän asettaa vakuuden, jonka harkitaan takaavan hänen saapumisensa oikeudenkäyntiin ja ehkä tuomittavien seuraamusten suorittamisen”. Kuten he tarkentavat, lisäksi usein edellytetään, että epäilty ei asu Suomessa, ja epäillään hänen pakenevan maasta ennen oikeudenkäyntiä tai rangaistusta [1]. Sekä yhdysvaltalaiselle että suomalaiselle järjestelmälle on yhteistä, että takuu tuomitaan menetettäväksi valtiolle, jos vapauden ehtoja rikotaan.

Kritiikkiä on esitetty molemmissa maissa osaltaan samoihin asioihin. Suomessa pykälää ei ole sovellettu, koska luultavasti sen tulkintaohjeet ovat niin niukat, kuten myös sääntely [1]. Yhdistävänä kritiikkinä sekä Zaniewski että esitutkinta- ja pakkokeinotoimikunta mainitsevat muun muassa sen, kuinka takuumaksujen toimeenpano vaikuttaa tai Suomen tapauksessa vaikuttaisi pienituloisten taloustilanteeseen [7, 1]. Suomalainen toimikunta esittää lisäksi monia muitakin ongelmakohtia, sikäli takuujärjestelmä haluttaisiin ottaa Suomessa käyttöön, esimerkiksi he toteavat, että vakuusmaksujen maksamiseen tulisi todennäköisesti liittymään ”epätoivottavia lieveilmiöitä” [1]. Tähän ongelmaan on Yhdysvalloissa jo osittain reagoitukin, sillä esimerkiksi Californian osavaltio päätti viime vuonna poistaa takuumaksut käytöstä [5].

3.3 ”Kausaalipäätteleminen uutena paradigmana”

Haluamme siirtyä assosiatiivisesta päättelystä kausaalipäätteeseen, koska definitiivisten päätöksien tekeminen muuten hankalaa. (Onko paraneminen seuraus lääkkeen käytöstä, nyt voimme sanoa vain että käyttäneillä on parempi ennusta. semantiikkaa vain?) Lisäksi on ylitettävä korrelaatio ei ole kausaali -kynnys, erityisesti [6].

* Kausaalipäätteleminen vaatii uutta laskentaa, *do*-laskento (calculus), myös Miksi-kirjan käännöksen tee-laskento. * Päätteleminen nojaa vahvasti / tarvitsee mallin, joka ilmaistaan (usein/aina) verkkona, josta voidaan suoraan lukea muuttujien väliset riippuvuussuhteet. * Usein funktionaalista muotoa ei määritellä, lisää tähän ne nuoliversiot yhtälöistä havainnollistamaan, että siirrytään yhtäsuuruudesta määrätymiseen [2] * Esimerkkejä

Miksi-kirjasta väärin määritellyistä malleista? Esimerkkejä aloista, joilla jo käytetty, oleelliset pointit historiasta

3.4 Valikoitumisharha

Datassa on valikoitumisharha, mistä Lakkaraju käyttää termiä ”*selective labels*” [3]. Datat harha johtuu luonnollisesti siitä, että rikoksen voi uusia vain, jos tuomari päättää vapauttaa takuita vastaan. Suorat päättelytavat – *counterfactual inference* – ovat ongelmallisia siinä mielessä, että jne jne.

Luku 4

Data

Tässä luvussa kuvaillaan käytetyt datasetit ja niiden ominaispiirteet.

4.1 COMPAS

Dataa broward Countysta

4.2 Synteettinen

Synteettinen data luotiin Lakkarajun artikkelissaan selostamalla tavalla [3]. Dataan simuloitiin kolme muuttujaa X , Z , ja W . Näistä muuttujista X vastaa informaatiota, joka on sekä mallin että tuomarin havaittavissa, eli informaatiota, joka on kirjattu oikeuden pöytäkirjoihin tai on kerättävissä muista rekistereistä, kuten vastaajan sukupuoli. Muuttujalla Z kuvataan tietoa, jonka vain tuomari voi havaita: kuten Lakkaraju havainnollistaa, tällaista voi olla esimerkiksi tieto siitä, onko vastaajalla perhettä mukana oikeussalissa [3]. W on mallissa havainnollistamassa reaali maailmaa. Muuttujalla esitämme datassa informaatiota, joka ei ole saatavilla päätöksentekijöille eikä mallille mutta vaikuttaa silti rikoksenuusimisriskiin. Datassa nämä ovat kaikki riippumattomia standardinormaalijakautuneita satunnaismuuttujia, eli $X, W, Z \sim N(0, 1) \perp$.

Yhdistämme henkilöt satunnaisesti kuhunkin $M = 500$ tuomariin, joista jokaiselle määritellään hyväksymisprosentti $r \in [0, 1]$. Tuomarin hyväksymisprosentti määritetään ottamalla arvoja tasajakaumasta suljetulta väliltä $[0, 1; 0, 9]$ ja sitten pyöristämällä ne 10 desimaalin tarkkuuteen. Tulostuottaja Y simuloidaan määrittämällä sen ehdollinen todennäköisyys seuraavasti: $\mathbb{P}(Y = 0 | X, Z, W) = \frac{1}{1 + \exp\{-(\beta_X X + \beta_Z Z + \beta_W W)\}}$, missä kertoimet β_X , β_Z ja β_W on asetettu arvoihin 1, 1 ja 0,2 vastaavassa järjestyksessä. [3]

Päätösmuuttujan T ehdollinen todennäköisyys $\mathbb{P}(T = 0|X, Z) = \frac{1}{1 + \exp\{-(\beta_X X + \beta_Z Z)\}} + \epsilon$ missä $\epsilon \sim N(0, 0, 1)$ vastaa pientä määrää kohinaa. Henkilöltä i kielletään takuut, eli $T_i = 0$ jos muuttujan T ehdollinen todennäköisyys on tuomarin j suurimman $(1-r) \cdot 100\%$ joukossa. Lopuksi koulutusdata suodatettiin siten, että saatavissa oli vain yksilöt, jotka päästettiin vapaaksi ($T = 1$). [3]

Luku 5

Metodit

Tässä kappaleessa selostan analyyseissa, mallinnuksessa ja validoinnissa käyttämäni menet.

5.1 Aiemmat tutkimukset

Aiemmat tutkimukset ovat lähestyneet monesta näkökulmasta, mutta ilman kausaatiota.

5.2 Validointimetodit

Ristiin taulukoinnit yms.

5.3 Verkkoteoria

Esitän tässä kappaleessa lyhyesti kaikki tarvittavat verkkoteoreettiset määritelmät, joita tulen hyödyntämään. Noudatan määritelmissä Oinosta [4].

Määritelmä 5.1 (Suunnattu verkko). *Suunnattu verkko* G on pari (V, E) , missä $V \neq \emptyset$ on solmujen joukko ja

$$E = \{(a, b) \in V \times V \mid \text{solmusta } a \text{ on nuoli solmuun } b\}$$

on *kaarien* joukko.

Määritelmä 5.2. Oletetaan, että $G = (V, E)$ on suunnattu verkko ja $a, b \in V$.

Merkintä $a \rightarrow b$ tarkoittaa, että $(a, b) \in E$. Tällöin sanotaan, että a on kaaren (a, b) *lähtösolmu* ja b on kaaren (a, b) *maalisolmu*. Sanotaan myös, että solmu b on solmun a *vierussolmu*.

Jos $(a, a) \in E$, sanotaan suunnatussa verkossa olevan *silmukka* solmussa a .

Määritelmä 5.3 (Vierekkäisyys). Oletetaan, että $G = (V, E)$ on suunnattu verkko ja $a, b \in V$.

Jos solmujen a ja b välillä on nuoli, niin solmujen a ja b sanotaan olevan *vierekkäisiä*.

Määritelmä 5.4 (Yksinkertainen suunnattu verkko). Oletetaan, että $G = (V, E)$ on suunnattu verkko, jossa ei ole yhtään silmukkaa eli $(v, v) \notin E$ kaikilla $v \in V$.

Tällöin sanotaan, että G on yksinkertainen suunnattu verkko.

Määritelmä 5.5 (Polku ja suunnattu polku). Oletetaan, että G on yksinkertainen verkko ja $n \in \mathbb{N}, n \geq 1$.

Verkon G solmujen jono v_1, \dots, v_n on *polku* solmusta v_1 solmuun v_n , jos jonon jokaisesta solmusta on kaari jonon seuraavaan solmuun. Polkua voidaan merkitä $a \rightsquigarrow b$.

Jos verkko G on suunnattu verkko ja kaikki polun $a \rightsquigarrow b$ kaaret kulkevat kaarien suuntien mukaisesti, voidaan täsmentää, että polku $a \rightsquigarrow b$ on *suunnattu polku*.

5.4 Kausaalipäättely

Erityisesti [6]. Esittele merkunnät, määritelmät ja malli. Käännökset Miksi-kirjaa mu-
kaillen?

5.4.1 Johdanto?

5.4.2 Merkinnät

Kausaalipäättelyssä käyttävät merkinnät noudattelevat pitkälle tavallista todennäköisyysskennan merkintöjä. Kun yritetään selvittää muuttujan X vaikutusta muuttujaan Y ja tehtään interventio, siten että muuttuja X asetetaan arvoon x_0 , merkitseen sitä $\mathbb{P}(Y|\text{do}(X = x_0))$.

5.4.3 Määritelmät

Määritelmä 5.6. Joukko \mathcal{S} sulkee / katkaisee (blocks) polun p , jos vähintään toinen seuraavista ehdoista on voimassa:

- (a) Polku p sisältää vähintään yhden solmun, joka on jonkin kaaren lähtösolmu ja kuuluu joukkoon \mathcal{S} . (arrow-emitting)
- (b) Polku p sisältää vähintään yhden käänteisen haarukkasolmun (collision node), joka ei kuulu joukkoon \mathcal{S} ja jolla ei ole jälkeläisiä joukossa \mathcal{S} .

Määritelmä 5.7. Oletetaan, että halutaan selvittää (satunnais)muuttujan X kausaalista vaikutusta muuttujaan Y . Joukko \mathcal{S} on riittävä adjustmenttiin, kun seuraavat ehdot ovat voimassa:

- (1) Yksikään joukon \mathcal{S} alkioista ei ole solmun X jälkeläinen.
- (2) Joukon \mathcal{S} alkiot ”blokkaavat” kaikki määritelmän 5.6 mukaiset ”takaovireitit” solmusta X solmuun Y .

5.4.4 Malli

Algoritmi 1 Kausaalialgoritmi

Syöte: Data $(\mathbf{x}, t, y) \in \mathcal{D}_t, \mathcal{D}_v$ ja hyväksymisaste $r \in [0, 1]$, missä \mathcal{D}_t on testidata ja \mathcal{D}_v validointidata.

Palauttaa: $\mathbb{P}(Y = 0 | \text{do}(R = r))$

- 1: Määritä $f(x) = \mathbb{P}(X = x)$ testidatasta.
 - 2: Ennusta vastetta Y selittävillä muuttujilla X käyttäen harjoitusdatan havaintoja, joilla $T = 1$.
 - 3: Määritä harjoitusdatan jokaiselle havainnolle $P(Y = 0 | X = x)$ käyttäen yllä olevaa mallia.
 - 4: Järjestä havainnot nousevaan järjestykseen edellisen kohdan todennäköisyyksien mukaan.
 - 5: Alusta muuttuja **summa** = 0.
 - 6: **for all** Jokaiselle parametriavaruuden pisteelle **do**
 - 7: $p_x \leftarrow P(X = x)$
 - 8: $\mathcal{D}_{\S} \leftarrow \{\mathcal{D} | X = x\}$
 - 9: Assign first $r \cdot 100\%$ observations from \mathcal{D}_{\S} to \mathcal{D}_{rx}
 - 10: $p_t \leftarrow \frac{|\{\mathcal{D}_{rx} | T = 1\}|}{|\mathcal{D}_{rx}|}$
 - 11: $\mathcal{D}_{tx} \leftarrow \{\mathcal{D}_x | T = 1\}$
 - 12: $p_y \leftarrow \frac{|\{\mathcal{D}_{tx} | Y = 0\}|}{|\mathcal{D}_{tx}|}$
 - 13: Lisää muuttujaan **summa** tulo $p_y \cdot p_t \cdot p_x$
 - 14: **end for**
 - 15: **return** **summa**
-

Luku 6

Tulokset

6.1 Synteettinen

6.2 Compas

Luku 7

Diskussio

Kirjallisuutta

- [1] Esitutkinta- ja pakkokeino-toimikunta: *Esitutkintalain, pakkokeinolain ja poliisilain kokonaisuudistus: esitutkinta- ja pakkokeino-toimikunnan mietintö*. Oikeusministeriö, Helsinki, 2009, ISBN 978-952-466-824-8. sivut 128–131.
- [2] Kalisch, Markus ja Peter Bühlmann: *Causal structure learning and inference: a selective review*. Quality Technology & Quantitative Management, 11(1):3–21, 2014.
- [3] Lakkaraju, Himabindu, Jon Kleinberg, Jure Leskovec, Jens Ludwig ja Sendhil Mullainathan: *The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables*. Teoksessa *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, sivut 275–284, New York, NY, USA, 2017. ACM, ISBN 978-1-4503-4887-4. <http://doi.acm.org.libproxy.helsinki.fi/10.1145/3097983.3098066>.
- [4] Oinonen, Lotta: *Johdatus yliopistomatematiikkaan*, Tammikuu 2016. Samannimisen kurssin kurssimateriaali.
- [5] Park, Madison: *California eliminates cash bail in sweeping reform*. CNN, Elokuu 2018. <https://edition.cnn.com/2018/08/28/us/bail-california-bill/index.html>.
- [6] Pearl, Judea: *An introduction to causal inference*. Int J Biostat, 6(2):Artikkeli 7, Helmikuu 2010. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2836213/>.
- [7] Zaniewski, Amanda: *Bail in the United States: A Review of the Literature*. <https://www.mass.gov/files/documents/2016/09/qx/bail-in-united-states-literature-review.pdf>, Marraskuu 2014. PDF, haettu 12.3.2019.

Liite A

Abstract in English?

The contents...