Aalto-yliopisto
Perustieteiden
korkeakoulu

Tarkan liikennedatan tilastollinen analyysi ja ennustaminen koneoppimismenetelmin

Suvi Laine

27.08.2020

Ohjaaja: *Henri Salmenjoki*
Valvoja: *Pauliina Ilmonen*

Systeemianalyysin
laboratorio

Tausta

Miksi liikennedatan analyysi on mielenkiintoista?

- Matka-aikojen optimointi
- Reittien suosittelu ruuhkien perusteella
- Tieliikenneverkon suunnittelu

Liikenneaineistoa voidaan kerätä eri tavoin:

- Perinteiset tavat: tiekamerat, mittauspisteet tien varrella
- Autoon kiinnitettävät sensorit (Floating Car Data)

Tausta

- Tutkimusaineisto RoadCloudilta: kerätty ajoneuvoihin kiinnitettävillä sensoreilla
  - Liikenteen tilan tunnistus ja ennustaminen
- Datasta saadaan myös informaatiota sääolosuhteista
  - mm. tarkka sijainti ja aika, nopeus, tien kunto, kiihtyvyys, kitkakerroin, lämpötila

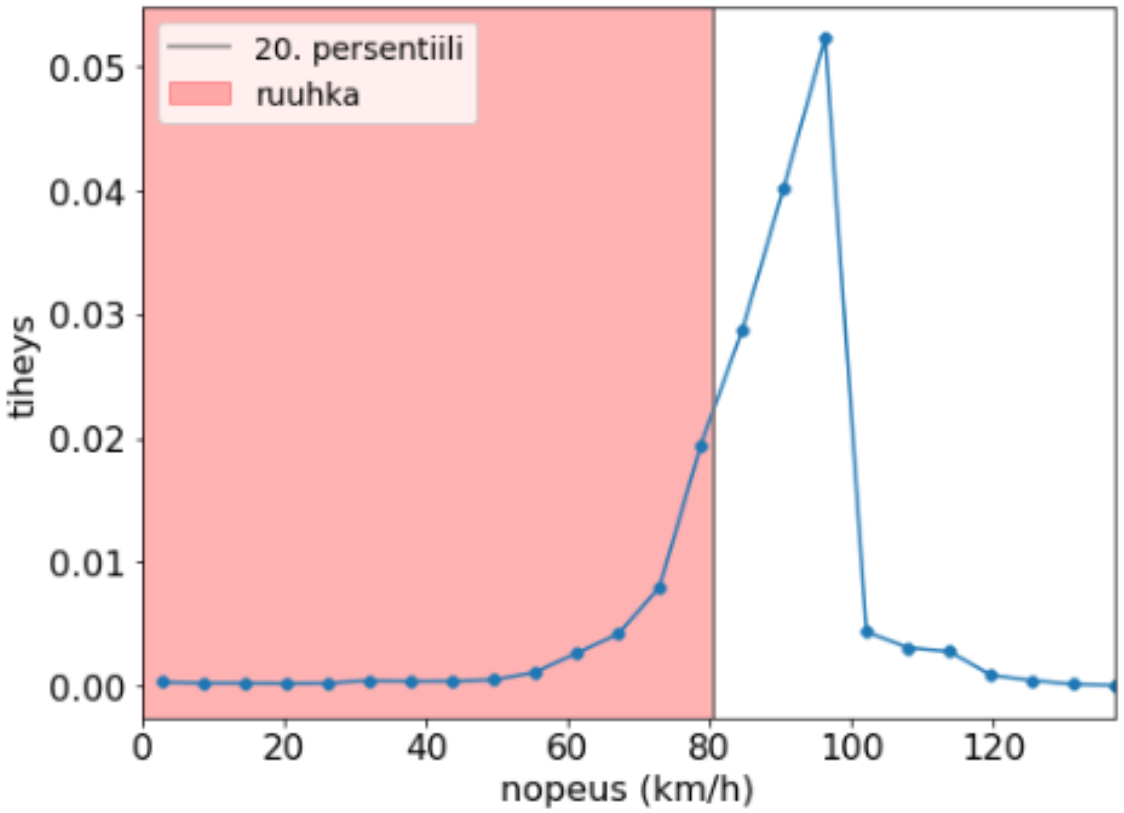
RoadCloud
WAY AHEAD

Työkalut ja aineisto

- Sensoriajoneuvoilla kerätty liikennedatasetti
  - Noin 11 miljoonaa havaintoa vuosilta 2019-2020
  - Tarkastelu rajattu Göteborgin lentokentän läheiselle moottoritielle
- Python-kirjastot (NumPy, pandas, Matplotlib, OSMnx)
- Koneoppimiseen scikit-learn
  - Pythonin avoimen lähdekoodin koneoppimiskirjasto

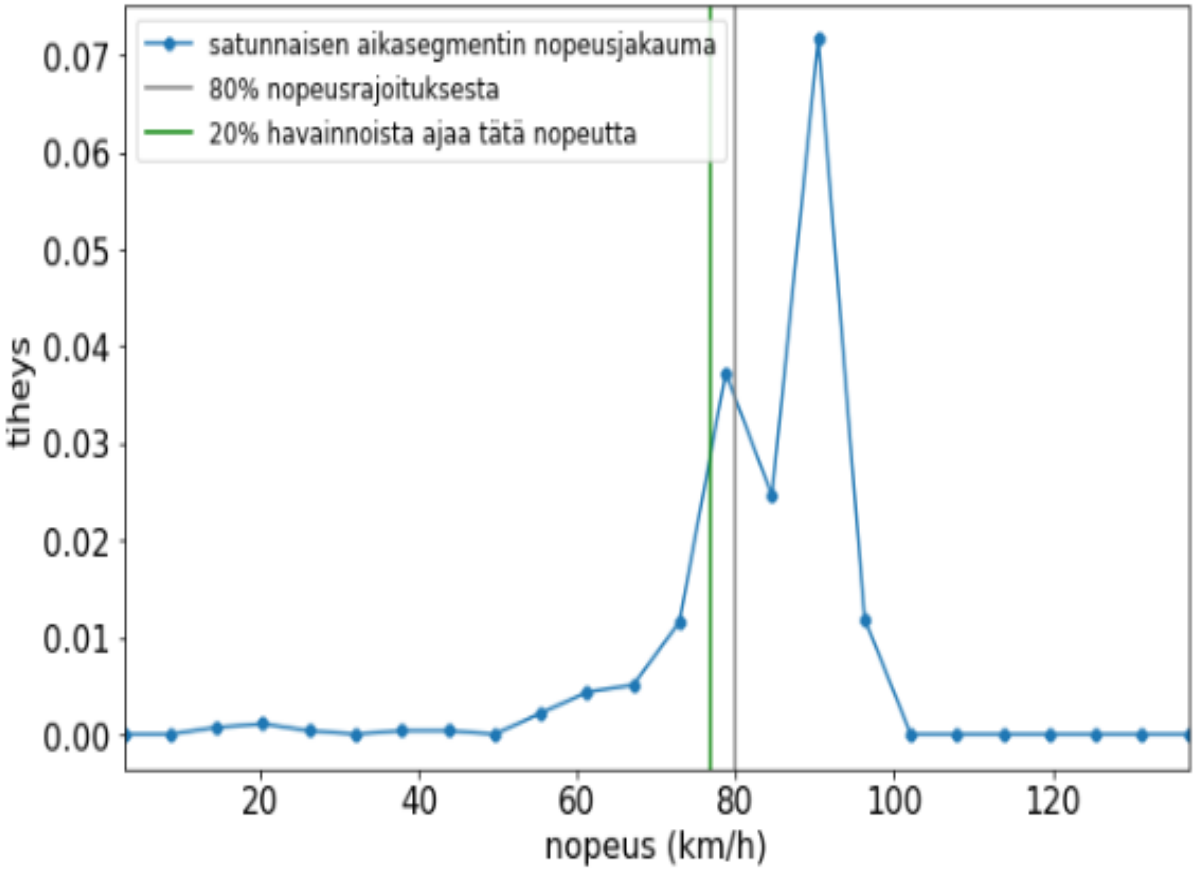
Ruuhkan määritelmät 1/2

- Ensimmäisen määritelmän mukaan liikenne on poikkeavaa, jos aikasegmentin havaintojen keskinopeus on pienempi kuin n:s persentiili koko tiesegmentin kaikista nopeuksista



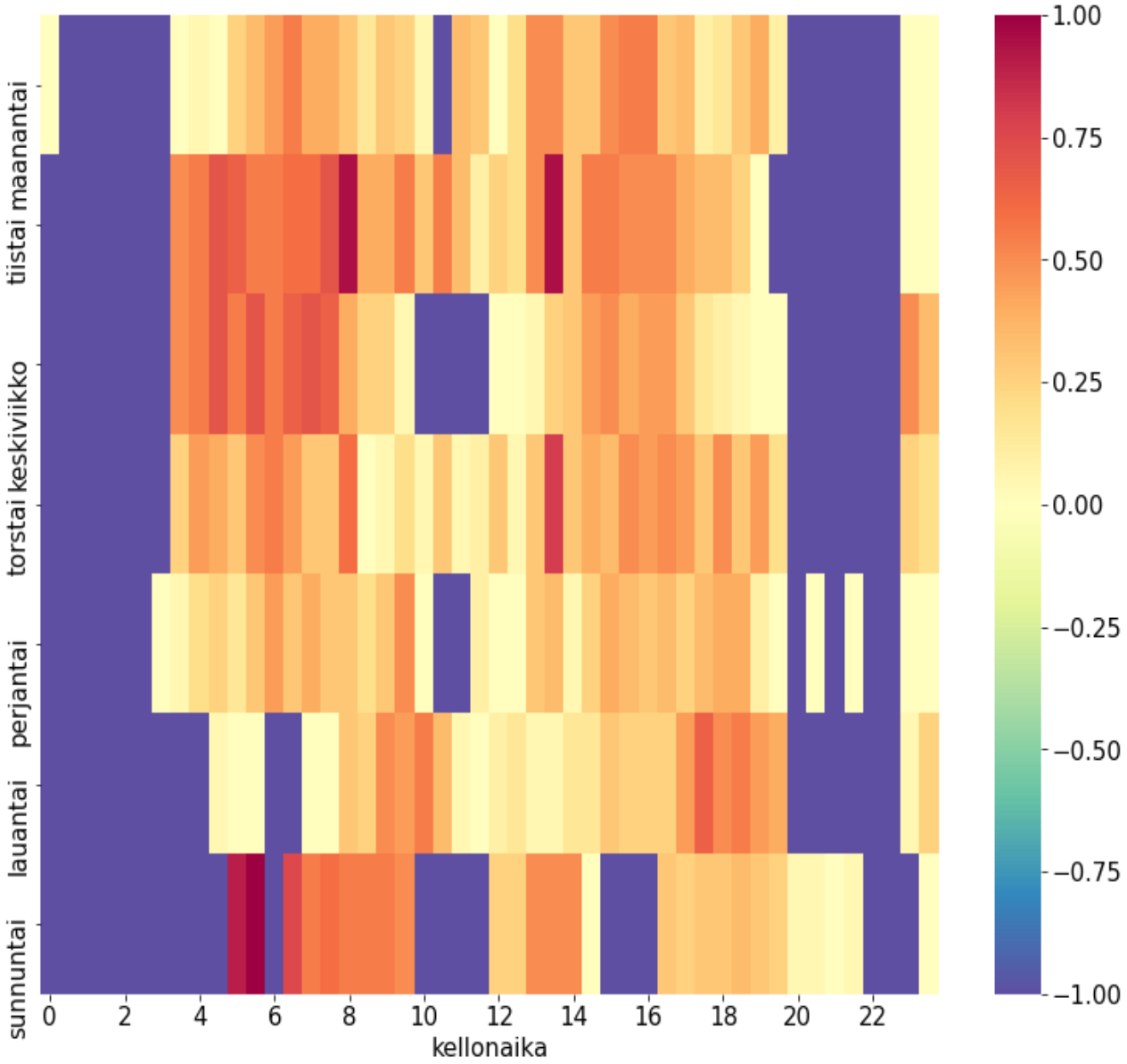
Ruuhkan määritelmät 2/2

- Toisen määritelmän mukaan liikenne on ruuhkaista, jos $p\%$ aikasegmentin havainnoista ajavat alle nopeutta x kertaa nopeusrajoitus



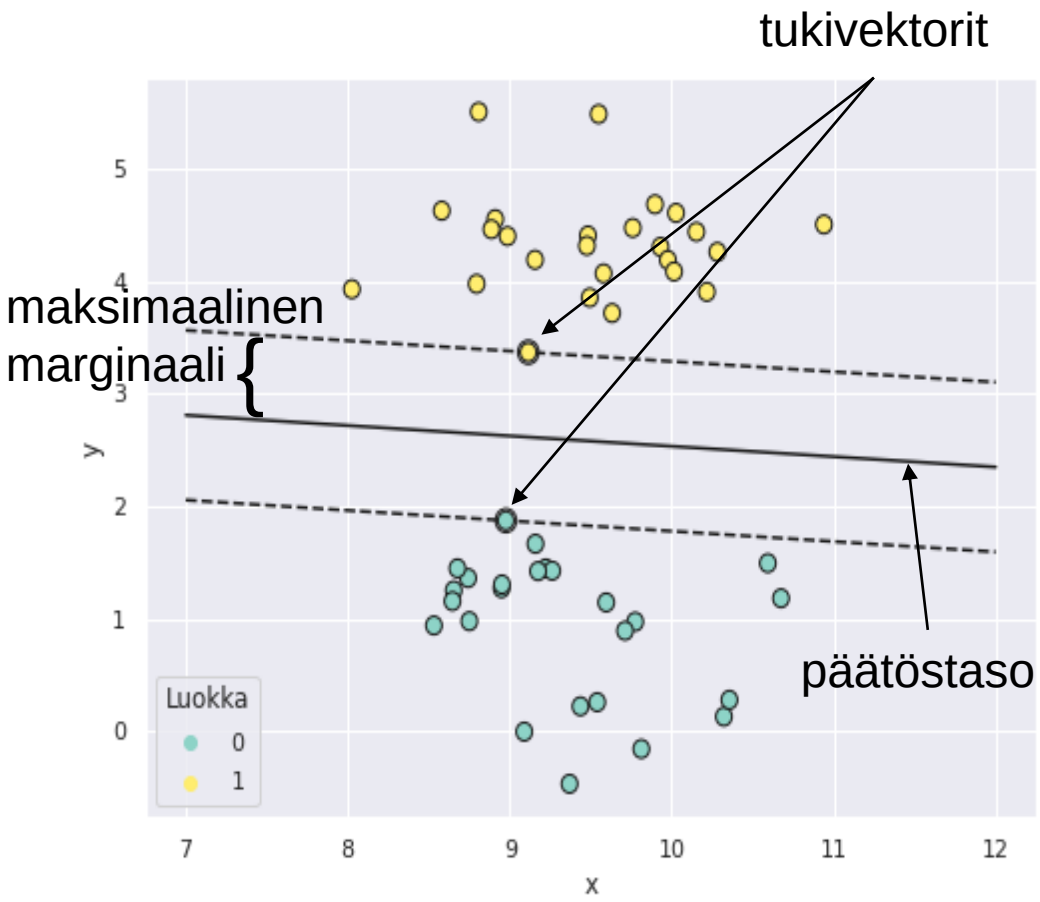
Esimerkki-viikko

- Aineisto jaettiin 15 minuutin aikasegmentteihin
- 1 = ruuhka
- 0 = ei ruuhkaa
- -1 = ei dataa



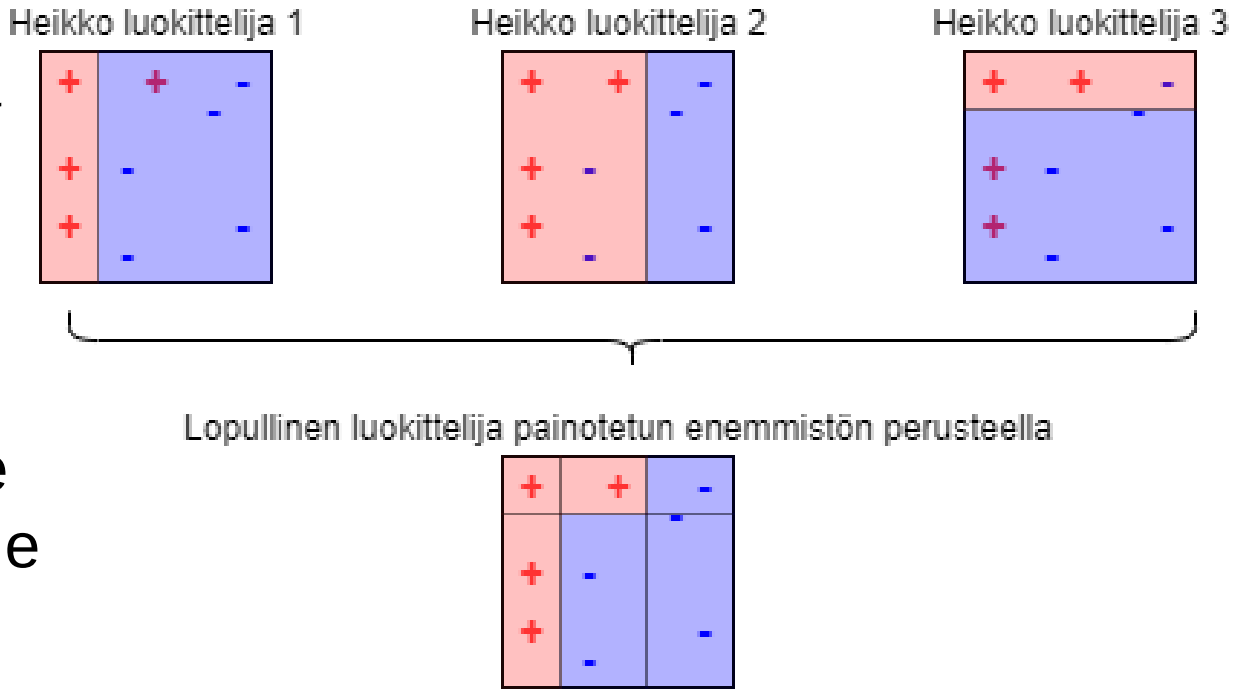
Tukivektorikone (support vector machine, SVM)

- Luokittelija muodostaa päätöstason ja marginaalit, joiden etäisyys tasoon maksimoidaan
- Aineisto ei separoidu täydellisesti → joustavat marginaalit
- Lineaarisesti separoitumattomalle datalle epälineaarinen ydinfunktio (kernel function)



AdaBoost eli mukautuva tehostamisalgoritmi

- Tehostaminen: heikoista oppijoista voidaan koostaa vahva oppija
- Toistetaan t iteraatiota ja annetaan jokaiselle heikolle luokittelijalle painokerroin → lopullinen malli



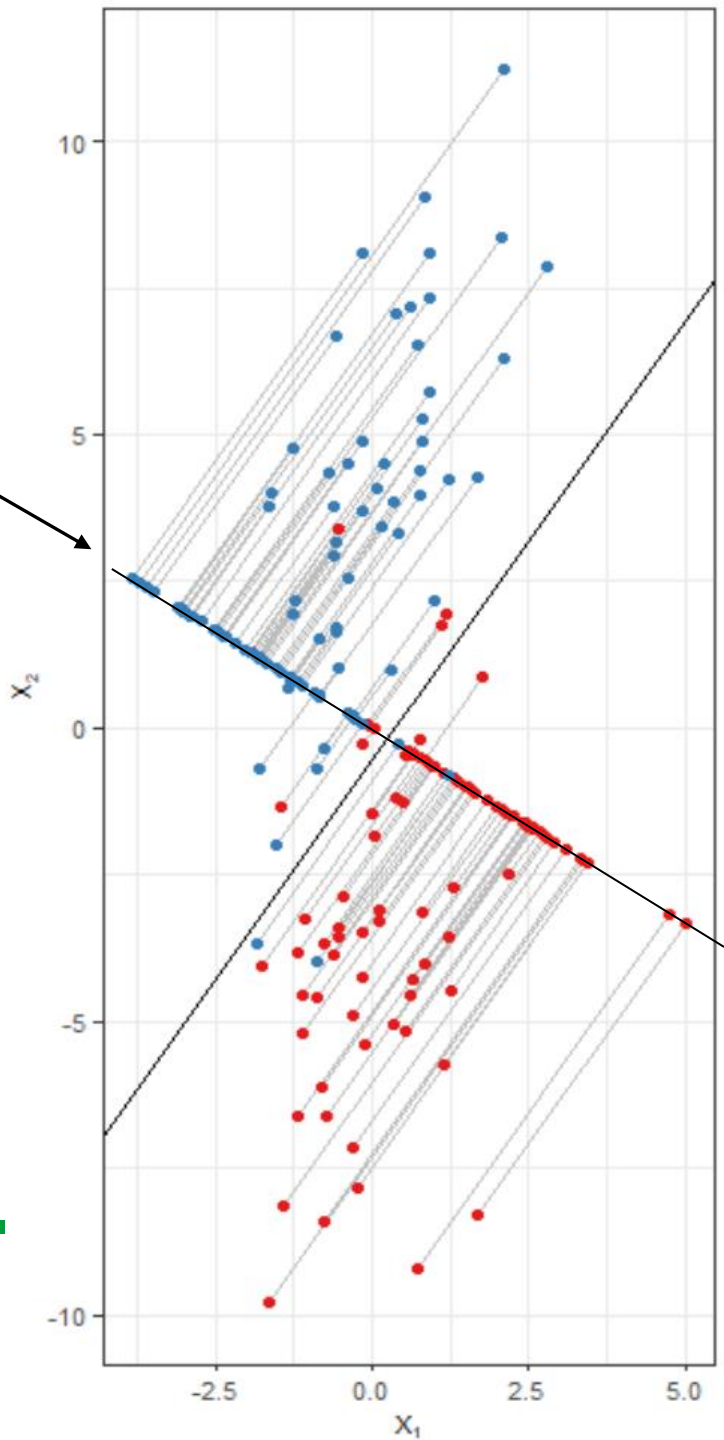
Lineaarinen erotteluanalyysi (LDA)

- Fisherin lineaarinen erottelufunktio

$$z = \boldsymbol{w}^T \boldsymbol{x},$$

josta etsitään lineaarinen painokerroinvektori $\boldsymbol{w}$

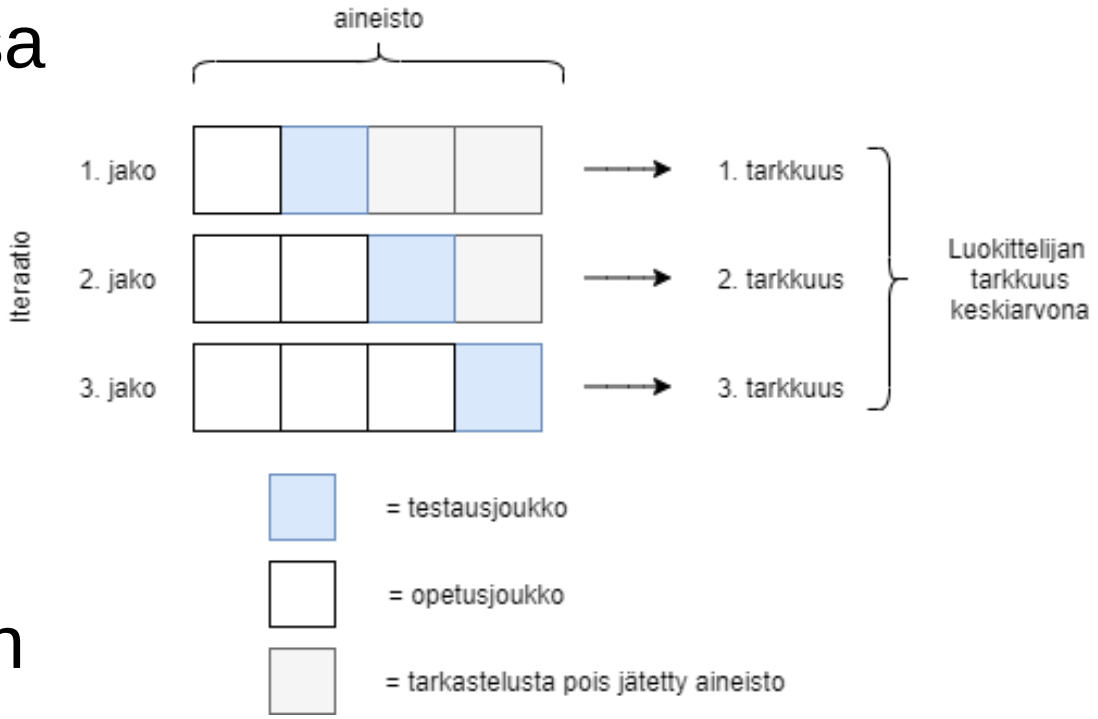
- Tavoitteena erotella datavektorin $\boldsymbol{x}$ pisteet
  - Luokkien keskipisteiden etäisyys maksimoituu
  - Luokkien sisäinen varianssi minimoituu



Tulokset

- Ennustusaineistossa huomioitiin 2 edeltävää aika-askelta (tien kunto, lämpötila, ruuhkaisuus, aika)
- Vain arkipäivät
- Tiesegmentti jaettiin itään ja länteen ajaviin

Ristiinvalidointi:



Ristiinvalidoinnin tulokset

Itä (suuremman luokan osuus)	SVM (%)	AdaBoost (%)	LDA (%)
n=20 (78%)	70,1±2,2	77,7±2,8	77,0±3,7
n=30 (62%)	59,4±2,0	62,3±5,5	63,2±2,8
n=40 (57%)	57,7±1,5	58,7±2,6	59,2±2,5
p=20,x=80 (58%)	59,6±1,2	61,9±2,5	60,6±2,4
p=30,x=80 (71%)	64,9±2,4	70,6±4,3	70,5±3,6
p=30,x=90 (72%)	65,3±3,5	69,7±2,5	71,8±3,8
Länsi (suuremman luokan osuus)	**SVM (%)**	**AdaBoost (%)**	**LDA (%)**
n=20 (74%)	72,5±3,5	77,2±3,3	74,9±4,1
n=30 (59%)	61,2±2,4	63,9±3,7	62,6±3,7
n=40 (60%)	56,3±6,9	63,1±3,1	59,7±5,5
p=20,x=80 (55%)	59,4±2,0	62,8±2,8	60,3±2, 9
p=30,x=80 (67%)	65,9±2,5	70,7±3,6	68,5±4,6
p=30,x=90 (77%)	59,1±1,7	76,2±3,2	76,6±2,6

Tulokset: koko aikasegmentti

- Esitettynä paras ennuste, joka saatiin koko 15 minuutin aikasegmenteille (länsi) parametreillä $p = 20\%$ ja $x = 0{,}8$
- Suuremman luokan osuus aineistosta oli 55%

AdaBoost

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**78**%	22%
Ruuhkaa	44%	**56**%

Lineaarinen erotteluanalyysi

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**78**%	22%
Ruuhkaa	55%	**45**%

Tukivektorikone

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**69**%	31%
Ruuhkaa	47%	**53%**

Tulokset: yksittäiset ajoneuvot

- Aineistosta poimittiin yksittäiset ajoneuvot, joiden ennusteita verrattiin koko segmentin ennusteisiin
- Esitettynä paras ennuste, joka saatiin yksittäisille ajoneuvoille (länsi) ensimmäisen ruuhkan määritelmän parametrilla $n = 30$
- Suuremman luokan osuus 62%

AdaBoost

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**85**%	15%
Ruuhkaa	56%	**44**%

Lineaarinen erotteluanalyysi

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**91**%	9%
Ruuhkaa	85%	**15**%

Tukivektorikone

Todellinen	Ennuste	
	Ei ruuhkaa	Ruuhkaa
Ei ruuhkaa	**67**%	32%
Ruuhkaa	55%	**45**%

Yhteenveto

- Tukivektorikone pärjäsi heikoiten, lineaarinen erotteluanalyysi ja AdaBoost lähes yhtä hyviä
- Yksittäisten autojen perusteella ennustustarkkuudet hieman matalampia
- Ei saavutettu yhtä hyviä ennusteita kuin sensoreilla kerätyllä liikennedatalla on mahdollista (de Fabritiis et al. (2008))
- Datan puutteellisuus vaikeuttaa liikenteen tilan tunnistusta

Tärkeimmät viitteet

- Corinna Cortes ja Vladimir Vapnik. Support-vector networks. Machine learning,20(3):273–297, 1995.
- Yoav Freund, Robert Schapire, ja Naoki Abe. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780):1612, 1999.
- Christopher M. Bishop.Pattern Recognition and Machine Learning. Information science and statistics. Springer, 1st ed. 2006. corr. 2nd printing edition, 2006. ISBN 9780387310732,0387310738
- Ruey Long Cheu, Chi Xie, ja Der-Horng Lee. Probe vehicle population and sample size for arterial speed estimation. Computer-Aided Civil and Infra-structure Engineering, 17(1):53–60, 2002. doi: 10.1111/1467-8667.00252
- Corrado de Fabritiis, Roberto Ragona, ja Gaetano Valenti. Traffic estimation andprediction based on real time floating car data. Teoksessa2008 11th InternationalIEEE Conference on Intelligent Transportation Systems, page 197–203, Oct 2008.