TAITOTALO

Lineaarinen regressio

10.10.2023

Kerttuli Ratilainen

Machine learning explained in 100 seconds

- https://www.youtube.com/watch?v=PeMlggyqz0Y

Tavoitteet

- Oppia tuntemaan lineaarisen regression periaatteet perustasolla
- Soveltaa periaatteita käyttämällä ohjatun koneoppimisen mallia lineaarinen regressio

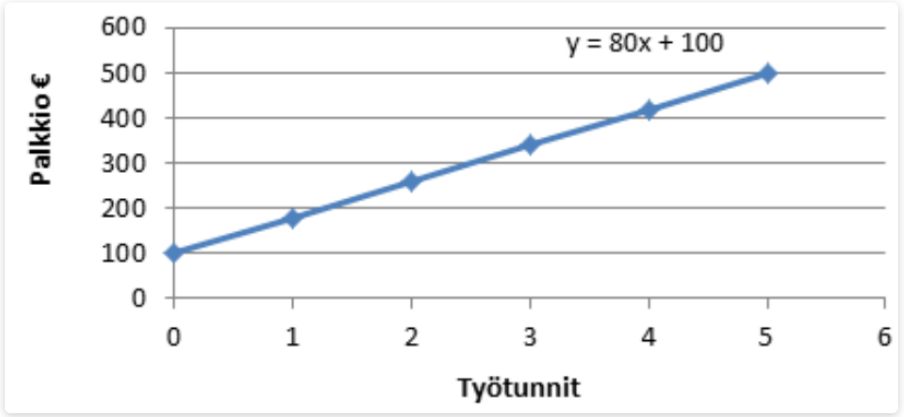
Mitä on lineaarinen regressio?

- Ennustetaan numeerisia arvoja
- Tutkitaan yhden tai useamman selittävän tekijän vaikutusta vastemuuttujaan
    - esimerkiksi mikä vaikuttaa asunnon hintaan
        - Sijainti
        - Pinta-ala
        - Huoneiden määrä

Esimerkki lineaarisesta regressiomallista

Esimerkki. Oletetaan, että konsultti perii palkkiota paikalle saapumisesta 100 euroa ja jokaiselta tehdyltä työtunnilta 80 euroa. Tällöin voin mallintaa konsultin kokonaispalkkiota suoralla y=80x+100, missä x on työtuntien määrä. Kyseisessä suoran yhtälössä

- vakiotermi 100 ilmoittaa y:n arvon, kun x=0 (eli esimerkissämme palkkio ilman varsinaisia työtunteja)
- kulmakerroin 80 ilmoittaa palkkion muutoksen, kun työtunnit lisääntyvät yhdellä.



Lähde: https://tilastoapu.wordpress.com/lineaarinen/

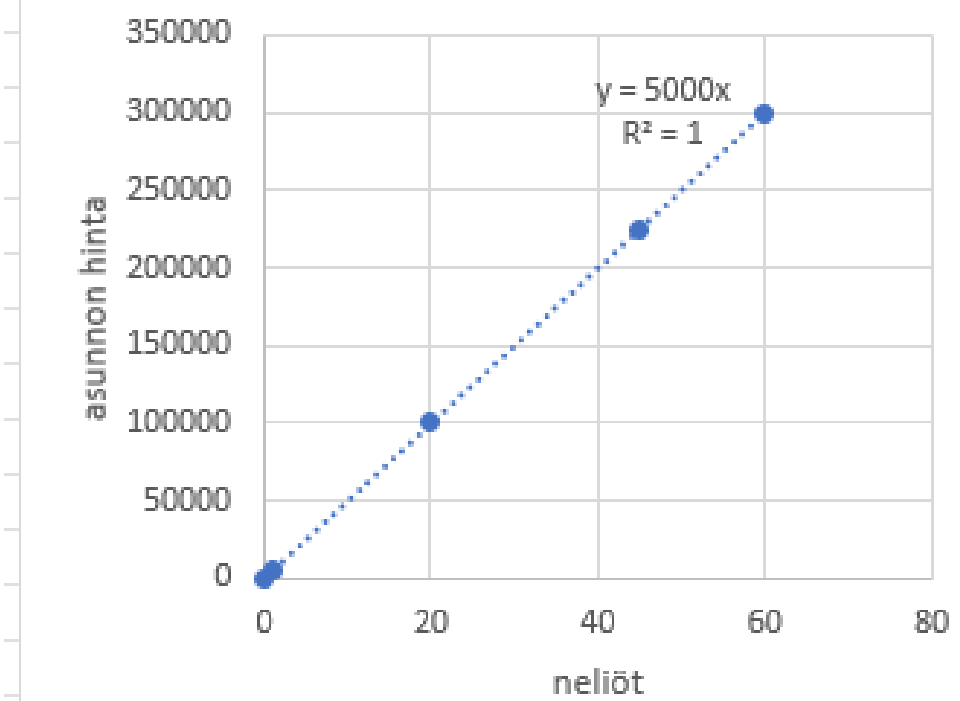
IoT TAITOTALO

Regressiosuora y = ax + b

- Osoittaa muuttujien välisen yhteyden
- Nousee ylöspäin -> positiivinen yhteys
- Laskee alaspäin -> negatiivinen yhteys
  - y on vastemuuttujan arvo -> asunnon hinta
  - a on regressiokerroin (regression coefficient) tai regressiosuoran kulmakerroin
    - Jos negatiivinen, laskeva
    - Jos positiivinen, nouseva
    - Jos kerroin on 0, ei muuttujien välillä ole lineaarista yhteyttä
  - x on selittävän muuttujan arvo -> esim. asunnon pinta-ala
  - b on vakiotermi, y-akselin leikkauspiste (intercept)

  - Esimerkiksi asunnon hinta Helsingissä, jos
    - neliöhinta 5000
    - neliöiden määrä 60 ja
    - lähtöhinta asuntojen hinnoille 0 euroa

      300 000 = 5000 * 60 + 0



Regressiomallin hyvyys

- Regressiomalli kuvaa ilmiötä sitä paremmin, mitä lähempänä pisteet ovat suoraa, toisiin sanoen muuttujien välinen riippuvuus on lineaarinen
- Jos pisteet ovat erityisen lähellä suoraa, on mallin ennustettavuus erittäin hyvä.

Residuaalit (jäännösarvot)

- Kunkin vastemuuttujan arvon etäisyys voidaan laskea regressiosuoran ennustamasta arvosta.
- Mitä suuremmat residuaalit mallissa on, sitä huonompi ennustettavuus mallilla on
- Residuaalien tulee olla riippumattomia toisistaan
- Residuaalit oletetaan olevan normaalisti jakautuneita
- Residuaalien keskiarvon tulee olla nolla
- Residuaalien varianssi on nolla, jos mallin avulla voidaan selittää kaikki havainnot täydellisesti eli mallille ei jää selittämättömiä eroja.
- Residuaalien kuvaajassa ei saisi näkyä säännönmukaisuutta, pisteiden tulisi sijaita satunnaisesti, ikään kuin ne olisi ”ammuttu haulikolla”

Mean squared error (MSE) - keskineliövirhe, (keskimääräinen neliövirhe, keskineliöpoikkeama)

- MSE:n avulla voidaan selvittää virheiden keskimääräinen suuruus ennusteiden ja todellisten arvojen välillä
- Mittaa ennusteen tarkkuutta verrattuna todellisiin arvoihin

- missä:
  - *n* on havaintojen lukumäärä,
  - *Yi* on todellinen arvo,
  - ^*Yi* ennustama arvo.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Palauttaa keskiarvon neliövirheistä
- Herkkä poikkeaville havainnoille (outliers)
- Auttaa arvioimaan miten hyvin malli suoriutuu

MSE-esimerkki

Oletetaan, että meillä on viisi havaintoa ja niiden todelliset ja ennustetut arvot ovat seuraavat:

$$Y = [3, 6, 8, 12, 15]$$
$$\hat{Y} = [2, 7, 8, 11, 14]$$

Lasketaan MSE:

$$MSE = \frac{1}{5}\sum_{i=1}^{5}(Y_i - \hat{Y}_i)^2$$

$$MSE = \frac{1}{5}[(3-2)^2 + (6-7)^2 + (8-8)^2 + (12-11)^2 + (15-14)^2]$$

$$MSE = \frac{1}{5}[1 + 1 + 0 + 1 + 1] = \frac{4}{5}$$

- Keskineliövirhe (MSE) on 4/5 tai 0,8.
- Mitä pienempi luku, sitä parempi ennustetarkkuus mallilla on
- Jos MSE on nolla, selittävät muuttujat ennustavat vastemuuttujan ilman virheitä
- MSE:llä ei ole ylärajaa

R2 (R-squared)

- Kertoo kuinka suuren osan selittävät muuttujat pystyvät selittämään vastemuuttujan vaihtelua
- Arvo aina 0-1
- Jos lähellä nollaa, selittävät muuttujat pystyvät selittämään vain vähän vastemuuttujaa
- Jos lähellä yhtä, muuttujat osuvat regressiosuoran lähelle
- R2 lukua käytetään, kun halutaan verrata kahden regressioanalyysin tuloksia keskenään.
- Uusien muuttujien lisääminen nostaa aina r2-lukua.

R2 kaava

- SSres on residuaalien neliön summa (havaittujen ja ennustettujen arvojen erotus)
- SStot neliöiden summa (havaittujen pisteiden etäisyys keskiarvo y-:stä).

- R2 voi laskea myös laskemalla korrelaatio potenssiin 2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Korrelaatiokerroin

- Välillä (-1) - (+1)
- Hajontakaavion kaikki pisteet sijaitsevat samalla nousevalla suoralla, puhutaan vahvasta positiivisesta korrelaatiosta
- Hajontakaavion kaikki pisteet sijaitsevat samalla laskevalla suoralla, puhutaan vahvasta negatiivisesta korrelaatiosta
- 0 tarkoittaa, että muuttujien välillä ei ole lainkaan korrelaatiota

Harjoitus 1 Regressiosuora vaihe 1.

- Avaa Excel Lineaarinen_regressio_tehtävä.xlsx
- Avaa Geogebra

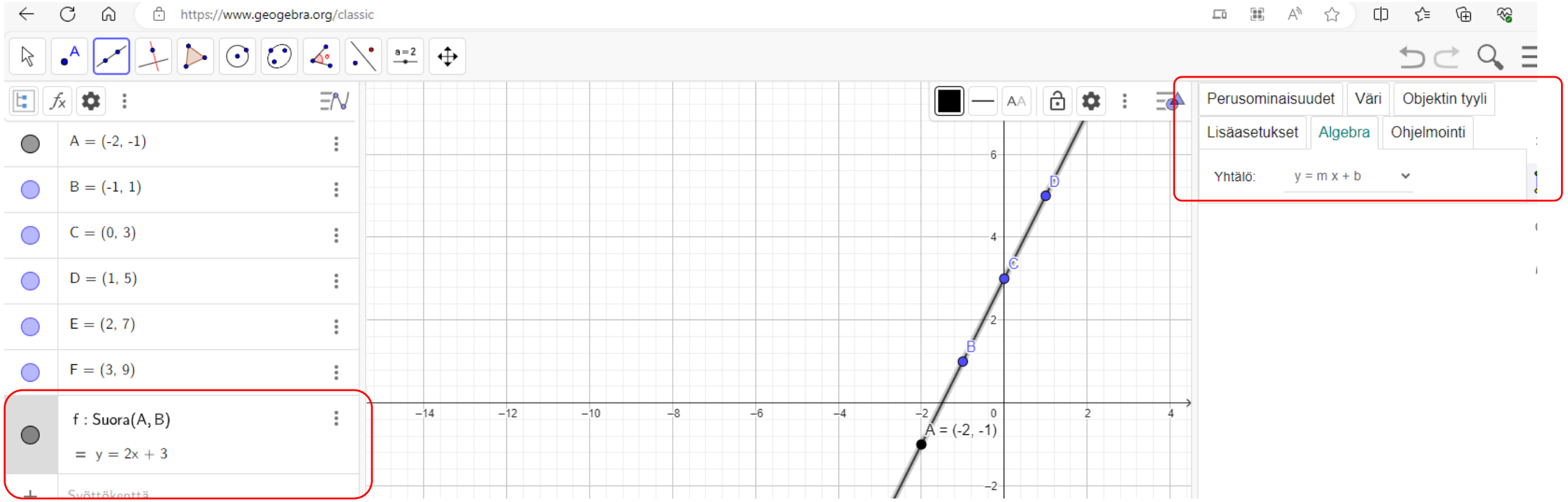
  https://www.geogebra.org/classic
- Piirrä pisteet koordinaatistoon
- Käytä Geogebraa, ruutupaperia tai Exceliä
- Piirrä pisteiden kautta kulkeva suora

Vaihe 2.

- Laske tai päättele pisteiden kautta kulkevan suoran
  - Kulmakerroin
  - Leikkauspiste
  - Korrelaatio

- Selvitä mikä on suoran selityskerroin r2 (excelissä pearsonin neliö)
- Laske ennuste satunnaiselle pisteelle, esimerkiksi x= 0,5 (excelissä ennuste)
- Kirjoita suoran yhtälö yhden desimaalin tarkkuudella

Esimerkki Geogebra



Esimerkki Excel

piste	x	y
A	-2	-1
B	-1	1
C	0	3
D	1	5
E	2	7
F	3	9

Laske pisteiden kautta kulkevan suoran

kulmakerroin	
leikkauspiste	
korrelaatio	
ennuste pisteella 0,5	
selityskerroin r2	

IOT AIT TAITOTALO

Vinkki Geogebran käyttöön

- Jos valitset vasemman reunan valikosta sen suoran, jonka olet piirtänyt, ja klikkaat kolmesta pisteestä Settings, oikeaan reunaan ilmestyy asetusvalikko. Sieltä välilehdeltä Algebra voit vaihtaa suoran esitystavan siihen muotoon, joka on (ainakin suomalaisille koululaisille) tutumpi, eli y = m*x + b. Tuossa kaavassa m on kulmakerroin, joka suomalaisissa koulukirjoissa on yleensä k, mutta Geogebrassa jostain syystä m. Ja b on vakiotermi eli leikkauspiste.
- Yläreunan valikosta täytyy klikata ensin pistettä A, jotta pääsee lisäämään koordinaatistoon pisteitä. Sitten klikataan siitä vierestä seuraavaa kuvaketta, jotta pääsee piirtämään kahden pisteen kautta kulkevan suoran.
- Kannattaa kokeilla sovittaa suoraa Geogrebrassa. Excel laskee sen valmiiksi.

Lineaarisuus koneoppimisessa

- Malliin otetaan mukaan useampi muuttuja
- Tarpeettomat muuttujat karsitaan, esimerkiksi jos:
  - Korreloivat vahvasti keskenään
  - Eivät selitä riittävästi vastemuuttujaa
  - Muuttujien korrelaatiota voidaan selvittää esimerkiksi scipy.stats-moduulin metodien avulla laskemalla p-arvo muuttujille
- Muuttujat yleensä välimatka-asteikoilla
- Luokittelu ja järjestysasteikkomuuttujat voidaan ottaa huomioon muuttamalla ne dummy-muuttujiksi

Muuttujat

- Yleensä välimatka-asteikoilla
- Luokittelu ja järjestysasteikkomuuttujat voidaan ottaa huomioon muutamalla ne dummy-muuttujiksi

	omena	banaani	appelsiini
omena	1	0	0
banaani	0	1	0
appelsiini	0	0	1
ei mikään edellisistä	0	0	0

```python
import pandas as pd

# Esimerkkidata
data = {'kaupunki': ['Helsinki', 'Tampere', 'Oulu', 'Helsinki', 'Oulu']}
df = pd.DataFrame(data)

# Muunnetaan kategorinen muuttuja dummy-muuttujaksi
df_dummies = pd.get_dummies(df, columns=['kaupunki']).astype('int')
print(df_dummies)
```

0.0s Open 'df_dummies' in Data Wrangler

```
   kaupunki_Helsinki  kaupunki_Oulu  kaupunki_Tampere
0                  1              0                 0
1                  0              0                 1
2                  0              1                 0
3                  1              0                 0
4                  0              1                 0
```

IOT AIT TAITOTALO

Usean muuttujan lineaarinen regressio

- Yhden muuttujan suhde useisiin selittäviin muuttujiin
- $y_1 = b_0 + b_1\, x_{i1} + b_2\, x_2 + \ldots + b_n\, x_n + \varepsilon$
- Y on vastemuuttuja
- b_0 on vakio (leikkauspiste)
- $b_1\ b_{2,} \ldots,\ b_n$ on selittävien muuttujien $x_1\ x_{2,} \ldots,\ x_n$ kertoimet
- E edustaa niitä havaitsemattomia tekijöitä, jotka vaikuttavat y:hyn, mutta eivät sisälly malliin
- Muuttujien suhde ei enää ole kaksiulotteinen vaan moniulotteinen

StandardScaler

- sklearn.preprocessing -moduuli
- Skaalaa tietoaineiston niin, että jokaisen muuttujan (feature) keskiarvo on 0 ja keskihajonta on 1
- Aineiston standardisointi auttaa monissa koneoppimisen algoritmeissa pitämään muuttujat vertailukelpoisina, koska suurempi vaihteluväli voi dominoida ja vääristää mallin oppimista.

StandardScaler muuntaa datan jokaisen muuttujan X arvot seuraavasti:

Tässä:
- mean(X) on ominaisuuden keskiarvo
- std(X) on ominaisuuden keskihajonta

$$X_{\text{scaled}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

Esimerkki 2

- Stack loss esimerkki
- Jupyter notebook

Harjoitus 3 wine

- Valitse jompikumpi viinitiedostoista
- Poista quality-sarake
- Valitse alkoholipitoisuus vastemuuttujaksi, muut selittäviä
- Jaa data opetus- ja testijoukkoon
- Sovita data lineaarisen regression malliin
- Tulosta mittarit ja tarvittavat kuvaajat
- Tutki voiko datan avulla selvittää mitkä tekijät vaikuttavat viinin alkoholipitoisuuteen ja jos voi niin, mitkä ovat tärkeimmät tekijät
- Keskustele havainnoista ryhmässä, hakeudu uuteen ryhmään
- Palauta työ ryhmätyönä teams-kansioon

Harjoitus 4 NBA

- Ota tarkasteluun nba:n tilastot ja kausi 22-23.
- https://www.nba.com/stats/teams/traditional?dir=A&sort=FG3A
- Ensimmäinen tehtävä on saattaa aineisto pandas-kirjastolla luettavaan muotoon. Löydätkö siihen python-koodin vai käytätkö esimerkiksi Power Query-editoria?
- Mitkä tekijät selittävät voittoa analyysin perusteella? Käytä testiaineistona kautta 23-24.
- Muista siivota aineisto ensin esimerkiksi poista turhat tai liikaa korreloivat sarakkeet, esimerkiksi prosentti-sarakkeet.
- Tarkista, että kaikki luvut ovat liukulukuja.
- Lyhenteiden selitykset löytyvät, kun viet hiiren lyhenteen päälle.
- Bonus: mieti suositusta valmentajalle ☺, mihin pelissä kannattaa keskittyä.

Lisätietoa:

https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvanti/regressio/analyysi/