**Harjoitukset**

Koneoppiminen, Syksy 2020

Tapani Alastalo

M1475

**Sisältö**

[1 Ohjeet 2](#_Toc55473345)

[2 Tehtävä 1 Lineaarinen regressio (videoesimerkki) 2](#_Toc55473346)

[3 Tehtävä 2. Tee edellisen tehtävän tilanteessa 3](#_Toc55473347)

[4 Tehtävä 3 3](#_Toc55473348)

[5 Tehtävät 4 Neuroverkot 3](#_Toc55473349)

[6 Tehtävä 5 (videoesimerkki) 4](#_Toc55473350)

[7 Tehtävä 6 4](#_Toc55473351)

[8 Tehtävä 7 (videoesimerkki) 5](#_Toc55473352)

[9 Ss 5](#_Toc55473353)

[10 Ss 5](#_Toc55473354)

[11 Ne kosteusarvot, jotka ovat välillä 90-92 (raja-arvot mukaanlukien). Ota listaukseen mukaan myös sensorin tunnus ja mittausajankohta 5](#_Toc55473355)

# Ohjeet

Kirjoita vastauksesi kuhunkin tehtävään WORD-ohjelmalla, johon liität mukaan kuvankaappaukset piirtämistäsi kuvaajista. Kirjoita mukaan myös sanallinen tai numeerinen vastaus tehtävässä kysyttyihin kysymyksiin. Kokoa kaikkien tehtävien ratkaisut samaan WORD-tiedostoon. Tee samaan WORD-tiedostoon kaikkien ratkaisujen jälkeen myös kappale ”Lähdekoodit”, johon kopioit tehtävien ratkaisuissa käyttämäsi lähdekoodit. Koodia ei tarvitse kommentoida eikä siistiä. Muunna tiedosto lopuksi PDF-muotoon ja palauta se Optiman palautuskansioon. Palautuslaatikko sulkeutuu 20.5. klo 16:00. Kustakin tehtävästä saa 0-2 pistettä.

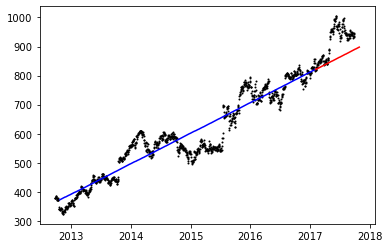
Kurssin arviointiasteikko:

* 36 p -> 5
* 32 p -> 4
* 28 p -> 3
* 24 p -> 2
* 16 p -> 1

Mikäli jokin tehtävä tuntuu vaikealta, jätä se aluksi väliin ja palaa siihen myöhemmin, mikäli motivaatiota ja energiaa riittää. Osa tehtävistä on videoesimerkkejä, jotka pystyt ratkaisemaan kopioimalla koodia videolta rivi riviltä. Videolla selitetään kunkin koodirivin sisältö. Sovella sitten oppimaasi tietoa muiden tehtävien ratkaisemiseen, joista ei ole videota. Kysy rohkeasti neuvoa tarvittaessa joko Teams-kanavalla tai sähköpostitse (tomi.nieminen@jamk.fi).

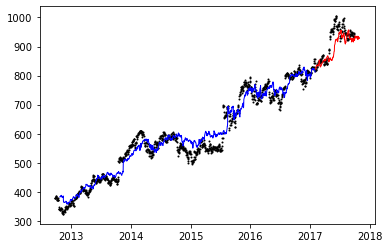
# Tehtävä 1 Lineaarinen regressio

a)



Ennusteen keskivirhe testidatassa on 70

b)



Ennusteen keskivirhe testidatassa on 35

## Lähdekoodit

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(30)

df\_test = df[:185]

df\_train = df[185:]

X = np.array(df\_train[['Time', 'Close']])

y = np.array(df\_train['CloseFuture'])

model = linear\_model.LinearRegression()

model.fit(X, y)

ennuste\_train = model.predict(X)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time', 'Close']])

ennuste\_test = model.predict(X\_test)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=1)

plt.plot((df\_train['Date'] + pd.DateOffset(days=30)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=30)).values, df\_test['Ennuste'].values, color='red')

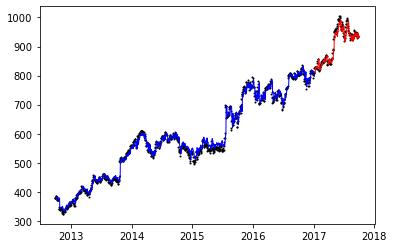
plt.show()

df\_validation = df\_test.dropna()

print("Ennusteen keskivirhe test datassa on %.f" % mean\_absolute\_error(df\_validation['CloseFuture'], df\_validation['Ennuste']))

# Tehtävä 2. Tee edellisen tehtävän tilanteessa

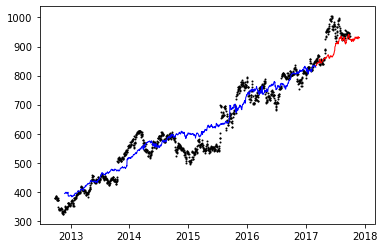
a) 7 päivän ennuste tulevaisuuteen



Ennusteen keskivirhe opetusdatassa on 16

Ennusteen keskivirhe testidatassa on 20

b) 60 päivän ennuste tulevaisuuteen



Ennusteen keskivirhe opetusdatassa on 34

Ennusteen keskivirhe testidatassa on 60

## Lähdekoodit

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

from sklearn.metrics import mean\_absolute\_error

days\_to\_forecast = 60

df = pd.read\_csv('data/Google\_Stock\_Price.csv')

df['Date'] = pd.to\_datetime(df['Date'])

df['Time'] = df.apply(lambda row: len(df) - row.name, axis=1)

df['CloseFuture'] = df['Close'].shift(days\_to\_forecast)

df\_test = df[:185]

df\_train = df[185:]

X = np.array(df\_train[['Time', 'Close']])

y = np.array(df\_train['CloseFuture'])

model = linear\_model.LinearRegression()

model.fit(X, y)

ennuste\_train = model.predict(X)

df\_train['Ennuste'] = ennuste\_train

X\_test = np.array(df\_test[['Time', 'Close']])

ennuste\_test = model.predict(X\_test)

df\_test['Ennuste'] = ennuste\_test

plt.scatter(df['Date'].values, df['Close'].values, color='black', s=1)

plt.plot((df\_train['Date'] + pd.DateOffset(days=days\_to\_forecast)).values, df\_train['Ennuste'].values, color='blue')

plt.plot((df\_test['Date'] + pd.DateOffset(days=days\_to\_forecast)).values, df\_test['Ennuste'].values, color='red')

plt.show()

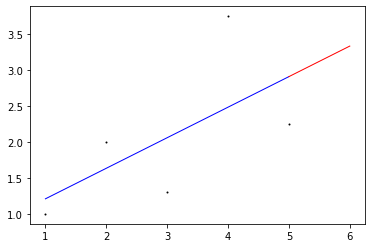
df\_train\_validation = df\_train.dropna()

df\_test\_validation = df\_test.dropna()

print("Ennusteen keskivirhe opetus datassa on %.f" % mean\_absolute\_error(df\_train\_validation['CloseFuture'], df\_train\_validation['Ennuste']))

print("Ennusteen keskivirhe test datassa on %.f" % mean\_absolute\_error(df\_test\_validation['CloseFuture'], df\_test\_validation['Ennuste']))

# Tehtävä 3



## Lähdekoodi

import pandas as pd

import numpy as np

from sklearn import linear\_model

import matplotlib.pyplot as plt

data = [[1.00,1.00],[2.00,2.00],[3.00,1.30],[4.00,3.75],[5.00,2.25],[6.00, None]]

df = pd.DataFrame(data, columns=['X', 'Y'])

df\_train = df[:5]

df\_test = df[4:]

X = np.array(df\_train['X'])

X = X.reshape(-1,1) # vain jos yksiulotteinen taulukko

y = np.array(df\_train['Y'])

model = linear\_model.LinearRegression()

model.fit(X, y)

df\_train['Ennuste'] = model.predict(X)

X\_test = np.array(df\_test['X'])

X\_test = X\_test.reshape(-1,1) # vain jos yksiulotteinen taulukko

df\_test['Ennuste'] = model.predict(X\_test)

plt.scatter(df['X'].values, df['Y'].values, color='black', s=1)

plt.plot(df\_train['X'].values, df\_train['Ennuste'].values, color='blue', linewidth=1)

plt.plot(df\_test['X'].values, df\_test['Ennuste'].values, color='red', linewidth=1)

plt.show()

# Tehtävät 4 Neuroverkot

Tee tehtävässä 1 olevan datan perusteella Googlen osakkeelle MLP-neuroverkkomalli, ja ennusta sen avulla osakekurssi 30 päivää tulevaisuuteen. Valitse input muuttujiksi aika sekä nykyhetken osakekurssi. Säästä vuoden 2017 tiedot test-dataksi, jonka perusteella arvioit mallin tarkkuutta. Esitä vastauksessa ennustekuvaaja sekä mallin keskivirhe erikseen training- ja test-datassa. Katso tehtävän johdannoksi neuroverkkojen teoriavideo: https://www.youtube.com/watch?v=APzICkVo2Q0

# Tehtävä 5 (videoesimerkki)

Muodosta edellisen tehtävän datasta neuroverkkomalli, jonka syötemuuttujana on pelkkä aika, ja jonka keskivirhe training datassa on alle 20 yksikköä eli selkeästi tarkempi kuin aiemmat mallit. Onko tämä malli tarkempi myös testidatassa, jota neuroverkko ei ole nähnyt mallin opetusvaiheessa? Esitä vastauksessa ennustekuvaaja sekä mallin keskivirhe erikseen training- ja test-datassa. (Vihje: kokeile muuttaa verkon hyperparametreja. Erityisesti aktivointifunktioksi kannattaa ehkä valita sigmoid tai tanh, jotta malli onnistuu paremmin epälineaaristen vaihteluiden kuvaamisessa.)

# Tehtävä 6

Oheisessa datassa (Kysynta.csv) on esitetty erään tuotteen kysyntä markkinointikampanjan aikana. Ennusta kysyntä ajanhetkellä 350 päivää käyttämällä

a) Lineaarista regressiomallia, jonka input-muuttujana on pelkkä aika.

b) MLP-neuroverkkoa, jonka input-muuttujana on pelkkä aika.

Lisää vähintään yhden piilotetun kerroksen aktivointifunktioksi tanh, jotta mallisi onnistuu paremmin epälineaarisen yhteyden kuvaamisessa. Kumpi malleista on mielestäsi luotettavampi ennuste? Liitä raporttiisi kuvaajat, joissa näkyy havaintopisteiden lisäksi mallisi ennusteet aikavälillä 0 – 350 päivää. Raportoi myös kummankin mallin keskivirhe training-datassa.

# Tehtävä 7 (videoesimerkki)

Liitteenä olevassa datassa (fruit\_data.csv) on 59 hedelmän mitattuja ominaisuuksia. Muodosta koneoppimismalli, joka ennustaa hedelmän nimen käyttäen input muuttujina massaa, leveyttä, korkeutta ja väripisteitä. Vertaile logistisen regressiomallin, SVM ja KNNmallien tarkkuutta. Raportoi vastaukseesi kunkin mallin osumatarkkuus, ja esitä myös pieni otos datariveistä, joissa näkyy kunkin hedelmän oikea tyyppi sekä kunkin mallin ennuste hedelmän tyypille.

# Ss

# Ss

# Tehtävä 20: Lähtevien asiakkaiden tunnistaminen

Oheisessa datassa (Telco.csv) on teleoperaattorin asiakastietokanta. Muuttuja churn=1 ilmaisee, että asiakas on lopettanut sopimuksen. Tehtäväsi on ennustaa kunkin asiakkaan ”churn”-kentän arvo käyttämällä input-muuttujina muiden kenttien arvoja. Jätä datasta 100 satunnaista asiakasta test-dataksi, ja valitse loput training-dataksi, jonka avulla muodostat haluamasi koneoppimismallin. Raportoi vastaukseesi seuraavat asiat:

• Käyttämäsi koneoppimismalli ja sen tarkkuus training- ja test-datassa.

• Huolehdi siitä, ettei malli ole liikaa ylisovitettu.

• Listaus 20 satunnaisesta test-datan asiakkaasta, jossa näkyy kunkin asiakkaan todellinen churn-arvo sekä mallisi ennuste (churn-riski numerona välillä 0-1)