Koneoppiminen – harjoitustehtävät

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).

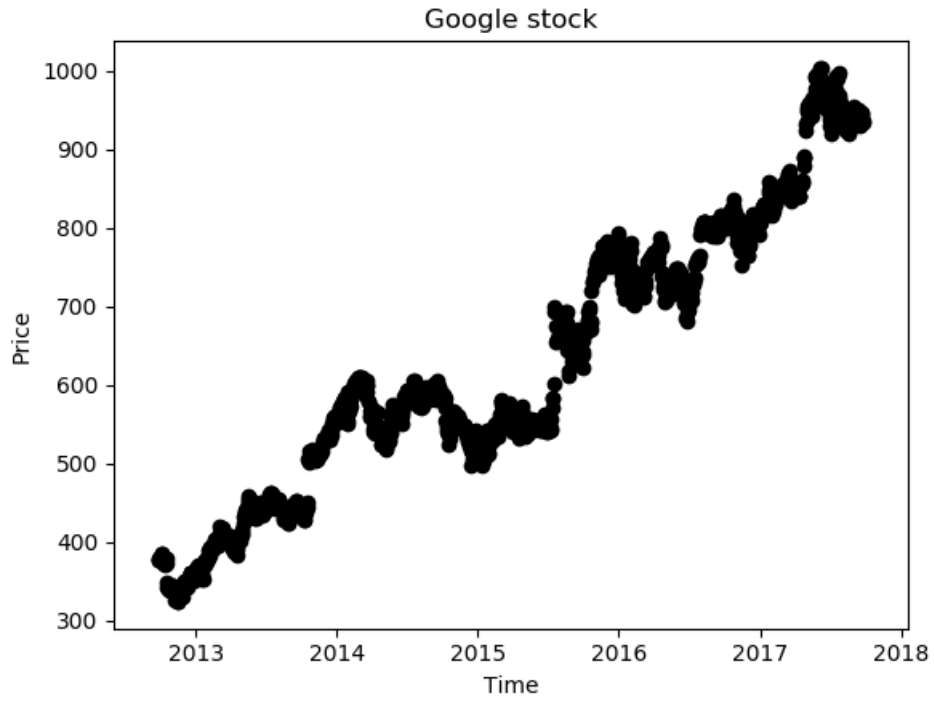
Antoisaa opiskelua tämän erittäin mielenkiintoisen aihepiirin parissa!

Lineaarinen regressio

Tehtävä 1 (videoesimerkki). Liitteenä olevassa datassa (Google_Stock_Price.csv) on Googlen osakkeen päivittäinen kurssi vuosina 2012 – 2017. Valitse training-dataksi vuosien 2012 – 2016 data ja test-dataksi vuoden 2017 data. Muodosta training datan perusteella lineaarinen regressiomalli, joka ennustaa kurssin 30 päivää eteenpäin. Laske sitten mallisi keskivirhe test-datassa. Valitse input-muuttujiksi

a) pelkkä aika.
b) aika sekä nykyhetken osakekurssi.

Piirrä ennusteesi selkeänä kuvaajana.



Tehtävä 2. Tee edellisen tehtävän tilanteessa

a) 7 päivän ennuste tulevaisuuteen
b) 60 päivän ennuste tulevaisuuteen

käyttämällä syötemuuttujina aikaa sekä nykyhetken osakekurssia. Esitä vastauksessa kuvaajat, joissa näkyy toteutuneet arvot sekä ennuste. Liitä vastaukseen mukaan myös mallien keskivirheet erikseen training- ja test-datassa.

Tehtävä 3. Tee lineaarinen regressiomalli oheisesta datasta. Valitse mallin input muuttujaksi x ja output muuttujaksi y. Ennusta muuttujan y arvo x:n arvolla 6. Liitä vastaukseesi kuvaaja, jossa näkyy sekä tunnetut arvot että mallisi ennuste.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Neuroverkot (regressio)

Tehtävä 4 (videoesimerkki). 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:

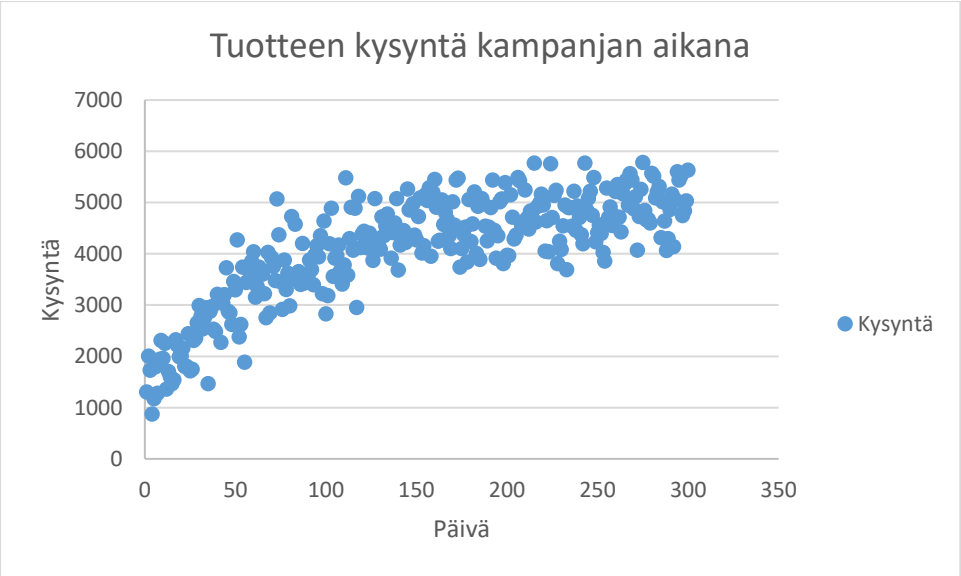
https://www.youtube.com/watch?v=APzICkVo2Q0

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.



Luokitteluongelma: Logistinen Regressio, Support Vector Machine, KNN

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 KNN-mallien tarkkuutta. Raportoi vastaukseesi kunkin mallin osumatarkkuus, ja esitä myös pieni otos datariveistä, joissa näkyy kunkin hedelmän oikea tyyppi sekä kunkin mallin ennuste hedelmän tyypille.

	fruit_label	fruit_name	fruit_subtype	mass	width	height	color_score
0	1	apple	granny_smith	192	8.4	7.3	0.55
1	1	apple	granny_smith	180	8.0	6.8	0.59
2	1	apple	granny_smith	176	7.4	7.2	0.60
3	2	mandarin	mandarin	86	6.2	4.7	0.80
4	2	mandarin	mandarin	84	6.0	4.6	0.79

Tehtävä 8 (videoesimerkki). Muodosta edellisen tehtävän datasta luokitteleva neuroverkko, joka ennustaa hedelmän nimen käyttäen input muuttujina massaa, leveyttä, korkeutta ja väripisteitä. Raportoi vastaukseesi mallin osumatarkkuus, ja esitä myös pieni otos datariveistä, joissa näkyy kunkin hedelmän oikea tyyppi sekä mallin ennuste hedelmän tyypille.

Tehtävä 9. Liitteenä olevassa datassa (Titanic.csv) on tietoja Titanicin matkustajista. Erityisesti tietokannassa näkyy, onko matkustaja selvinnyt hengissä (Survived 1) vai ei (Survived 0). Valitse matkustajista satunnaisesti 200 test-dataan, ja jätä loput training-dataan. Muodosta training-datasta 3 koneoppimismallia (Logistinen regressio, SVM ja KNN) jotka ennustavat matkustajien selviytymistä. Valitse input muuttujiksi Pclass, Sex, Age, SibSp, Parch ja Embarked. Raportoi kunkin mallin tarkkuus training- ja test-datassa. Liitä vastaukseesi myös ote parhaan mallisi ennusteesta listaamalla 20 satunnaista matkustajaa test-datasta siten, että listassa näkyy matkustajan numero (PassengerId), Survived-kentän arvo sekä mallisi ennusteen arvo.

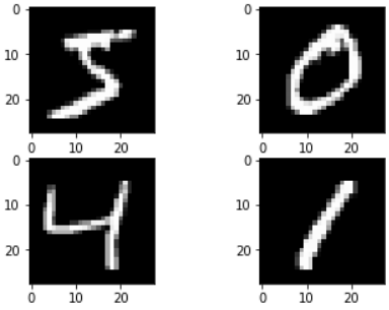
PassengerI	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, M	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr	male		0	0	330877	8.4583		Q
7	0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S

(Huom. Luokitteluasteikolliset input-muuttujat täytyy muuntaa ”one hot”-muotoon.)

Tehtävä 10. Muodosta edellisen tehtävän datasta neuroverkkomalli, jonka avulla ennustat matkustajien selviämistä. Valitse test-dataan 200 satunnaista matkustajaa ja jätä loput training-dataan. Valitse input muuttujiksi Pclass, Sex, Age, SibSp, Parch ja Embarked. Raportoi mallisi tarkkuus training- ja test-datassa. Liitä vastaukseesi myös ote mallisi ennusteesta listaamalla 20 satunnaista matkustajaa test-datassa siten, että listassa näkyy sekä Survived-kentän arvo että mallisi ennusteen arvo. Muista muuntaa kaikki luokitteluasteikolliset muuttujat ”one hot”-muotoon.

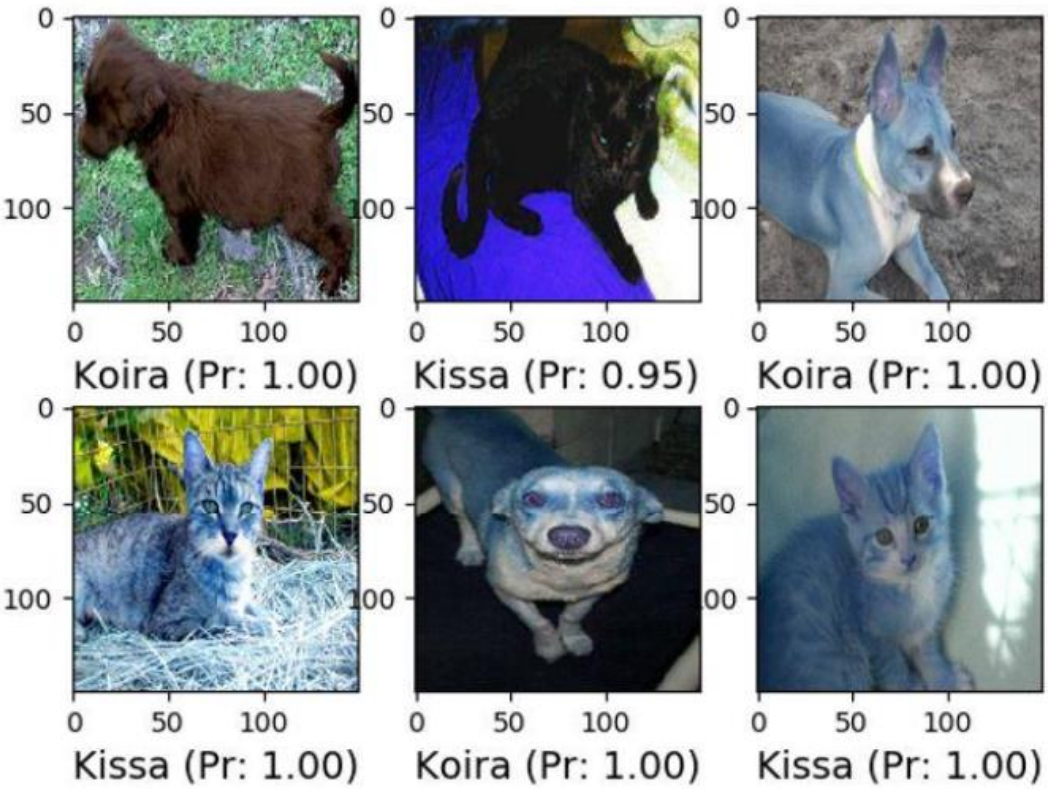
Kuvantunnistus ja konvoluutioneuroverkot

Tehtävä 11 (videoesimerkki). MNIST-tietokantaan on kerätty käsin kirjoitettujen numeroiden kuvia. Training-datassa on 60000 kuvaa ja test-datassa 10000 kuvaa. Rakenna MLP-neuroverkko, joka tunnistaa kuvassa olevan numeron. Raportoi vastaukseen mallisi keskivirhe training- ja test-datassa. Esitä myös muutama kuva numeroista, jotka mallisi tunnistaa väärin.



Tehtävä 12 (videoesimerkki). Ratkaise edellinen tehtävä *konvoluutioneuroverkon* avulla. Onko sen tarkkuus parempi kuin MLP-neuroverkon tarkkuus? Raportoi vastaukseesi mallin tarkkuus sekä kuva muutamasta numerosta, jonka tavallinen MLP-verkko tunnistaa väärin mutta konvoluutioverkko oikein.

Tehtävä 13. Lataa sivustolta https://www.kaggle.com/c/dogs-vs-cats/data kissojen ja koirien kuvia. Train-kansiossa on 25000 kuvaa. Valitse niistä 2000 koiran ja 2000 kissan kuvaa test-dataan, ja jätä loput training-dataan. Rakenna training-datan avulla konvoluutioneuroverkko, joka tunnistaa, onko kuvassa kissa vai koira. Raportoi mallisi tarkkuus training- ja test-datassa, ja anna esimerkki 6 test-datan kuvasta, jotka mallisi tunnistaa oikein, ja toisaalta 6 test-datan kuvasta, jotka mallisi tunnistaa väärin. Kerro myös näille kuville ennustetodennäköisyydet. Hyväksyttyyn ratkaisuun vaaditaan, että mallisi kokonaistarkkuus testidatassa on vähintään 80 % (selvästi parempi kuin arvaus).



VIHJEITÄ:

- Tämä tehtävä on hieman haastavampi.
- Sinun on rekisteröidyttävä Kaggle-sivustolle, jotta voit ladata datasetin.
- Kuvat ovat jpg-formaatissa. Sinun täytyy pystyä lukemaan ne NumPy-array muotoon. Eräs työkalu tähän tarkoitukseen on OpenCV-kirjasto.
- Kuvat ovat eri kokoisia. Sinun on skaalattava ne saman kokoisiksi, jotta ne kelpaavat syötteeksi saman neuroverkon input-kerrokseen. OpenCV-kirjasto kelpaa tähänkin tarkoitukseen.

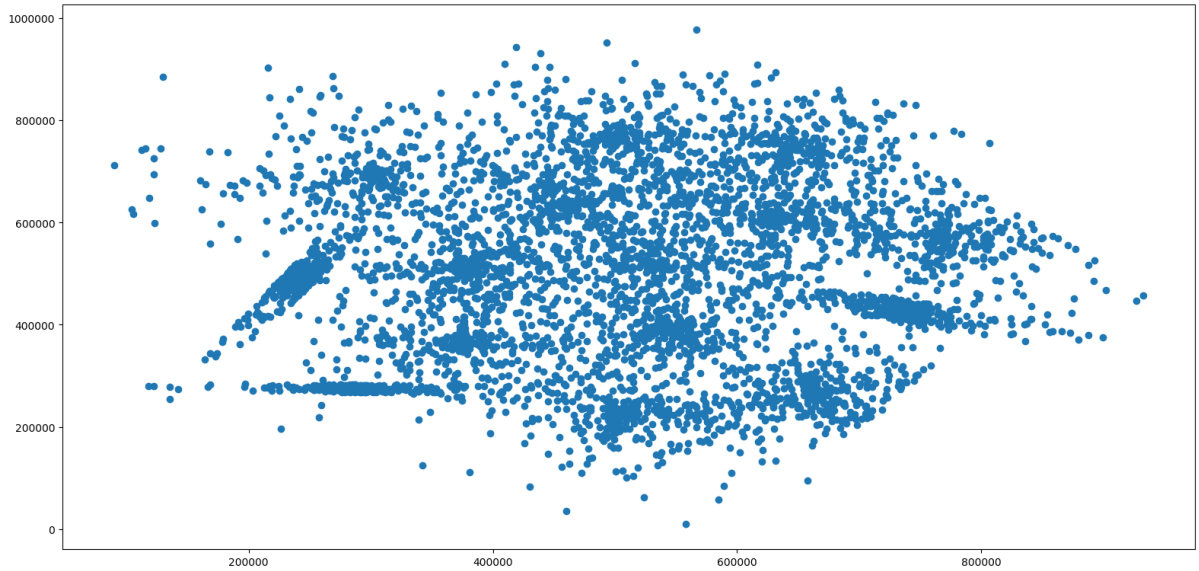
Ohjaamaton koneoppiminen: klusterointi KMeans-algoritmilla

Tehtävä 14 (videoesimerkki). Oheisessa datassa (Mall_Customers.csv) on erään yrityksen asiakastietokanta. Jaa asiakkaat neljään ryhmään (klusteriin) K-means-algoritmin avulla. Käytä muuttujia Age, Annual Income ja Spending Score. Visualisoi ryhmät 3D-koordinaatistossa.

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3

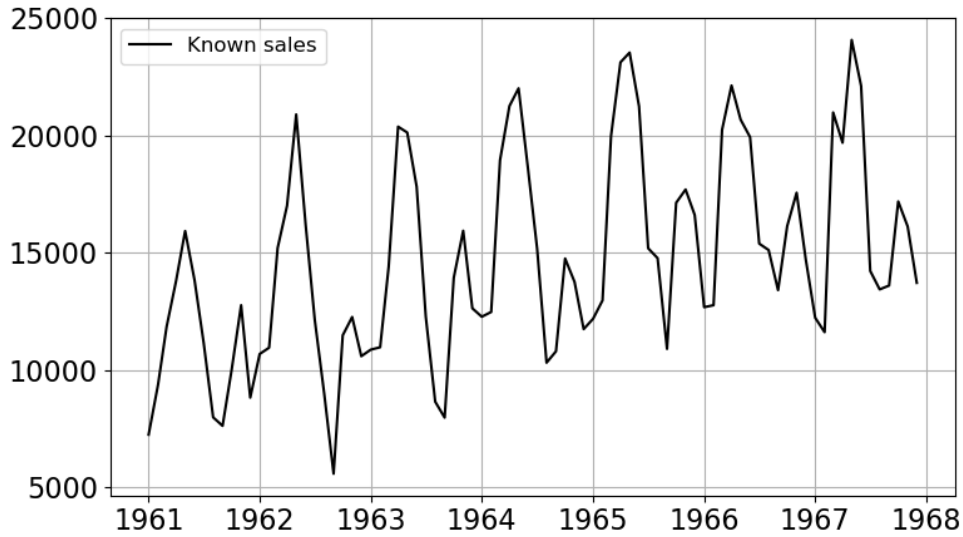
Tehtävä 15 (videoesimerkki). Selvitä edellisen tehtävän tilanteessa optimaalinen määrä tietokannan klustereille. Tee tämä piirtämällä kuvaaja, jossa vaaka-akselilla on klustereiden määrä ja pystyakselilla K-means-algoritmin määrittämä ”inertia”-arvo (inertia mittaa, kuinka kaukana klusterin pisteet ovat klusterin keskipisteestä keskimäärin).

Tehtävä 16. Määritä alla olevalle datalle (2dclusters.csv) sopiva klustereiden määrä. Visualisoi klusterit eri värein.

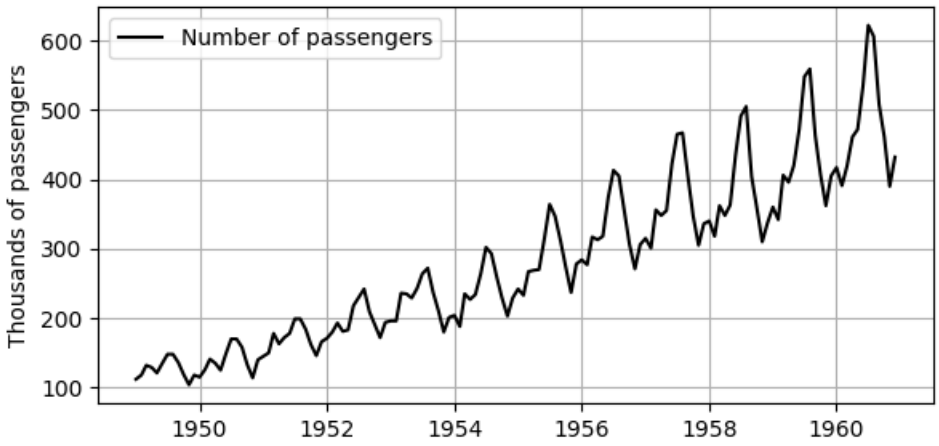


Aikasarjat ja rekursiiviset LSTM-neuroverkot

Tehtävä 17 (videoesimerkki). Oheisessa datassa on kuukausittain myytyjen autojen määrä Amerikassa vuosina 1961 – 1967. Ennusta tämän datan avulla myynti vuoden 1968 kullekin kuukaudelle. Käytä rekursiivista LSTM-neuroverkkoa, joka huomioi myös kausivaihtelun. Datassa on mukana myös vuoden 1968 myynti, mutta jätä se test-dataksi, jonka avulla arvioit mallisi tarkkuutta. Liitä vastaukseesi kuvaaja, jossa näkyy sekä ennuste, että toteutunut myynti.



Tehtävä 18. Oheisessa datassa on kansainvälisten lentomatkustajien määrä tuhansissa vuosina 1949 – 1960. Ennusta vuosien 1949 – 1959 historian perusteella lentomatkustajien määrä vuoden 1960 kullekin kuukaudelle. Käytä rekursiivista LSTM-neuroverkkoa. Piirrä kuvaaja, jossa näkyy sekä ennuste että toteutunut myynti vuodelle 1960. Laske myös ennusteesi keskivirhe.



Soveltava tehtävä: vikaantumisen ennustaminen

Tehtävä 19. Oheisessa datassa (MachineData.csv) on koneiden vikaantumistietoa. Datassa näkyy kyseisen koneen käyttäjätiimi, koneen toimittaja, tämänhetkinen käyttöikä, paine-, kosteus- ja lämpöindeksit. Lisäksi dataan on kirjattu koneen tila (broken).

Machine ID	Team	Provider	Lifetime	PressureInd	MoistureInd	TemperatureInd	Broken
1	TeamA	Provider4	56	92.17885406	104.2302045	96.51715873	0
2	TeamC	Provider4	81	72.07593772	103.0657014	87.27106218	1
3	TeamA	Provider1	60	96.27225443	77.80137602	112.1961703	0
4	TeamC	Provider2	86	94.40646126	108.4936078	72.02537441	1
5	TeamB	Provider1	34	97.75289859	99.413492	103.7562706	0
6	TeamA	Provider1	30	87.67880097	115.7122623	89.79210466	0
7	TeamB	Provider2	68	94.61417404	85.70223564	142.8270014	0
8	TeamB	Provider3	65	96.48330289	93.04679747	98.31619045	1

Ennusta haluamasi koneoppimismallin avulla, mitkä 10 konetta ovat suurimmassa riskissä vikaantua seuraavaksi. Raportoi vastaukseesi seuraavat asiat:

- Käyttämäsi koneoppimismalli.
- Listaus 20 satunnaisesta datarivistä, jossa näkyy muiden annettujen muuttujien lisäksi myös mallisi ennuste -sarake (vikariski numerona välillä 0-1).
- Listaus 10 koneen Machine ID -numeroista, joilla on suurin riski vikaantua seuraavaksi.

Soveltava tehtävä: Lähtevien asiakkaiden tunnistaminen (churn analysis)

Tehtävä 20. 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).

	region	tenure	age	marital	income	employ	gender	tollfree	wireless	cardten	logtoll	logcard	custcat	churn
1	2.000	13.000	44....	1.000	64.000	5.000	0.000	0.000	0.000	110.000	3.240	2.015	1.000	1.000
2	3.000	11.000	33....	1.000	136.000	5.000	0.000	1.000	1.000	125.000	3.033	2.725	4.000	1.000
3	3.000	68.000	52....	1.000	116.000	29.000	1.000	1.000	0.000	2150....	2.890	3.409	3.000	0.000
4	2.000	33.000	33....	0.000	33.000	0.000	1.000	0.000	0.000	0.000	3.240	2.854	1.000	1.000
5	2.000	23.000	30....	1.000	30.000	2.000	0.000	0.000	0.000	0.000	3.240	2.854	3.000	0.000