Innlevering 2 i MAT102 - frist mandag 5/11

Denne innleveringen er det tredje arbeidskravet i MAT102 (etter den første innleveringen og skoleprøven). Den skal leveres på canvas som en kommentert Python-fil:

- Etter at hjelpefilene RSA.py, regression.py og pca.py er kjørt skal besvarelsen være kjørbar.
- Svar på oppgaver som ikke forventes løst av programmet skal stå som kommentar i besvarelsen. Eksempel:

  ```
  #1a: Den dekodede beskjeden blir ...
  ```

- Svar på oppgaver som skal løses av programmet skal stå under koden som gir svaret, som kommentar. Eksempel:

  ```
  #1c: Kode for å kryptere HEI SJEF:
  ... Koden dere skriver
  # Svaret når koden kjøres er: ...
  ```

- Om dere ikke får frem grafikken rett i Spyder, prøv Jupyter notebook. Da kan dere få bruk for denne kodelinjen:

  ```
  %matplotlib inline
  ```

Lever i grupper på tre. Det er tre oppgaver, en om hvert av de store temaene vi ikke tester til eksamen.

1 RSA

I et oppsett for RSA er den offentlige nøkkelen (n, e) gitt ved

$$n = 160169311 \quad \text{og} \quad e = 1737$$

For å sende beskjeder kodes en tekststreng (i engelsk alfabet, uten æøå) ved $A \leftrightarrow 00, B \leftrightarrow 01, \ldots, Z \leftrightarrow 25$. I tillegg lar vi mellomrom være representert ved 99. Strengen deles opp i fire og fire tegn. Hvert firetuppel er dermed representert som et tall mellom 0 (AAAA) og 99999999 (fire mellomrom).

a) En kodet beskjed har representasjonen $[1041706, 4139999]$. Dekod beskjeden.

b) Kod meldingen `HEI SJEF`.

c) Krypter meldingen `HEI SJEF`.

Vi skal så knekke dette oppsettet, og finne ut hva den hemmelige beskjeden U vi har snappet opp betyr:

$$U = [112718817, 85128008, 148479246, 91503316, 26066602, 95584344, 142943071]$$

d) Finn primtallene p og q slik at $n = pq$.

e) Hva må til for at (n, e) skal være en korrekt valgt nøkkel for RSA? Sjekk dette for den oppgitte nøkkelen (n, e).

f) Regn ut dekrypteringsnøkkelen (n, d). Kontroller svaret ved å dekryptere resultatet fra deloppgave c).

g) Dekrypter og dekod den hemmelige beskjeden U.

2 Regresjon

I denne oppgaven skal vi vurdere ulike modeller for et gitt datasett. Datasettet består av 21 punkter og viser temperaturmålinger i et rom over en periode der utetemperaturen synker, mens det så settes på en ovn i et naborom. Tiden måles med tre minutters mellomrom over en periode på en time. Temperaturmålingene T er gitt ved denne tabellen, som også legges ut på canvas for lettere klipping:

$$T = [13.14, 12.89, 12.26, 12.64, 12.22, 12.47, 12.51, 12.80, 12.24, 12.77, 13.35,\\ 12.82, 13.57, 13.38, 14.41, 14.00, 15.68, 15.41, 15.51, 15.86, 15.72]$$

a) Sett opp et scatterplot av dataene. Virker det som dataene passer med beskrivelsen over?

b) Sett opp en lineær tilpasning til datapunktene. Regn også ut determinasjonskoeffisienten.

c) Sett videre opp kvadratisk og kubisk (andre- og tredjegrads) tilnærming. Regn ut determinasjonskoeffisientene for begge modellene.

d) Hvilken modell tror du best beskriver den virkelige situasjonen? Baser svaret ditt på plot og utregninger tidligere i oppgaven.

3 Prinsipalkomponentanalyse

I denne oppgaven skal vi bruke noen numeriske indikatorer for å gruppere fylker. Ettersom noen av tallene jeg har funnet ikke er oppdaterte for i år, da Sør- og Nord-Trøndelag har slått seg sammen til Trøndelag fylke, bruker jeg bare data fra 2017 med nitten norske fylker. Indikatorene vi skal se på er areal (km^2), befolkningstall, antall sysselsatte, BNP per innbygger og BNP per sysselsatt (i kroner). De relevante dataene er med i oppgaveteksten her, men legges også ut i en separat fil på canvas.

Fylker = ['Akershus','Aust-Agder','Buskerud','Finnmark','Hedmark','Hordaland','Møre og Romsdal','Nordland','Nord-Trøndelag','Oppland','Oslo','Rogaland','Sogn og Fjordane','Sør-Trøndelag','Telemark','Troms','Vest-Agder','Vestfold','Østfold']
Indikatorer = ['Areal','Folketall', 'BNP/kapita','BNP/sysselsatt', 'Sysselsatte']
Areal = [4917.95,9155.36,14912.19,48631.38,27397.85,15436.98,15101.07,38478.13,22414, 25192.09,454.10,9376.77,18622.44,18848,15298.23,25876.85,7278.71,2225.38,4187.22]
Folketall =[604368,116673,279714,76149,196190,519963,266274,242866,137233,189479, 666759,472024,110266,317363,173307,165632,184116,247048,292893]
BNPKap=[435982,337974,397080,438594,364944,488515,433030,428402,367157,363111, 820117,488463,455872,473954,371886,451887,403893,364007,331575]
BNPSyss =[918710,771973,831298,808765,777248,922939,834642,850163,759414,731136, 1125019,899272,846111,886057,817060,824648,811833,792748,778412]
Sysselsatte=[270338,47868,125938,37143,86627,254290,127060,116020,62621, 86968,468375,233986,54490,166479,74749,84537,86997,106931,118320]
X = np.transpose(np.array([Areal,Folketall,BNPKap,BNPSyss,Sysselsatte]))

Matrisen X vil etter dette ha riktig format som utgangspunkt for PCA.

a) Som preprosessering, pass på å normalisere matrisen X. Som svar på oppgaven, angi den preprosesserte X.

b) Utfør PCA med to komponenter (dvs. $a = 2$) på denne datamengden. Som svar på oppgaven, angi de to matrisene T og P.

c) Hvilket plot av dataene lar oss se grupperinger av fylkene? Hvilket plot lar oss se grupperinger av indikatorene? Hva heter måten å plotte begge disse tingene samlet?

I resten av oppgaven skal du svare basert på plot og visuell inspeksjon.

d) Hvilke to fylker er likest?

e) Ett fylke skiller seg klart fra de andre. Hvilket? Diskuter hvorfor i gruppen (men dere trenger ikke svare skriftlig på hvorfor).

f) Indikatoren *Areal* er spesielt god til å skille ut ett fylke, og to andre også godt, men ikke like godt som det første. Hvilket fylke? Hvilke er de to neste? Diskuter hvorfor i gruppen (men dere trenger ikke svare skriftlig på hvorfor).

g) Gi minst en ekstra interessant opplysning basert på plottene.

Til slutt

Ikke glem å spørre om hjelp på regneøvelser eller i forbindelse med forelesninger. Dette er ikke en eksamen, men en obligatorisk oppgave som det er meningen dere skal lære av å arbeide med.

Lykke til!

Jon Eivind Vatne