Prosjekt 1 rapport

BiHui Chen

Rapporten inneholder:

- Resultat fra 3 klassifikator, med 3 datasett. Der kombinasjon av egenskap er funnet av nærmeste-nabo klassifikator.
- Svar til avsluttende spørsmål.

Resultat fra klassifikator:

```
-------------------------------------
Best combination of data set 1:  [[0], [0, 3], [0, 1, 3], [0, 1, 2, 3]]
Best combination of data set 2:  [[0], [0, 1], [0, 1, 2]]
Best combination of data set 3:  [[1], [1, 2], [1, 2, 3], [0, 1, 2, 3]]
-------------------------------------
-----------------------------------------------
Data set:  1
Dimensjon:  1
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.81
Error:  0.19

-----------------------------------------------
Data set:  1
Dimensjon:  2
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.89
Error:  0.11

-----------------------------------------------
Data set:  1
Dimensjon:  3
Beste klassifikator er  Minimum_squared_error_Classifier
Accuracy:  0.91
Error:  0.09

-----------------------------------------------
Data set:  1
Dimensjon:  4
Beste klassifikator er  Minimum_squared_error_Classifier
Accuracy:  0.93
Error:  0.07

-----------------------------------------------
Data set:  2
Dimensjon:  1
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.89
Error:  0.11

-----------------------------------------------
Data set:  2
Dimensjon:  2
Beste klassifikator er  Nearest_neighbor_Classifier
Accuracy:  0.99
Error:  0.01
```

```
-----------------------------------------------
Data set:  2
Dimensjon:  3
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.98
Error:  0.02

-----------------------------------------------
Data set:  3
Dimensjon:  1
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.78
Error:  0.22

-----------------------------------------------
Data set:  3
Dimensjon:  2
Beste klassifikator er  Nearest_neighbor_Classifier
Accuracy:  0.91
Error:  0.09

-----------------------------------------------
Data set:  3
Dimensjon:  3
Beste klassifikator er  Nearest_neighbor_Classifier
Accuracy:  0.93
Error:  0.07

-----------------------------------------------
Data set:  3
Dimensjon:  4
Beste klassifikator er  Minimum_failrate_Classifier
Accuracy:  0.93
Error:  0.07
```

Gjennomføring av oppgaven:

Oppgave 1:

Bruk nærmeste-nabo klassifikatoren til å estimere feilraten for alle kombinasjoner av egenskaper av en gitt dimensjon:

NB! Legg merk til skjermbilde av terminalen som forteller beste kombinasjoner av egenskap bruker Python indeks, som begynner fra 0. Dvs. 0 betyr egenskap 1, osv.

Beste kombinasjoner av egenskapene fra data sett 1, gitt ved dimensjon 1 til 4:

Dimensjon 1, egenskap: 1

Dimensjon 2, egenskap: 1, 4

Dimensjon 3 egenskap: 1, 2, 4

Dimensjon 4 egenskap: 1, 2, 3, 4

Beste kombinasjoner av egenskapene fra data sett 2, gitt ved dimensjon 1 til 3:

Dimensjon 1, egenskap: 1

Dimensjon 2, egenskap: 1, 2

Dimensjon 3 egenskap: 1, 2, 3

Beste kombinasjoner av egenskapene fra data sett 3, gitt ved dimensjon 1 til 4:

Dimensjon 1, egenskap: 2

Dimensjon 2, egenskap: 2, 3

Dimensjon 3 egenskap: 2, 3, 4

Dimensjon 4 egenskap: 1, 2, 3, 4

Oppgave 2:

Bruker de beste kombinasjonene som ble funnet av nærmeste-nabo

Se skjermbilde. Det ble gjord sammenlikning av alle tre klassifikator i hvert gitt data sett og hvert dimensjon. Klassifikator med høyest accuracy, dvs. minst feilrate, som er den beste klassifikator blant de tre og ble skrevet ut.

Avsluttende spørsmål:

Bruk av nærmeste-nabo klassifikatoren til å finne gunstige egenskapskombinasjoner:

Oppgave 1:

Fra bevis om konvergens av feilrate for nærmeste-nabo klassifikatoren. Den vil aldri gjør dårligere enn 2 gange av den optimale feilraten, dersom den optimale feilraten er små. Fra observasjon av feilraten med å sette inn random kombinasjon av egenskapene, så er feilraten små. Derfor kan det være fornuftig å bruke nærmeste-nabo klassifikatoren til å finne gunstige egenskapskombinasjoner.

Oppgave 2:

Lineær og kvadratisk klassifikator mer praktisk enn nærmeste-nabo klassifikator:

Den største ulempe med nærmeste-nabo klassifikatoren er at den er for sakte sammenlikner med lineær og kvadratisk klassifikator. Nærmeste-nabo klassifikatoren trenger å regne avstanden til alle punktene i trening sett før den kan gjør en predikasjon. For hver data punkt som skal predikeres, så må avstanden mellom punktet og alle data punktene i trening sett bregnes på nytt. Dette krever mye mer beregningstid sammenlikner med de to andre klassifikatorene. Dermed i praksis så vil vi ofte benytte minimum feilrate eller minste kvadrater klassifikator.

Oppgave 3:

Grunn til å ikke bruke samme datasettet både til trening og evaluering av klassifikatoren:

Klassifikatoren skal bygges slik til å predikere data som den aldri har sett før. Med å bruke samme datasettet til både trening og evaluering vil ikke gir noe fornuftig resultat som forteller hvor godt klassifikatoren er til å predikere nye data. For eksempel, dersom vi bruke samme datasett til både trening og evaluering i nærmeste-nabo klassifikator, så vil den alltid gi 0.0 feilrate. Men dette er en feil estimering av feilraten til å predikere data utenfor trening settet.

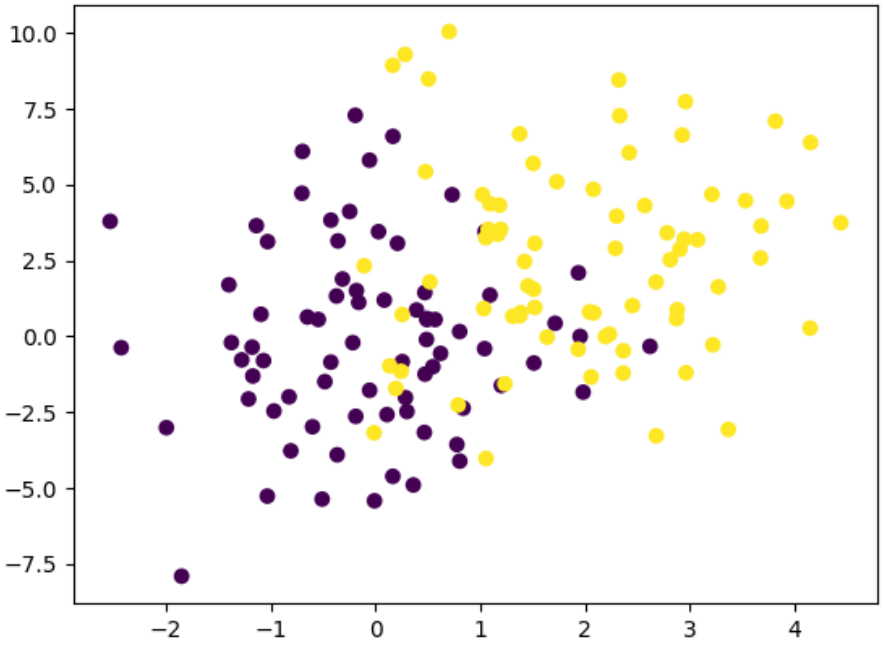
Oppgave 4:

Grunnen til at lineær klassifikator er dårlig til å klassifisere datasett 2:

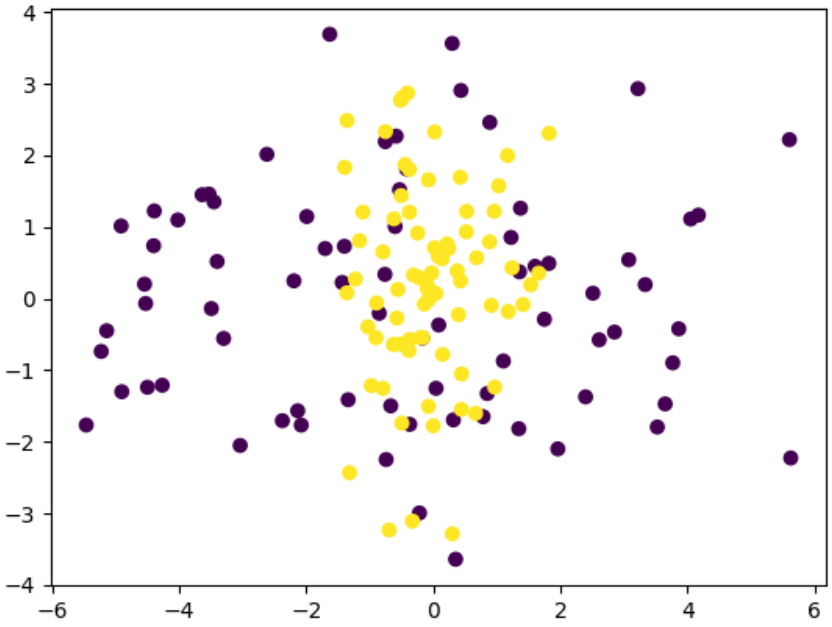
Lineær klassifikator gjør vanligvis dårlig når datasett er ikke lineær separabel. Så i data sett 2, så er lineær klassifikator er dårligere enn de to andre klassifikator. Vi ser at det er bare nærmeste nabo og minimum feilrate klassifikator som gjorde best prediksjon i datasett 2. Og grunnen til at lineær klassifikator gjorde så dårlig er på grunn av data sett 2 er enda mer ikke lineært separabelt enn de to andre data settene.

Jeg har plukket ut 2 egenskaper fra alle tre datasettene og plotta de. Vi ser at data sett 2 har egenskap som kommer midt i en annet egenskap. Det gjør blant annet at det er vanskelig å dele desisjonsregion med en linje. Dermed gjør lineære klassifikatoren dårlig på data sett 2

Plott for datasett 1 med egenskap 0 og 1:



Plott for datasett 2 med egenskap 1 og 2:



Plott for datasett 3 med egenskap 0 og 1: