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, 1, 2, 3]
Best combination of data set 2:  [0, 1]
Best combination of data set 3:  [1, 2, 3]
-------------------------------------
-------------------------------------
Minimum_failrate_Classifier
Data 1:
Accuracy:  0.92
Error rate:  0.08

Data 2:
Accuracy:  0.98
Error rate:  0.02

Data 3:
Accuracy:  0.87
Error rate:  0.13

-------------------------------------
-------------------------------------
Nearest_neighbor_Classifier
Data 1:
Accuracy:  0.91
Error rate:  0.09

Data 2:
Accuracy:  0.99
Error rate:  0.01

Data 3:
Accuracy:  0.93
Error rate:  0.07

-------------------------------------
-------------------------------------
Minimum_squared_error_Classifier
Data 1:
Accuracy:  0.93
Error rate:  0.07

Data 2:
Accuracy:  0.88
Error rate:  0.12

Data 3:
Accuracy:  0.84
Error rate:  0.16

-------------------------------------
```

Gjennomføring av oppgaven:

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

NB! Indeksering av egenskap nummer starter fra 0.

Beste kombinasjon av egenskapene fra data sett 1: 0, 1, 2, 3

Beste kombinasjon av egenskapene fra data sett 2: 0, 1

Beste kombinasjon av egenskapene fra data sett 3: 1, 2, 3

Bruker de beste kombinasjonene som ble funnet av nærmeste-nabo klasssifikatoren, og bruke alle tre klassifikatorene til å klassifisere data settene:

Data sett 1: Minste kvadraters metode klassifikatoren gir minst feilrate med 0.07.

Data sett 2: Nærmeste-nabo klassifikatoren gir minst feilrate med 0.01.

Data sett 3: Nærmeste-nabo klassifikatoren gir minst feilrate med 0.07.

Generelt ser vi at nærmeste-nabo klassifikatoren gir minst feilrate for data settene, selv om minst kvadraters metode klassififkator er bedre enn nærmeste-nabo kalssifikator til å klassifisere data sett 1, så differansen mellom feilratene små.

Avsluttende spørsmål:

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

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.

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.

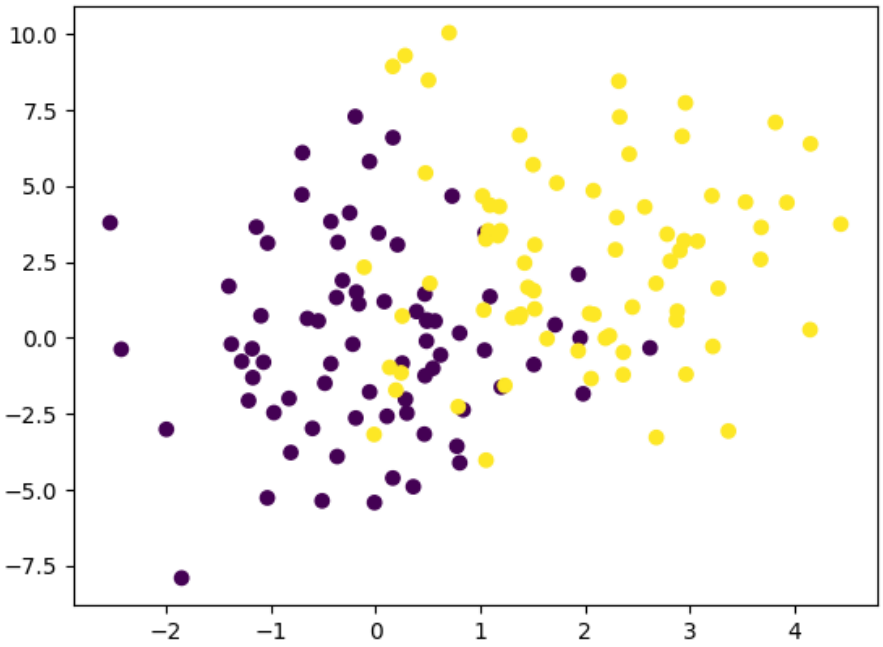
Grunn til å ikke bruke samme datasettet både til trening og evaluering av klassifikatoren:

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.

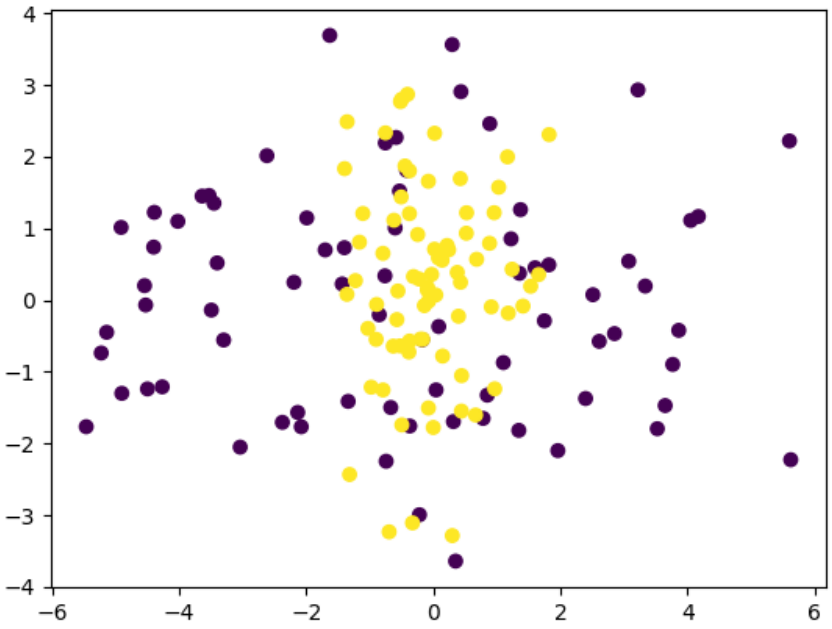
Grunnen til at lineær klassifikator er dårlig til å klassifisere datasett 2:

Lineær klassifikator gjør vanligvis dårlig når datasett er ikke lineær separabel. Ved plotting av datasett1 med egenskap 0 og 1 ser vi selv om den ikke er lineær separabel, men likevel kan vi separere datasett bra med en linje mellom. Mens for plotting av datasett 2 med egenskap 1 og 2 ser vi at den kan ikke separeres med en linje. Derfor er lineær klassfikator dårlig til å klassifisere ikke lineær separable datasett.

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: