Prosjektoppgave 1
Mønstergjenkjenning TEK5020

Simen Sørlie

November 15, 2021

1 Oppgavebeskrivelse

Møsntergjennkjenning handler om å analysere store datasett for å finne et mønster. Denne oppgaven går ut på å finne den beste egenskapskombinasjonen for et gitt datasett ved hjelp av nærmeste-nabo klassifikatoren. Deretter utfører man videre analyser på disse egenskapskombinasjonene ved hjelp minimum feilrate klassifikator og minste kvadraters metode. Disse klassifikatorene blir vurdert ut i fra feilrate. Som feilrateestimat benyttes forholdet mellom antall feilklassifiserte objekter og det totale antall objekter i testsettet:
$\hat{P}(e) = \frac{n_{feil}}{n_{totalt}}$

Nærmeste nabo-klassifikatoren benyttes for å finne beste egenskapskombinasjon for hver dimensjon. Dette gjøres ved å beregne feilraten for hver kombinasjon av de ulike egenskapene i en gitt dimensjon, slik at man får ut den beste kombinasjonen med lavest feilrate. Man bruker deretter den beste egenskapskombinasjonen for å finne den beste klassifikatoren. Man sammenligner da feilraten til de 3 ulike metodene: Nærmeste nabo-klassifikatoren, minimum feilrate-klassifikatoren og minste kvadraters metode. Den klassifikatoren med lavest feilrate blir da ansett som den beste for den gitte egenskapskombinasjonen.Metoden blir brukt på tre datasett.Datasett 1 inneholder 300 objekter med 4 egenskaper, datasett 2 består av 300 objekter med 3 egenskaper og datasett 3 har 400 objekter med 4 egenskaper

2 Resultater

Starter med å finne den beste egenskapskombinasjonen ved hjelp av nærmeste nabo-klassifikatoren for alle egenskapsdimensjoner i hvert av de tre datasettene.

Tabell 1. Feilraten for hver egenskapskombinasjon i d=1 for hvert datasett.

d=1	Datasett 1	Datasett 2	Datasett 3
1	0.24	0.18	0.33
2	0.36	0.28	0.31
3	0.433	0.4933	0.345
4	0.387	—-	0.395

Egenskap med minste feilrate for hvert av datasettene er følgende: Datasett 1: 0.24, datasett 2: 0.18 og datasett 3: 0.31

Tabell 2. Feilrate for hver egenskapskombinasjon i d=2 for hvert datasett.

d=2	Datasett 1	Datasett 2	Datasett 3
1 2	0.18	0.013	0.215
1 3	0.193	0.193	0.17
1 4	0.167	—-	0.285
2 3	0.32	0.2867	0.095
2 4	0.267	—-	0.24
3 4	0.3	—-	0.19

Egenskap med minste feilrate for hvert av datasettene er følgende: Datasett 1: 0.167, datasett 2: 0.013 og datasett 3: 0.095.

Tabell 2: feilrate for hver egenskapskombinasjon i d=3 for hvert datasett.

d=2	Datasett 1	Datasett 2	Datasett 3
1 2 3	0.1467	0.02	0.1
1 2 4	0.1	—-	0.2
1 3 4	0.1267	—-	0.15
2 3 4	0.213	—-	0.075

Egenskap med minste feilrate for hvert av datsettenere er følgende: Datasett 1: 0.1, datasett 2: 0.02, datasett 3: 0.075.

Tabell 4. Feilrate for hver egenskapskombinasjon i d=4 for hvert datasett.

d=2	Datasett 1	Datasett 2	Datasett 3
1 2 3 4	0.093	—-	0.095

Feilrate for hver klassifikator gitt den beste egenskapskombinasjonen for hver dimensjon i datasett 1.

Egenskaper	Nærmeste nabo	Minste feilrate	Minste kvadraters
1	0.24	0.187	0.187
1 4	0.167	0.113	0.113
1 2 4	0.1	0.1	0.0933
1 2 3 4	0.0933	0.08	0.0733

Feilrate for hver klassifikator gitt den beste egenskapskombinasjonen for hver dimensjon i datasett 2.

Egenskaper	Nærmeste nabo	Minste feilrate	Minste kvadraters
1	0.18	0.107	0.107
1 2	0.0133	0.02	0.12
1 2 3	0.02	0.02	0.12

Feilrate for hver klassifikator gitt den beste egenskapskombinasjonen for hver dimensjon i datasett 3.

Egenskaper	Nærmeste nabo	Minste feilrate	Minste kvadraters
2	0.31	0.225	0.335
2 3	0.095	0.2	0.2
2 3 4	0.075	0.13	0.16
1 2 3 4	0.095	0.07	0.12

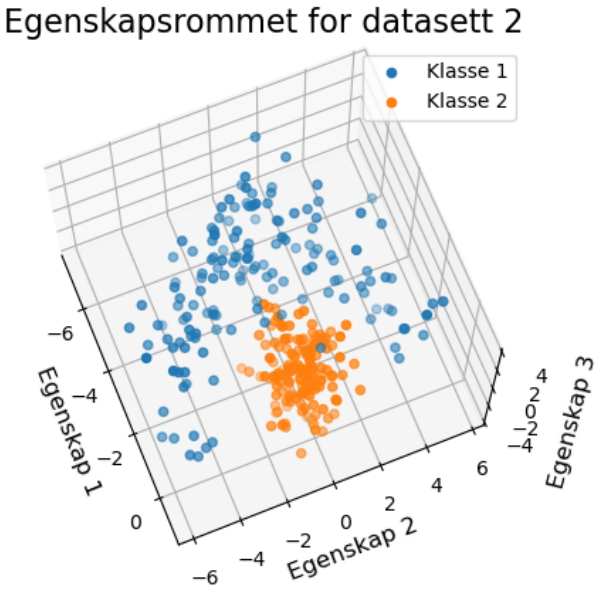

Figure 1: Egenskapsrommet for datasett 2. Objekter tilhørende klasse 2 i klynge nær origo, mens objekter tilhørende klasse 1 spredt rundt.

3 Svar på spørsmål

1. Hvorfor er det fornuftig å benytte nærmeste-nabo klassifikatoren til å finne gunstige egenskapskombinasjoner?

   Nærmeste-nabo baserer seg på om egenskapene befinner seg nærme hverandre. Om man tar et datasett hvor egenskap 1 og 2 er blandet sammen, mens egenskap 1 og 3 er lokalisert i hvert sitt område, så vil nærmeste-nabo oppdage dette. I tillegg så er nærmeste-nabo ikke-parametrisk og gjør få antagelser, som gir en grei feilrate, uavhengig av fordelingen i datasettet.

2. Hvorfor kan det i en praktisk anvendelse være fornuftig å finne en lineær eller kvadratisk klassifikator til erstatning for nærmeste-nabo klassifikatoren?

   Behovet for regnekraft er en faktor i nærmeste-nabo klassifikatoren, samtidig blir det ikke gjort noe antagelser om datasettet. I en praktisk anvendelse kan det være at slike antagelser kan være en fordel å ta. Det vil derfor være nyttig å bruke en klassifikator som utnytter denne informasjonen.

3. Hvorfor er det lite gunstig å bruke samme datasettet både til trening og evaluering av en klassifikator?

   Evalueringen vil da basere seg på data som allerede er blir brukt til trening av modellen. Det vil derfor være bedre å trene en klassifikator med et datasett, og se hvordan klassifikatoren presterer med et annet testsett. Man kan da sammenligne hvor gode forskjellige klassifikatorer er.

4. Hvorfor gir en lineær klassifikator dårlige resultater for datasett 2? Om vi ser på figur 1, kan man se at det er vanskelig å skille objektene i det lineære området. Klasse 1 bretter seg rundt klasse 2, som er samlet i en klynge.