TEK9020 - Prosjektoppgave 1

Kristoffer Langstad
kristoffer.langstad@its.uio.no

1 Introduksjon

I denne oppgaven er tanken å illustrere hva som skjer i mønstergjenkjenning. Ved å bruke tre forskjellige klassifikatorer; Minimum feilrate med normalfordelings antagelse, Minste kvadraters metode og Nærmeste-nabo, skal jeg forsøke å finne ut hvilken av disse tre mønstergjenkjenningsklassifikatorene som er den beste klassifikatoren av tre ulike datasett med forskjellige objekter og egenskaper. Klassifikatorene blir vurdert ved å estimere feilrater med et to-klasse problem. Klassifikatorene trenes opp med et trenings sett og blir evaluert med et test sett for å finne den beste klassifikatoren. Dette gjøres med programmeringsverktøyet Python (A). Til slutt diskuteres resultatene, og det svares på noen relevante spørsmål angående resultatene som er funnet og om klassifikatorene som er brukt.

2 Metode

2.1 Datasett

Hvert av de tre datasettene inneholder forskjellige antall objekter og antall egenskaper, som er fordelt på to klasser. Datasett 1 inneholder 300 objekter med 4 egenskaper, datasett 2 inneholder 300 objekter med 3 egenskaper og datasett 3 inneholder 400 objekter med 4 egenskaper. Datasettene blir så individuelt delt inn i 50% trening og 50% test datasett, der trenings data består av odde nummererte objekter mens test data består av like nummererte objekter.

2.2 Trening og evaluering

Først skal nærmeste-nabo klassifikatoren brukes til å regne ut de beste egenskapskombinasjonene for hver dimensjon for hvert datasett. Her regnes det ut feilrate for hver egenskapskombinasjon, der den egenskapskombinasjonen hver dimensjon som har den laveste feilraten dømmes som den beste egenskapskombinasjonen.

Når dette er gjort for samtlige egenskapskombinasjoner for et datasett, så skal alle klassifikatorene evalueres. Kun de beste egenskapskombinasjonene hver dimensjon brukes til å bestemme hvilken av de tre klassifikatorene som best klarer å klassifisere objektene i datasettene riktig. Den klassifikatoren med lavest feilrate for den gitte egenskapskombinasjonen anses som den beste klassifikatoren i dette tilfellet. Dette gjøres da for alle tre datasettene.

For å evaluere klassifikatorene benyttes feilrateestimatet som regnes ut som forholdet mellom antall feilklassifiserte objekter og det totale antall objekter i testsettet:

$$\hat{P}(e) = \frac{n_{feil}}{n_{totalt}} \tag{1}$$

3 Resultater

3.1 Finne best egenskapskombinasjonene

Først finner vi de beste egenskapskombinasjonene med nærmest-nabo klassifikatoren for hver dimensjon for all tre datasettene. Feilrate resultatene av disse vises i Tabell 1 til 4.

Dim=1	Datasett 1	Datasett 2	Datasett 3
1	**0.240**	**0.180**	0.330
2	0.360	0.280	**0.310**
3	0.4333	0.4933	0.345
4	0.3867	-	0.395

Table 1: Feilratene for alle egenskapene i 1 dimensjon, for de tre datasettene. Den beste egenskapen med lavest feilrate for hvert datasett er markert i svart, manglende egenskaper er markert med en strek.

Dim=2	Datasett 1	Datasett 2	Datasett 3
12	0.180	**0.0133**	0.215
13	0.1933	0.1933	0.170
14	**0.1667**	-	0.285
23	0.320	0.2867	**0.095**
24	0.2267	-	0.240
34	0.300	-	0.190

Table 2: Feilratene for alle egenskapene i 2 dimensjoner, for de tre datasettene. Den beste egenskapen med lavest feilrate for hvert datasett er markert i svart, manglende egenskaper er markert med en strek.

Dim=3	Datasett 1	Datasett 2	Datasett 3
123	0.1467	**0.020**	0.100
124	**0.100**	-	0.200
134	0.1267	-	0.150
234	0.2133	-	**0.075**

Table 3: Feilratene for alle egenskapene i 3 dimensjoner, for de tre datasettene. Den beste egenskapen med lavest feilrate for hvert datasett er markert i svart, manglende egenskaper er markert med en strek.

Dim=4	Datasett 1	Datasett 2	Datasett 3
1234	**0.0933**	-	**0.095**

Table 4: Feilratene for alle egenskapene i 4 dimensjoner, for de tre datasettene. Den beste egenskapen med lavest feilrate for hvert datasett er markert i svart, manglende egenskaper er markert med en strek.

3.2 Evaluere klassifikatorene

Med de beste egenskapskombinasjonene for hver dimensjon med nærmeste-nabo klassifikatoren, brukes så minimum feilrate og minste kvadraters metode klassifikatorene til å beregne feilrate for å finne ut hvilken av de tre klassifikatorene som gir lavest feilrate og beste resultat for hver egenskapskombinasjon for hvert datasett. Tabellene 5, 6 og 7 viser henholdsvis resultatene av klassifikator evalueringene for datasett 1, 2 og 3.

Beste komb.	N.N.	M.F.	M.K.
1	0.240	**0.187**	**0.187**
14	0.167	**0.113**	**0.113**
124	0.100	0.100	**0.0933**
1234	0.0933	0.0800	**0.0733**

Table 5: *Datasett 1*: Feilratene for de beste egenskapskombinasjonene med nærmest-nabo (N.N.), minimum feilrate (M.F.) og minste kvadraters metode (M.K.) klassifikatorene. Den beste klassifikatoren med lavest feilrate for hver egenskapskombinasjon er markert i svart.

Beste komb.	N.N.	M.F.	M.K.
1	0.180	**0.107**	**0.107**
12	**0.0133**	0.020	0.120
123	**0.020**	**0.020**	0.120

Table 6: *Datasett 2*: Feilratene for de beste egenskapskombinasjonene med nærmest-nabo (N.N.), minimum feilrate (M.F.) og minste kvadraters metode (M.K.) klassifikatorene. Den beste klassifikatoren med lavest feilrate for hver egenskapskombinasjon er markert i svart.

Beste komb.	N.N.	M.F.	M.K.
2	0.310	**0.225**	0.335
23	**0.095**	0.200	0.200
234	**0.075**	0.130	0.160
1234	0.095	**0.070**	0.120

Table 7: *Datasett 3*: Feilratene for de beste egenskapskombinasjonene med nærmest-nabo (N.N.), minimum feilrate (M.F.) og minste kvadraters metode (M.K.) klassifikatorene. Den beste klassifikatoren med lavest feilrate for hver egenskapskombinasjon er markert i svart.

3.3 Datasett 2 egenskapsrom visualisert

I Figur 1 ser vi egenskapsrommet til datasett 2 visualisert i 3 dimensjoner, vinklet for å best mulig vise plasseringene av klasse 1 og 2 objektene utifra egenskapene.

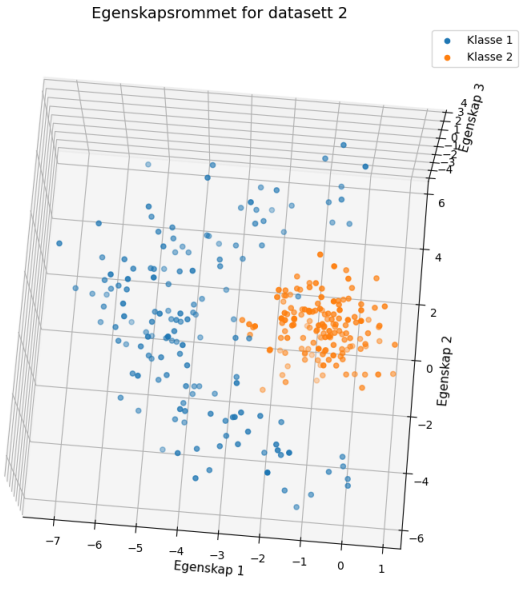

Figure 1: Plot av egenskapsrommet til datasett 2 i 3D som viser fordelingen av objekter med klasse 1 (blå) og klasse 2 (oransje). Klasse 2 objektene smaler seg som en klase rundt et punkt, mens klasse 1 objektene ligger i en halvsirkel rundt.

4 Diskusjon

4.1 Evaluering av nærmeste-nabo klassifikator

Noe av det som gjør at nærmeste-nabo klassifikatoren egner seg godt til å finne gunstige egenskapskombinasjoner er at den finner egenskaper med dataene som best skiller de forskjellige klassene. Den prøver å finne det punktet i datasettet som er nærmest et annet gitt punkt eller klasse. Den egner seg spesielt godt når det er objekter med flere egenskaper involvert. Den bruker heller ingen antagelser om fordelingen til dataene siden den er en ikke-parametrisk algoritme, som gjør at den fungerer fint på data med ingen eller lite tidligere kunnskap som kan være passende for ikke syntetiske data.

Den har også en relativt rask trenings fase, siden den ikke trenger eksplisitt trening slik at kun treningsegenskapene og tilhørende klassene lagres istedenfor en modell. Når det kommer til test dataene, så må den derimot gå gjennom hele trenings datasettet for å finne det objektet og klassen som er nærmest.

4.2 Praktisk anvendelse

Ofte når det kommer til praktiske anvendelser, så kan fort antall dimensjoner på egenskapsrom og størrelsene på trenings dataene blir veldig store. I slike tilfeller kan det fort oppstå feilklassifiseringer siden avstandsberegningene ofte blir mer unøyaktige når dimensjonene øker. Den vil også bruke lang tid under testing, som nevnt allerede, da nærmeste-nabo klassifikatoren må gå gjennom hele trenings settet for å finne kortest avstanden til et objekt for klassifisering. I slike tilfeller kan det da lønne seg å bruke en lineær eller kvadratisk klassifikator. Disse klassifikatorene bruker ofte noe lenger tid på å trenes opp, men har til fordel at de så kan brukes på hele test settet når de først har blitt trent. Dermed holder det å gå gjennom trenings settet kun en gang under treningen, som vil spare tid og minne som trengs.

4.3 Trening og test data

Å splitte data inn i trening og test sett er veldig vanlig i mønstergjenkjenning for å trene opp en klassifikator. Dersom test settet skulle vært en del av trenings settet, oppstår det veldig fort overtrening som er lite gunstig for en klassifikator. Når en klassifikator er overtrent på et datasett, så vil det gi veldig bra resultater på det test settet som var en del av treningen men vil den ikke klare å klassifisere et nytt datasett like bra. Med godt trente klassifikatorer vil vi at skal kunne gjøre det bra på generelle datasett og ikke kun på de dataene klassifikatoren har blitt trent opp på.

4.4 Rangering av klassifikatorer

4.4.1 Datasett 1

Fra Tabell 5 er det enkelt å se at minste kvadraters metode er den beste klassifikatoren på datasett 1. Den er enten like god som minimum feilrate i små dimensjoner eller bedre enn begge de to andre. Siden minste kvadraters metode er en lineær klassifikator, så kan vi konkludere med at datasett 1 som er et syntetisk datasett med trekninger fra kjente tetthetsfordelinger er lineært separabelt. Feilraten minker for alle klassifikatorene når antall egenskaper øker, men minste kvadraters metode har den laveste feilraten og er best på datasett 1 med 4 egenskaper.

4.4.2 Datasett 2

Som for datasett 1, er minimum feilrate og minste kvadraters metode klassifikatorene like gode på datasett 2 i en dimensjon. Begge disse to gir dårligere feilrate når dimensjonen øker, mens nærmeste-nabo gir bedre resultater og er best på datasett 2 med 2 egenskaper. Dette kan forklares ved å se på Figur 1. Der ser vi at de to klassene i datasettet ikke er lineært separable i høyere dimensjoner, og kan ikke bli optimalt separert med en rett linje. Dette passer bedre for f.eks. en kvadratisk klassifikator, som vi kan se i tabellen for datasett 2.

4.4.3 Datasett 3

For datasett 3 i Tabell 7 er nærmeste-nabo best for dimensjon 2 og 3, mens minimum feilrate er best i dimensjon 1 og 4. Den laveste feilraten på datasett 3 oppnås ved å bruke minimum feilrate klassifikatoren i 4 dimensjoner som er det komplette datasettet med alle 4 egenskapene til objektene.

5 Konklusjon

Med nærmeste-nabo klassifikatoren har jeg funnet de beste egenskapskombinasjonene for tre datasett med forskjellig antall objekter og antall egenskaper. Klassifikatorene ble trent med et trenings datasett og evaluert med et test datasett. Med de beste egenskapskombinasjonene har jeg evaluert nærmeste-nabo, minimum feilrate og minste kvadraters metode klassifikatorene for å finne ut hvilken klassifikator som egner seg best på de tre datasettene.

På datasett 1 gav den lineære klassifikatoren minste kvadraters metode den laveste feilraten for alle dimensjoner, med lavest feilrate i 4 dimensjoner. På datasett 2 og 3 gir minste kvadraters metode de dårligste resultatene, mens de to andre ble bedre. På datasett 2 ble den laveste feilraten oppnådd med nærmeste-nabo klassifikatoren med 2 dimensjoner. På datasett 3 ble den laveste feilraten oppnådd med minimum feilrate klassifikatoren med 4 dimensjoner.

A Python kode

Koden som er brukt i dette prosjektet kan bli funnet i GitHub på: https://github.com/krilangs/TEK9020/tree/main/Project1