En introduksjon til mønstergjenkjenneingsanvendelser

Elin Finstad

Institutt for informatikk
Universitetet i Oslo
Norge
November 13, 2020

I. INTRODUKSJON

Mønstergjenkjenning handler om å lete gjennom store datasett etter mønstre. Overalt er det mulig å finne mønstre, nor som gir faget et stort bruksområde. Det benyttes blant annet innen statistiske analyser, signalbehandling, bildegjenkjenning, bioinformatikk og maskinlæring. Ved å innføre mønstergjenkjenning kan man lære opp maskiner til å få en tilnærmet menneskelig gjenkjenningsevne, men mer effektivt på store mengder data.

Som en introduksjon til hvordan en mønstergjenkjenningsanvendelse foregår vil jeg i denne oppgaven forsøke å finne den beste egenskapskombinasjonen for et gitt datasett ved hjelp av nærmeste nabo-klassifikatoren, før videre analyser på disse egenskapskombinasjonene blir gjennomført med minimum feilrate klassifikator og minste kvadraters metode. Klassifikatorene blir vurdert basert på feilrate. Ettersom dette kun er en introduksjon, betraktes kun problemer med to klasser.

II. METODE

Hvert datasett deles opp i treningsdata og testdata. I dette tilfellet er objektene fordelt slik at treningssettet består av odde nummererte objekter mens testsettet består av like nummererte objekter. Både treningssettet og testsettet har derfor like mange datapunkter.

Målet med oppgaven er å sammenlikne ulike klassifikatorer basert på feilrate. Her brukes forholdes mellom antall feilklassifiserte objekter og det totale antallet objekter som feilrateestimat:

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

I første omgang benyttes nærmeste nabo-klassifikatoren til å finne den beste egenskapskombinasjonen for hver dimensjon. Dette gjøres ved å beregne feilraten for hver kombinasjon av egenskaper i en gitt dimensjon, hvor den beste kombinasjonen er den med lavest feilrate. Videre benyttes kun den beste egenskapkombinasjonen for hver dimensjon til å finne den beste klassifikatoren. Her sammenliknes da feilraten til nærmeste nabo-klassifikatoren, minimum feilrate-klassifikatoren og minste kvadraters metode. Den klassifikatoren med den laveste feilraten anses da som den beste for den gitte egenskapskombinasjonen.

Metoden blir anvendt på tre datasett. Datasett 1 og datasett 2 er dannet ved trekninger fra kjente tetthetsfordelinger og inneholder henholdsvis 300 objekter med 4 egenskaper og 300 objekter med 3 egenskaper. Datasett 3 er generert ved uttrekking av formegenskaper fra segmenter av to ulike bilmodeller. Dette datasettet har 400 objekter med 4 egenskaper.

III. RESULTATER

Begynner først med å finne den beste egenskapskombinasjonen basert på nærmeste nabo-klassifikatoren. Dette gjøres for alle egenskapsdimesjoner for hvert av de tre datasettene. Resultatene er presentert i tables I to IV.

Table I. Feilrate for hver egenskapskombinasjon i dimensjon 1 for hvert av de tre datasettene. Egenskapen med lavest feilrate er markert i blått for hvert av de tre datasettene. Ugyldige egenskaper er markert med en strek.

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	0.493	0.345
4	0.387	—	0.395

Table II. Feilrate for hver egenskapskombinasjon i dimensjon 2 for hvert av de tre datasettene. Egenskapskombinasjonen med lavest feilrate er markert i blått for hvert av de tre datasettene. Ugyldige egenskapskombinasjoner er markert med en strek.

d = 2	**Datasett 1**	**Datasett 2**	**Datasett 3**
12	0.18	0.013	0.215
13	0.193	0.193	0.17
14	0.167	—	0.285
23	0.32	0.287	0.095
24	0.227	—	0.24
34	0.3	—	0.19

Table III. Feilrate for hver egenskapskombinasjon i dimensjon 3 for hvert av de tre datasettene. Egenskapskombinasjonen med lavest feilrate er markert i blått for hvert av de tre datasettene. Ugyldige egenskapskombinasjoner er markert med en strek.

d = 3	**Datasett 1**	**Datasett 2**	**Datasett 3**
123	0.147	0.02	0.1
124	0.1	—	0.2
134	0.127	—	0.15
234	0.213	—	0.075

Table IV. Feilrate for hver egenskapskombinasjon i dimensjon 4 for hvert av de tre datasettene. Egenskapskombinasjonen med lavest feilrate er markert i blått for hvert av de tre datasettene. Ugyldige egenskapskombinasjoner er markert med en strek.

d = 4	**Datasett 1**	**Datasett 2**	**Datasett 3**
1234	0.093	—	0.095

Med utgangspunkt i den beste egenskapskombinasjonen innen hver dimensjon, benyttes så minimum feilrate-klassifikatoren og minste kvadraters metode, for å finne den klassifikatoren som gir lavest estimert feilrate. tables V to VII viser henholdsvis resultatene for datasett 1, 2 og 3.

Table V. Feilrate for hver av de tre klassifikatorene gitt den beste egenskapskombinasjon innen hver dimensjon for datasett 1. Klassifikatoren med lavest feilrate innen hver egenskapsdimensjon er markert i blått.

Egenskaper	**Nærmeste nabo**	**Min. feilrate**	**Minste kvadraters**
1	0.24	0.187	0.187
14	0.167	0.113	0.113
124	0.1	0.1	0.0933
1234	0.0933	0.08	0.0733

Table VI. Feilrate for hver av de tre klassifikatorene gitt den beste egenskapskombinasjon innen hver dimensjon for datasett 2. Klassifikatoren med lavest feilrate innen hver egenskapsdimensjon er markert i blått.

Egenskaper	Nærmeste nabo	Min. feilrate	Minste kvadraters
1	0.18	0.107	0.107
12	0.0133	0.02	0.12
123	0.02	0.02	0.12

Table VII. Feilrate for hver av de tre klassifikatorene gitt den beste egenskapskombinasjon innen hver dimensjon for datasett 3. Klassifikatoren med lavest feilrate innen hver egenskapsdimensjon er markert i blått.

Egenskaper	Nærmeste nabo	Min. feilrate	Minste kvadraters
2	0.31	0.225	0.335
23	0.095	0.2	0.2
234	0.075	0.13	0.16
1234	0.095	0.07	0.12

Til slutt et plot som illustrerer egenskapsrommet for datasett 2.

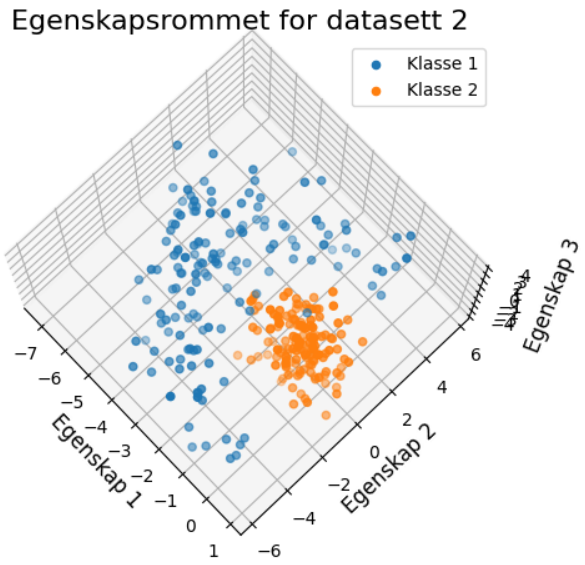

Figure 1. Egenskapsrommmet for datasett 2. Objekter tilhørende klasse 2 i en tett klynge nær origo, med objekter tilhørende klasse 1 mer spredt i en halvsirkel rundt.

IV. DISKUSJON

A. Nærmeste nabo-klassifisering

Nærmeste nabo-klassifikatoren er en ikke-parametrisk algoritme, den bruker med andre ord ingen antakelser om fordelingen til inngangsdataen. Parametriske metoder antar at mesteparten av inngangsdataene følger teoretiske begrensninger, men det er ikke gitt at reell data faktisk følger disse antakelsene. Nærmeste nabo-klassifikatoren er derfor et godt valg til å finne gunstige egenskapskombinasjoner når det er lite eller ingen tidligere kunnskap tilgjengelig.

Algoritmen krever heller ikke et eksplisitt treningssteg, ettersom det ikke læres opp en modell. Treningsdelen krever kun lagring av treningsegenskapene og den tilhørende klassen, noe som gjør treningsfasen rask. Til gjengjeld er testfasen tidkrevende, ettersom man for hvert testobjekt må løpe gjennom hele treningssettet for å finne det treningsobjektet med den korteste avstanden til testobjektet.

Nærmeste nabo-klassifisering krever homogene egenskaper. Ved bruk av f.eks. euklidsk norm for beregning av avstand er man avhengig av at en gitt avstand for egenskap 1 er den samme som for egenskap 2, ettersom absolutt differanse av egenskaper vektes likt. Det kan derfor være hensiktsmessig å skalere dataene på forhånd.

Metoden fungerer best på balansert data, altså at hver klasse er representert med omtrent like mange treningsobjekter. Dersom de fleste treningsobjektene tilhører klasse ω_1, mens bare noen få tilhører klasse ω_2 kan det føre til at objekter som egentlig tilhører klasse ω_2 blir feilklassifisert. I tillegg er klassifikatoren sensitiv for avvikere, siden det drastisk endrer klassebestemmelsen.

Nærmeste nabo-klassifikatoren feiler når dimensjonen på egenskapsrommet blir stor. Når dimensjonen øker blir avstandsberegningen ofte mindre nøyaktige, og følgelig blir forskjellen på et nært objekt og et objekt langt unna liten. Dette fører fort til feilklassifiseringer.

B. Praktiske anvendelser

I en praktisk anvendelse kan det være fornuftig å velge en lineær eller kvadratisk klassifikator til erstatning for nærmesete nabo-klassifikatoren. Som forklart tidligere kan nærmeste nabo-klassifikatoren være veldig tidkrevende, ettersom man for hvert testobjekt må løpe gjennom hele treningssettet. Dersom man ved en praktisk anvendelse har et stort dataset kan det å velge en lineær eller kvadratisk klassifikator minske kjøretiden kraftig. En lineær eller kvadratisk klassifikator krever mer treningstid, men treningsmodellen kan til gjengjeld brukes på alle testobjektene. Det betyr at man kun trenger å løpe gjennom treningsdataene til å lage modellen i stedet for å løpe gjennom hele treningssettet for hvert objekt i testsettet.

C. Trening- og testdata

Det er lite hensiktsmessig å bruke det samme datasettet både til trening og evaluering av en klassifikator, ettersom dette fort fører til overtilpasning. Dersom det samme datasettet brukes til både trening og evaluering av klassifikatoren vil resultatet bli veldig bra. Dersom klassifikatoren videre blir brukt på andre datasett vil ikke resultatene fra evalueringen være representative, ettersom klassifikatoren nå er trent til å fungere veldig bra på treningssettet, men ikke på et generelt datasett.

D. Datasett 1

Table V viser at minste kvadraters metode er den beste klassifikatoren på datasett 1. Datasettet er med andre ord lineært separabelt. Ved dimensjon 1 og 2 gir minimum feilrate-klassifikatoren samme resultat som minste kvadraters metode, mens nærmeste nabo-klassifikatoren gir dårligst resultat uansett egenskapsdimensjon. Minste kvadraters metode brukt på alle egenskapene gir den mest nøyaktige klassifiseringen, med en feilrate på 0.0733.

E. Datasett 2

Fra Table VI viser at nærmeste nabo-klassifikatoren med egenskapskombinasjonen 12 gir den mest nøyaktige klassifiseringen, med en feilrate på 0.0133. Tabellen viser at den lineære klassifikatoren minste kvadraters metode gir et vesentlig dårligere resultat enn de to andre klassifikatorene i dimensjon 2 og 3. Dette kan lett forklares ved å se på plottet i Figure 1. Her ser man tydelig at de to klassene ikke er lineært separable, da klasse 2 omringer klasse 1. En kvadratisk klassifikator gir derfor et bedre resultat for dette datasettet.

F. Datasett 3

Table VII viser at ved dimensjon 1 og 4 er minimum feilrate-klassifikatoren best, mens ved dimensjon 2 og 3 er nærmeste nabo-klassifikatoren den beste. Det beste resultatet oppnås ved å bruke minimum feilrate-klassifikatoren på alle egenskapene, hvor feilraten er nede på 0.07, men feilraten er også nede i 0.075 for nærmeste nabo-klassifikatoren for egenskapskombinasjonen 234.

V. KONKLUSJON

Datasett 1 klassifiseres best ved bruk av en lineær klassifikator, mens den samme klassifikatoren fungerer dårlig på datasett 2. Heller ikke datasett 3 klassifiseres bra ved bruk av en lineær klassifikator. De to sistnevnte får vesentlig bedre resultater ved bruk av nærmeste nabo-klassifikatoren eller minimum feilrate-klassifikatoren. Den laveste feilraten for datasett 1 oppnås ved bruk av minste kvadraters metode på alle de fire egenskapene. Klassifisering av testobjektene til datasett 2 oppnår lavest feilrate ved bruk av nærmeste nabo-klassifikatoren på egenskapskombinasjonen 12. Til slutt gir minimum feilrate-klassifikatoren på alle de fire egenskapene minst feilrate for datasett 3.

Appendix A: Kode for reprouksjon

Koden brukt i dette prosjektet er tilgjengelig på følgende github-konto https://github.com/elinfi/FYS-STK4155/tree/master/Project2.