# Mønstergjenkjenning Prosjekt 1

Trym Nordal

November 2017

## 1 Introduksjon

I denne oppgaven har tre klassifikatorer blitt implementert og trent for klassifisering av tre datasett. For hvert datasett har nærmeste-nabo klassifikatoren blitt brukt til å estimere feilraten for alle egenskapskombinasjoner. De tre klassifikatorene nærmeste-nabo (NN), minste kvadraters metode (MKM) og minimum feilrate (MF) har blitt rangert basert på feilraten i klassifiseringen for den beste egenskapskombinasjonen i hver egenskapsdimensjon.

Programmeringsspråket som ble brukt for implementering av klassifikatorene, evalueringen og lesing fra datasett er C++ med Armadillo, et C++ bibliotek for lineær algebra.

## 2 Rangering av egenskapskombinasjoner

For hver egenskapskombinasjon i hvert datasett har nærmeste-nabo-klassifikatoren blitt brukt til å estimere feilrate. Under er egenskapskombinasjonene rangert fra lavest til høyest feilrate.

### Datasett 1:

| Egenskapskombinasjon | Feilrate |
|---|---|
| [ 1, 2, 3, 4 ] | 0.093 |
| [ 1, 2, 4 ] | 0.100 |
| [ 1, 3, 4 ] | 0.127 |
| [ 1, 2, 3 ] | 0.147 |
| [ 1, 4 ] | 0.167 |
| [ 1, 2 ] | 0.180 |
| [ 1, 3 ] | 0.193 |
| [ 2, 3, 4 ] | 0.213 |
| [ 2, 4 ] | 0.227 |
| [ 1 ] | 0.240 |
| [ 3, 4 ] | 0.300 |
| [ 2, 3 ] | 0.320 |
| [ 2 ] | 0.36 |
| [ 4 ] | 0.387 |
| [ 3 ] | 0.433 |

### Datasett 2:

| Egenskapskombinasjon | Feilrate |
|---|---|
| [ 1, 2 ] | 0.013 |
| [ 1, 2, 3 ] | 0.020 |
| [ 1 ] | 0.180 |
| [ 1, 3 ] | 0.193 |
| [ 2, 3 ] | 0.287 |
| [ 2 ] | 0.280 |
| [ 3 ] | 0.493 |

### Datasett 3:

| Egenskapskombinasjon | Feilrate |
|---|---|
| [ 2, 3, 4 ] | 0.075 |
| [ 1, 2, 3, 4 ] | 0.095 |
| [ 2, 3 ] | 0.095 |
| [ 1, 2, 3 ] | 0.100 |
| [ 1, 3, 4 ] | 0.150 |
| [ 1, 3 ] | 0.170 |
| [ 3, 4 ] | 0.190 |
| [ 1, 2, 4 ] | 0.200 |
| [ 1, 2 ] | 0.215 |
| [ 2, 4 ] | 0.240 |
| [ 1, 4 ] | 0.285 |
| [ 2 ] | 0.310 |
| [ 1 ] | 0.330 |
| [ 3 ] | 0.345 |
| [ 4 ] | 0.395 |

# 3 Rangering av klassifikatorer

## Datasett 1:

| Egenskapsdimensjon | MKM | MF | kNN |
|---|---|---|---|
| 1 | 0.187 | 0.187 | 0.240 |
| 2 | 0.113 | 0.113 | 0.167 |
| 3 | 0.093 | 0.100 | 0.100 |
| 4 | 0.073 | 0.080 | 0.093 |

Den beste klassifikatoren for dette datasettet er minste kvadraters metode med egenskapsdimensjon 4, hvor alle egenskaper er aktive. De andre klassifikatorene presterer også best ved denne egenskapsdimensjonen. Nærmeste-nabo-klassifikatoren kommer dårligst ut.

## Datasett 2:

| Egenskapsdimensjon | MKM | MF | kNN |
|---|---|---|---|
| 1 | 0.107 | 0.107 | 0.180 |
| 2 | 0.120 | 0.020 | 0.013 |
| 3 | 0.120 | 0.020 | 0.020 |

For datasett 2 ser vi at nærmeste-nabo-klassifikatoren presterer best, ved egenskapsdimensjon 2. Vi ser også at minste kvadraters metode presterer langt dårligere enn de to andre klassifikatorene.

## Datasett 3:

| Egenskapsdimensjon | MKM | MF | kNN |
|---|---|---|---|
| 1 | 0.335 | 0.225 | 0.310 |
| 2 | 0.200 | 0.200 | 0.095 |
| 3 | 0.160 | 0.130 | 0.075 |
| 4 | 0.120 | 0.070 | 0.095 |

Minimum feilrate klassifikatoren presterer best i dette datasettet, for egeskapsdimensjon 4. Minste kvadraters metode presterer ikke veldig bra, mens nærmeste-nabo får best resultat på egenskapsdimensjon 3.

# 4 Avsluttende spørsmål

## Spørsmål 1

Det er fornuftig å benytte nærmeste-nabo klassifikatoren til å finne gunstige egenskapskombinasjoner fordi nærmeste-nabo klassifikatoren ikke farger feilraten på bakgrunn av formen til desisjonsgrensen. Ved bruk av for eksempel en lineær klassifikator vil kombinasjoner med lineær desisjonsgrense komme mye bedre ut enn kombinasjoner med sirkulær desisjonsgrense.

## Spørsmål 2

I en praktisk anvendelse kan det være fornuftig å bruke en kvadratisk eller lineær klassifikator til fordel for nærmeste-nabo fordi nærmeste-nabo er en tidskrevende og prosesseringstung klassifikator. Ved klassifiksering med nærmeste-nabo må man iterere gjennom hele treningssettet for hvert nye testobjekt. For en praktisk andvendelse, hvor tid og prosessering kanskje er begrenset, vil det være fornuftig å benytte en klassifikator hvor matrisene i diskriminantfunksjonen allerede er trent opp fra treningssettet.

## Spørsmål 3

Det er lite gunstig å bruke samme datasett i trening og evaluering av en klassifikator fordi da evaluerer man med data som klassifikatoren har "sett før". Diskriminantfunksjonen er trent opp fra treningsettet og vil derfor gi ufortjent gode resultater ved evalueringen av det samme datasettet.

## Spørsmål 4

En lineær klassifikator gir dårlig resultat for datasett 2 fordi desisjonsgrensen er sirkulær. Det gjør at en klassifikator med en kvadratisk diskriminantfunksjon, som minimum feilrate, vil være godt egnet. I figuren under er egenskap 1 og 2 plottet. Objekter i klasse 1 er plottet i rødt og objekter i klasse 2 er plottet i blått. Her ser vi tydelig at desisjonsgrensen er sirkulær.