PROSJEKTOPPGAVE 1

TEK5020/9020 – MØNSTERGJENKJENNING

Høsten 2023

1 Innledning

Denne oppgaven er ment å illustrere gangen i en tenkt mønstergjenkjenningsanvendelse. Den går i korte trekk ut på å finne beste egenskapskombinasjon for et gitt datasett, konstruere ulike klassifikatorer, samt evaluere resultatene. For enkelthets skyld vil vi se på problemer med bare to klasser.

For å løse oppgaven må det skrives programvare som leser inn data, splitter datasettet i *treningssett* og *testsett*, trener en klassifikator av ønsket type på angitt egenskapskombinasjon, tester denne ved å klassifisere objektene i testsettet og estimerer feilrater. Du står fritt til å benytte det som måtte være tilgjengelig av programmeringshjelpemidler (f. eks. Matlab eller Python).

Det skal skrives en kortfattet *rapport* (maksimalt ca. 10 sider), som beskriver hva som er gjort og hvilke resultater som er oppnådd. For hvert datasett skal følgende oppgis:

1. Rangering av egenskapskombinasjoner, basert på nærmeste-nabo regelen.
2. Rangering av klassifikatorene (tre klassifikatortyper) for de beste egenskapskombinasjonene.

I begge tilfeller skal evalueringsmålene oppgis, dvs. estimerte feilrater for hver av de testede egenskapskombinasjonene og hver av klassifikatortypene som er brukt. Dessuten skal det også svares kortfattet på noen spørsmål. Legg gjerne ved ved utskrift av kode, men dette er ikke noe krav.

Oppgaven kan utføres som et gruppearbeid, der 2-3 studenter leverer en felles rapport.

2 Datasett

Datasettene som skal brukes er lagret som tekstfiler, der hver linje svarer til ett objekt. Første element i hver linje angir klassetilhørigheten (verdien 1 eller 2) og de øvrige elementene i linjen angir egenskapsverdiene knyttet til det samme objektet.

Tre datasett skal brukes i oppgaven (*Datasett_1*, *Datasett_2* og *Datasett_3*). Datafilene er vedlagt oppgaven. De to første settene er syntetiske, dvs. at de er dannet ved trekninger fra kjente tetthetsfordelinger. Det siste er generert ved uttrekking av formegenskaper fra segmenter (silhuetter) av to ulike bilmodeller. 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.

Klassifikatorene skal i denne oppgaven trenes og testes på forskjellige utvalg av objekter, dvs. hvert datasett må splittes i treningssett og testsett. For å gjøre det enkelt å sammenlikne resultater er det viktig at alle studenter bruker samme oppdeling. Hvert av datasettene skal derfor deles opp slik at treningssettet består av *odde* nummererte objekter (dvs. objektene 1, 3, 5, 7 osv.) mens de øvrige objektene (nr. 2, 4, 6, 8 osv.) plasseres i testsettet.

3 Trening av klassifikatorer

I denne oppgaven skal følgende klassifikatorer benyttes:

1. Minimum feilrate klassifikatoren med normalfordelingsantagelse,
2. Minste kvadraters metode,
3. Nærmeste-nabo klassifikatoren.

Det skal her bare gis en kort gjennomgang av disse klassifikatorene.

3.1 Minimum-feilrate klassifikatoren

Denne klassifikatoren tilordner et objekt med egenskapsvektor $\boldsymbol{x}$ til klassen ω_k dersom

$$P(\omega_k|\boldsymbol{x}) = \max_i P(\omega_i|\boldsymbol{x}), \quad i = 1, \ldots, c,$$

der c er antall klasser. Med antagelsen om normalfordelte klassebetingede tetthetsfunksjoner, fås diskriminantfunksjonene (se Duda, Hart & Stork side 41 eller avsnitt 2.9.1 i forelesningsnotatene)

$$g_i(\boldsymbol{x}) = \boldsymbol{x}^t W_i \boldsymbol{x} + \boldsymbol{w}_i^t \boldsymbol{x} + w_{i0}, \quad i = 1, \ldots, c,$$

der

$$W_i = -\frac{1}{2}\Sigma_i^{-1}, \tag{1}$$

$$\boldsymbol{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i \tag{2}$$

og

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i). \tag{3}$$

Her er Σ_i og $\boldsymbol{\mu}_i$ henholdsvis kovariansmatrisen og forventningsvektoren til klasse ω_i. Disse størrelsene er i de aller fleste tilfeller ukjente og må derfor estimeres. Her skal *maksimum-likelihood* estimatene benyttes, dvs. forventningsvektoren estimeres ved

$$\hat{\boldsymbol{\mu}}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{x}_k$$

og kovariansmatrisen ved

$$\hat{\Sigma}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}}_i)^t.$$

I disse uttrykkene er n_i antall objekter i treningssettet for klasse ω_i og $\boldsymbol{x}_k$ egenskapsvektoren til objekt k fra samme klasse. Fordi vi her bare ser på toklasseproblemer, kan

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$$

benyttes som diskriminantfunksjon. Med denne tilordnes $\boldsymbol{x}$ til ω_1 hvis $g(\boldsymbol{x}) \geq 0$ og til ω_2 ellers.

Som et ledd i prosjektoppgaven skal det lages en funksjon som trener en minimum feilrate klassifikator som skissert ovenfor, dvs. du må lage en funksjon som ut fra *treningsobjektene* beregner matrisene W_i, vektorene $\boldsymbol{w}_i$ og konstantene w_{i0} som gitt i likningene (1), (2) og (3), for i=1, 2. Siden apriorisannsynlighetene $P(\omega_i)$ som inngår i likn. (3) er ukjente her, må de estimeres på bakgrunn av antall treningsobjekter fra hver klasse i det aktuelle datasettet.

Funksjonen skal brukes til å konstruere klassifikatorer for utvalgte egenskapskombinasjoner fra alle tre datasett.

3.2 Minste kvadraters metode

Denne metoden gir en lineær klassifikator med toklasse diskriminantfunksjon

$$g(\boldsymbol{x}) = \boldsymbol{a}^t \boldsymbol{y} \tag{4}$$

der $\boldsymbol{a} = [a_0, a_1, \ldots, a_d]^t$ er den utvidede vektvektoren og $\boldsymbol{y} = [1, x_1, \ldots, x_d]^t$ en utvidet egenskapsvektor. La nå Y være en matrise av dimensjon $n \times (d+1)$ som inneholder alle de utvidede treningsvektorene lagret radvis, dvs.

$$Y = \begin{bmatrix} \boldsymbol{y}_1^t \\ \vdots \\ \boldsymbol{y}_n^t \end{bmatrix}.$$

Videre definerer vi en n-dimensjonal vektor $\boldsymbol{b} = [b_1, \ldots, b_n]^t$ der hvert element er gitt som $b_k = 1$ for $\boldsymbol{y}_k$ fra klasse ω_1 og $b_k = -1$ ellers.

Vektvektoren $\boldsymbol{a}$ velges nå slik at lengden på feilvektoren $\boldsymbol{e} = Y\boldsymbol{a} - \boldsymbol{b}$ blir minimalisert, dvs. kostfunksjonen

$$J(\boldsymbol{a}) = ||Y\boldsymbol{a} - \boldsymbol{b}||^2$$

skal minimaliseres. Dette gir løsningen

$$\boldsymbol{a} = (Y^t Y)^{-1} Y^t \boldsymbol{b}.$$

Det skal med andre ord skrives en funksjon som ut fra *treningsobjektene* beregner vektoren $\boldsymbol{a}$ for en vilkårlig kombinasjon av egenskaper, slik at klassifikatoren derved er gitt ved diskriminantfunksjonen i likning (4). Se side 240 i Duda, Hart & Stork (eller avsnitt 5.7 i forelesningsnotatene) for detaljene i metoden.

3.3 Nærmeste-nabo klassifikatoren

Dette er en programmeringsmessig enkel klassifikator, som samtidig gir gode og pålitelige klassifiseringer. For hvert objekt $\boldsymbol{x}$ som skal klassifiseres, beregnes avstanden til alle objektene i treningssettet, og $\boldsymbol{x}$ tilordnes samme klasse som det nærmeste objektet $\boldsymbol{x}_k$ der

$$||\boldsymbol{x} - \boldsymbol{x}_k|| = \min_i ||\boldsymbol{x} - \boldsymbol{x}_i||, \quad i = 1, \ldots, n.$$

En grundigere beskrivelse finnes i Duda, Hart & Stork, side 177-179 (se også avsnitt 4.2.2 i forelesningsnotatene).

4 Evaluering

Evalueringen av en klassifikator skal her foretas ved å klassifisere objektene i *testsettet*, og sammenligne klassifiseringsresultatet med den oppgitte klassetilhørigheten for objektet.

Som feilrateestimat benyttes vanligvis forholdet mellom antall feilklassifiserte objekter og det totale antall objekter i testsettet:

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

Det er dette feilrateestimatet som skal brukes i denne oppgaven.

5 Gjennomføring av oppgaven

For hvert av de tre datasettene skal du:

1. Bruke nærmeste-nabo klassifikatoren til å estimere feilraten for alle kombinasjoner av egenskaper av en gitt dimensjon (systematisk utprøving). Dette skal gjøres for alle mulige dimensjoner (d = 1, 2 ,..., antall egenskaper).

2. For beste kombinasjon innen hver mulige egenskapsdimensjon finne den beste klassifikatoren av de tre som er implementert.

6 Avsluttende spørsmål

Til slutt skal du svare kortfattet på følgende spørsmål:

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

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

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

4. Hvorfor gir en lineær klassifikator dårlige resultater for datasett 2?