PROSJEKTOPPGAVE 1

Unik4590/Unik9590/TTK 4205 - MØNSTERGJENKJENNING

6. august 2015

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, men klassifikatorene som skal brukes kan også benyttes i flerklasseproblemer.

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 både treningssettet og testsettet, og estimerer feilrater. Du står fritt til å benytte det som måtte være tilgjengelig av programmeringshjelpemidler (f. eks. Matlab).

Besvarelsen skal, foruten programlistingen, for hvert datasett bestå av:

1. Rangering av egenskapskombinasjoner
2. Rangering av klassifikatorene for beste egenskapskombinasjon.

I begge tilfeller skal evalueringsmålene oppgis. Dessuten skal du også svare kortfattet på noen spørsmal.

2 DATASETT

Datasettene du skal benytte ligger lagret som tekstfiler på ASCII-format. En datafil inneholder $n+1$ tekstlinjer der n er antall objekter i filen, og det er like mange poster i hver linje.

Første linje inneholder antall objekter og antall egenskapskandidater som er knyttet til hvert objekt. I de øvrige linjene ligger det lagret opplysninger om hvert objekt, dvs. en linje pr. objekt. Opplysningene består av klassetilhørighet i første post, etterfulgt av m egenskapskandidater, dvs i alt $m+1$ poster pr. objekt. Filformatet er illustrert i figuren nedenfor.

Linje nr.	Post 1	Post 2	Post 3	Post 4	....	Post m+1
1	ant. obj. (n)	ant. egensk.	-	-	....	-
2	klasse obj. 1	egenskap 1	egenskap 2	egenskap 3	....	egenskap m
....	....	....	....	....	....	....
n+1	klasse obj. n	egenskap 1	egenskap 2	egenskap 3	....	egenskap m

Filformat for datasettene.

Tre datasett skal brukes i oppgaven. Linker til disse filene (Datasett_1, Datasett_2 og Datasett_3) er å finne på kursets hjemmeside.

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 (Nissan Sunny og Nissan Prairie). Segmentene er generert ved bildebehandling av digitale videoopptak av kjøretøyene.

Klassifikatorene skal i denne oppgaven trenes og testes på forskjellige utvalg av objekter. Følgelig må hvert datasettet splittes i et treningssett og et 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 GENERERING 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 $\vec{x}$ til klassen ω_k dersom

$$P(\omega_k|\vec{x}) = \max_j P(\omega_j|\vec{x}), \quad j = 1,\ldots,c$$

der c er antall klasser. Med antagelser om normalfordelte klassebetingede tetthetsfunksjoner, fås diskriminantfunksjonene (se Duda, Hart & Stork side 41):

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

der

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

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

og

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

Her er Σ_i og $\vec{\mu}_i$ henholdsvis kovariansmatrise og forventningsvektor til klasse ω_i. Disse størrelsene er i de aller fleste tilfeller ukjente og må derfor estimeres. Her skal *maximum likelihood* estimatene benyttes, dvs. forventningsvektoren estimeres ved:

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

og kovariansmatrisen ved:

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

Her er n_i antall objekter i treningssettet for klasse ω_i, og $\vec{x}_k$ egenskapsvektoren til objekt k fra samme klasse. Fordi vi her bare ser på toklasseproblemer, kan $g(\vec{x}) = g_1(\vec{x}) - g_2(\vec{x})$ benyttes som diskriminantfunksjon. Med denne tilordnes $\vec{x}$ til ω_1 hvis $g(\vec{x}) \geq 0$ og til ω_2 ellers.

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

Prosedyren skal brukes til å konstruere klassifikatorer for ulike egenskapskombinasjoner fra alle tre datasett.

3.2 Minste kvadraters metode

Denne metoden gir en lineær klassifikator med toklasse diskriminantfunksjon:

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

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

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

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

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

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

skal minimaliseres. Dette gir løsningen:

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

Det skal med andre ord skrives en prosedyre som ut fra *treningsobjektene* beregner vektoren $\vec{a}$ for en vilkårlig kombinasjon av egenskaper, slik at klassifikatoren derved er gitt ved diskriminantfunksjonen i likning (4). For detaljer, se Duda, Hart & Stork, side 240.

3.3 Nærmeste nabo klassifikatoren

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

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

En grundigere beskrivelse finnes f.eks. i Duda, Hart & Stork, side 177 - 179.

4 EVALUERING

Evalueringen av en klassifikator skal her foretas ved å klassifisere objektene i datasettene. *Treningssettet* og *testsettet* skal behandles *hver for seg*. Klassifiseringsresultatene kan registreres fortløpende i "forvirringsmatriser" (confusion matrices). En forvirringsmatrise C er generelt en $c \times c$ matrise, der c er antall klasser ($c = 2$ i denne oppgaven). Komponenten C_{ij} er antall objekter fra ω_i som har blitt klassifisert til ω_j, dvs. raden i forvirringsmatrisen angir sann klassetilhørighet og kolonnen antatt (klassifisert) klassetilhørighet. Som feilrateestimat benyttes vanligvis:

$$P(e) = \sum_{i=1}^{c} P(\omega_i) P(e|\omega_i).$$

Dette estimatet kan beregnes på bakgrunn av forvirringsmatrisen. Dersom á priori sannsynlighetene beregnes som fraksjonen av objekter fra ω_i som forekommer i matrisen, og de klassebetingede feilratene som de tilsvarende fraksjoner av feilklassifiseringer innen hver klasse, er feilraten gitt ved forholdet mellom summen av de ikkediagonale komponentene i matrisen og summen av alle komponenter.

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 med 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 egenskaps-kombinasjoner?
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?