UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I:	**ST 366 – Statistisk mønstergjenkjenning**
	Del I av to deler
VED:	**Nils Lid Hjort**
TID FOR EKSAMEN:	**Del I: 19.12.1996 kl. 10:05 – 13.1.1997 kl. 16:05.**
	Del II: Foredrag basert på artikler, 17.1.1997.

Oppgavene til **Del I** deles ut torsdag 19. desember kl. 10:05, ved K. Pethon, 7. etasje, Matematisk Institutt; de blir dessuten sendt pr. elektronisk post til hver av kandidatene. Skriftlig besvarelse, helst tekstbehandlet, skal leveres samme sted, eller direkte til Nils Lid Hjort, senest mandag 13. januar 1997 kl. 16:05. Relevante figurer bør inngå i besvarelsen. Også kopier av programmer som er benyttet skal legges ved. Kandidatene skal arbeide uavhengig av hverandre.

Del II av eksamen i dette kurset består i å holde et foredrag, basert på utlevert journalartikkel, fredag 17. januar. Se separat ark for opplysninger om dette.

Dette er oppgavesettet til **Del I**. Det inneholder fire oppgaver og er på fire sider.

Del I, oppgave 1

Visse objekter kan stamme fra en av to like sannsynlige klasser, og man skal studere klassifikasjonsregler basert på egenskapsvektoren $X = (X_1, \ldots, X_d)^{\mathrm{t}}$, som kan ekstraheres fra objektene. I denne oppgaven skal vi se på noen situasjoner der klassefordelingene er kjente for statistikeren. Man kan for eksempel tenke seg at parametrene er estimert med meget god presisjon basert på store mengder treningsdata.

(a) La oss først anta at $X \sim N_d(\mu_1, \Sigma)$ hvis opphavet er klasse 1 og at $X \sim N_d(\mu_2, \Sigma)$ hvis opphavet er klasse 2. Her er altså μ_1, μ_2, Σ kjente parametre, og kovariansmatrisen Σ er positiv definit. Vis at den vanlige klassifikasjonsregelen blir som følger: Hvis

$$D(X) = (\mu_2 - \mu_1)^{\mathrm{t}}\Sigma^{-1}(X - \bar{\mu}) \begin{cases} > 0, & \text{så påstå at det er klasse 2,} \\ \leq 0, & \text{så påstå at det er klasse 1.} \end{cases}$$

Her er $\bar{\mu} = \frac{1}{2}(\mu_1 + \mu_2)$. I hvilken forstand er dette den beste metoden?

(b) Vis at suksessraten blir

$$\mathrm{pcc} = \Phi(\tfrac{1}{2}\delta),$$

der δ er Mahalanobis-avstanden $\{(\mu_2 - \mu_1)^{\mathrm{t}}\Sigma^{-1}(\mu_2 - \mu_1)\}^{1/2}$ og der Φ er den kumulative standard-normalfordelingsfunksjonen.

(c) Anta nå at $D(X)$-regelen over brukes i en situasjon der de to klassefordelingene er henholdsvis $N_d(\mu_1, \Sigma_1)$ og $N_d(\mu_2, \Sigma_2)$, med $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$. (Dette svarer jo til et fellesestimat for kovariansmatrisen basert på et stort treningssett med like mange fra hver av klassene.) Generaliser resultatet fra (b) ved å finne en formel for suksessraten pcc i denne situasjonen.

(d) La så de to klassetetthetene være henhodsvis

$$N_2(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}) \quad \text{og} \quad N_2(\begin{pmatrix}1.3\\1.3\end{pmatrix}, \begin{pmatrix}3.61 & 0.38\\0.38 & 0.16\end{pmatrix}).$$

Finn først den beste lineære klassifikasjonsregel, og bestem dens suksessrate. Konstruer så den optimale regelen og bestem dennes suksessrate. Tegn opp beslutningsområdene i en figur, både for den beste lineære og for den optimale metoden, og kommenter.

Oppgave 2

Anta at de sanne fordelingene er som i punkt (d) i forrige oppgave, men, nå mer realistisk, at parametrene i de to multinormale fordelingene er ukjente. Altså skal de estimeres fra treningsdata, si $X_{1,1}, \ldots, X_{1,n}$ fra klasse 1 og $X_{2,1}, \ldots, X_{2,n}$ fra klasse 2. Følgelig står man overfor et klassisk valg mellom to modeller, den multinormale med lik kovariansmatrise og den multinormale med forskjellige kovariansmatriser.

(a) Sett nøyaktig opp klassifikasjonsregler som resulterer fra disse to modellene. Sett så opp et modellvalgskriterium basert på AIC, Hirotugu Akaikes *johoryo-tokeigaku*.

(b) Er det andre modellvalgskriterier du kunne tenke deg å anvende her?

(c) Spørsmålet om det er den empiriske lineære eller den empiriske kvadratiske regelen som er best avhenger av sample-størrelsen. Hvilken tror du vinner for lav n, og hvilken vinner for høy n? Gi korte begrunnelser. Stemmer dette med det modellvalgsråd du får fra AIC?

(d) Statistiker-ambisjonene bør jo være høyere enn å kunne synse om 'lav n' og 'høy n'. Prøv å finne noe mer informativt om hvordan suksessratene til de to metodene avhenger av n, og prøv å finne ut for hvilke n den forventede suksessrate blir bedre med den kvadratiske metoden.

Del I, oppgave 3

Denne oppgaven skal handle om aspekter ved treningsmengdens størrelse i forhold til dimensjonen på egenskapsvektoren. Det er er klart at større dimensjon avkrever relativt sett flere treningsdata for rimelig estimering av parametre i klassefordelingene. Altså vil det finnes situasjoner der det ikke lønner det seg å 'måle for mye' hvis man har et fast antall objekter å trene på.

For å illustrere dette ser vi på et idealisert skrivebordseksempel som følger. Man har objekter fra hver av to a priori like sannsynlige klasser, og man kan utføre så mange målinger man vil på dem. Resultatet er vektoren $X = (X_1, \ldots, X_d)^{\mathrm{t}}$, der klassefordelingene er henholdsvis

$$f_1 = N_d(0, I_d) \quad \text{og} \quad f_2 = N_d(\mathbf{m}_2, I_d),$$

hvor $\mathbf{m}_2 = (\mu_1, \ldots, \mu_d)^{\mathrm{t}}$ har komponenter $\mu_j = 1/\sqrt{j}$. (En mer realistisk situasjon ville nok være at nye målinger ble mer og mer lineært avhengige av de tidligere. Situasjonen her kan man se på som en transformert versjon av dette, der man altså stadig kan få tak i nye uavhengige målinger, men der den relative nytte av disse blir mindre og mindre.)

(a) Først ser vi på den ideelle klassifikasjonsregel der parametrene er kjente og ikke trenger treningsdata for å bli estimert. Sett opp denne, og vis at dens feilrate

$$\varepsilon_0 = \varepsilon_{0,d}$$

avtar monotont mot null når d vokser. Sett også opp en passende approksimasjon for denne feilraten.

I en realistisk situasjon er imidertid feilraten resultatet av et mer nyansert samspill mellom dimensjon d og treningssettets størrelse. Anta at senter-parametrene $\mathbf{m}_1$ og $\mathbf{m}_2$ er estimert fra treningsdata, $X_{1,1}, \ldots, X_{1,n}$ fra klasse 1 og $X_{2,1}, \ldots, X_{2,n}$ fra klasse 2. For enkelhets skyld antar vi at kovariansmatrisen er kjent, og altså lik I_d. (Resultater som ligner dem som kommer under vil kunne oppnås i den mer realistiske situasjon der også kovariansmatrisen Σ blir estimert.)

(b) Vis at den relevante estimerte utgave av den beste lineære regelen blir som følger: Hvis

$$\widehat{D}(X) = (\widehat{\mathbf{m}}_2 - \widehat{\mathbf{m}}_1)^{\mathrm{t}}(X - \tfrac{1}{2}(\widehat{\mathbf{m}}_1 + \widehat{\mathbf{m}}_2)) \begin{cases} > 0, & \text{så påstå at det er klasse 2,} \\ \leq 0, & \text{så påstå at det er klasse 1.} \end{cases}$$

Finn betinget forventning og varians for $\widehat{D}(X)$ gitt treningssettet, når X kommer fra henholdsvis klasse 1 og klasse 2.

(c) Finn så eksplisitte uttrykk for ubetinget forventning og varians for $\widehat{D}(X)$, for X fra henholdsvis klasse 1 og klasse 2. Bruk dette til å vise at den forventede feilrate

$$\varepsilon_n = \varepsilon_{n,d}$$

faktisk konvergerer mot det verst tenkelige resultat, nemlig $\frac{1}{2}$, når d vokser; 'signalet drukner i støy'.

(d) Illustrer med en figur hvordan den teoretiske $\varepsilon_{0,d}$ og den realistiske $\varepsilon_{n,d}$ oppfører seg som funksjon av d, og kommenter.

Oppgave 4

Denne siste oppgaven er en utfordring av mer praktisk art! Dataene er hentet fra en anvendelse Norsk Regnesentral arbeidet med for noen år siden for et kartografifirma. Problemet er å gjenkjenne håndskrevne tall automatisk. Passende preprosesseringsalgoritmer sørger for at hvert kandidatsymbol blir isolert fra andre og representert på en bestemt vektoriell form. Ut fra denne representasjonen kan man så bestemme seg for å trekke ut ulike egenskapstall. En viktig del av slike prosjekter er å finne slike atributter med stor diskriminerende evne.

Dataene vi skal arbeide med her er hentet fra 200 'null'-symboler, 200 'tre'-symboler og 200 'åtte'-symboler. For hvert symbol er det beregnet en vektor $X = (X_1, X_2, X_3)^{\mathrm{t}}$ basert på visse lengder av kandidatsymbolet innenfor visse delruter av det rektangel som omslutter det. Dataene foreligger i tre filer, og er av typen vist under. Disse tre datafiler vil bli emailet til hver av kandidatene

`kjetilka@math.uio.no, jornma@math.uio.no, bhm@math.uio.no .`

Oppgaven består nå i å lage noen gode klassifikasjonsalgoritmer for dette problemet! For hver metode du lager, gi en kort begrunnelse og motivasjon. Man kan gjerne prøve flere ulike metoder, men man bør innskrenke seg til å rapportere om dem som synes mest lovende. Man skal også redegjøre for kvaliteten av de diskriminantregler som foreslås.

Fotnote 1: Diskrimineringsoppgaven var i det virkelige prosjekt lettere enn det man kan få inntrykk av her. Man kunne trekke inn ytterligere karakteristika som hadde større diskriminerende suksess enn de tre målinger som foreligger her. Den aktuelle og altså slett ikke optimale (X_1, X_2, X_3) er valgt av pedagogiske grunner.

Fotnote 2: Sensor vil ha en god mekanisme for å vurdere hvilke løsningsforslag som er gode og hvilke som kanskje ikke er det. Data finnes over ytterligere 3×200 symboler, 'null', 'tre' og åtte' (men holdes altså skjult for kandidatene). Altså vil det bli interessant å utføre klassifikasjon på dette uavhengige testsettet med de algoritmer kandidatene foreslår!

Fotnote 3: Dataene er altså av typen under:

```
# These are three-dimensional feature vector data
# for each of 200 handdrawn "zero" symbols.
   1   69.30 128.00 120.8
   2   70.70 114.50 114.3
   3   67.97 129.40 132.2
   4   67.20 138.40 129.5
   5   71.40 127.00 128.9
   6   68.45 132.80 122.1
   7   70.89 109.40 119.6
   8   69.15 125.60 120.5
   9   73.90 108.90 126.8
  10   68.67 116.40 133.6
   ...
   And so on.  [Kurt Vonnegut.]
   ...
```

En slik fil kan leses inn i S-Plus som følger:

```
fil0data <- matrix( scan("tall0", skip=2), ncol=4, byrow=T)
tall0data <- fil0data[ ,2:4]
```

QUANTUM SATIS