UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I:	**ST 366 – Statistisk mønstergjenkjenning, høstsemesteret 1991**
VED:	*Nils Lid Hjort*
FOR:	*Dr. Scient. og Cand. Scient.-studenter*
TID FOR EKSAMEN:	*Utlevering mandag 27. januar 1991 kl.* 9^{15}, *innlevering av besvarelse fredag 7. februar 1991 kl.* $\leq 16^{00}$ *[samme dag som kandidatene skal legge frem de utdelte journalartikler for eksamenskommisjonen].*

Oppgavesettet inneholder fem oppgaver, og er på fem sider (inkludert en side med data).

GENERELT: *Pensum og annen orientering om eksamensprosessen i dette kurset er på eget ark. Journalartiklene som skal fremlegges fås av Nils Lid Hjort. Kandidatene kan bruke alle hjelpemidler, men skal arbeide uavhengig av hverandre. Besvarelsen skal gjerne være pent tekstbehandlet.*

Oppgave 1

Visse objekter skal gjenkjennes automatisk, så godt det lar seg gjøre. Hvert objekt er av type 1, eller type 2, eller type 3, og disse tre kategorier er like sannsynlige a priori. En bestemt tellemåling $X \in \{0, 1, 2, \ldots\}$ lar seg registrere med et apparat, for hvert objekt, og tidligere eksperimenter har godtgjort at X er Poisson-fordelt med parameter $\theta_1 = 10$ hvis objektet er av type 1, Poisson-fordelt med parameter $\theta_2 = 20$ hvis objektet er av type 2, og Poisson-fordelt med parameter $\theta_3 = 30$ hvis objektet er av type 3.

(a) Finn klassifikasjonsprosedyren som oppnår minst mulig feilsannsynlighet.

(b) Beregn de klassebetingede suksessratene pcc(1), pcc(2), pcc(3), samt den totale suksessrate pcc [som står for 'probability of correct classification'].

(c) Finn generelt affiniteten $\sum_{x \geq 0} \{f_1(x) f_2(x)\}^{1/2}$ mellom to Poisson-fordelinger, og beregn deretter de generaliserte Mahalanobis-avstandene $\omega_{1,2}$, $\omega_{1,3}$, $\omega_{2,3}$ mellom de forskjellige par av klasser i den aktuelle situasjon. [Se Section 10.2.C i Hjort (1986).]

(d) For en viss ekstra utgift er det mulig å observere en tellemåling Y i tillegg til X for hvert objekt, der Y er uavhengig av X og har samme fordeling. Hvis (X, Y) stammer fra et objekt av type 1 er de uavhengige og begge Poisson-fordelte med parameter $\theta_1 = 10$, for eksempel. Finn ut hvor meget de generaliserte Mahalanobis-avstander forandrer seg hvis Y også anvendes.

(e) Finn også ut hva de nye suksessratene blir, for den optimale metode som bruker både X og Y.

(f) Foreta eventuelle rimelige generaliseringer, og gi (korte) tilleggskommentarer du måtte finne relevante.

Oppgave 2

I foregående oppgave var det antatt at klassefordelingene var eksplisitt kjente. Forutsett nå at det bare anses kjent at X er Poisson-fordelt for hver klasse, men at de tre parametrene θ_1, θ_2, θ_3 er ukjente. Anta videre at det foreligger fem X-målinger som garantert er av type 1, fem X-målinger som garantert er av type 2, og fem X-målinger som garantert er av type 3.

(a) Skriv opp en rimelig 'plug-in' klassifikasjonsprosedyre som bygger på disse treningsdataene.

(b) Gjør også rede for hvordan en prediktiv klassifikasjonsprosedyre ser ut, (i) når informasjon om θ-ene foreligger i form av Gamma-apriorifordelinger, (ii) når den noninformative apriorifordelingen θ^{-1} blir brukt for hver klasse.

(c) Gi noen kommentarer om forskjellen på løsningene fra (a) og (b).

Oppgave 3

[Hvis denne oppgaven faller deg vanskelig på grunn av manglende øvelse og fortrolighet med det multivariable, kan du nøye deg med å gi en en-dimensjonal løsning, eller en løsning der Σ-matrisen under er på diagonalform.]

La oss anta at egenskapsvektoren $X = (X_1, \ldots, X_d)'$ kan stamme fra en av to like sannsynlige klasser, og at $X \sim \mathcal{N}_d\{\mu_1, \Sigma\}$ hvis opphavet er klasse 1 og at $X \sim \mathcal{N}_d\{\mu_2, \Sigma\}$ hvis opphavet er klasse 2. Her er μ_1, μ_2, Σ foreløpig kjente parametre (basert på store mengder treningsdata, kan vi tenke oss), og kovariansmatrisen Σ er positiv definit.

(a) Vis at den vanlige klassifikasjonsregelen blir som følger: Hvis

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

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

(b) Tegn opp beslutningsområdene for denne metoden, i det enkle eksempel der data er todimensjonale, $\mu_1 = (0, 0)'$, $\mu_2 = (2, 2)'$, og Σ er proporsjonal med identitetsmatrisen.

(c) Vis at suksessraten for klasse 1 blir

$$\text{pcc}_{\text{ideal}}(1) = \Phi(\tfrac{1}{2}\delta), \qquad \text{der } \delta = \{(\mu_2 - \mu_1)'\Sigma^{-1}(\mu_2 - \mu_1)\}^{1/2}$$

og der $\Phi(.)$ er den kumulative standard-normalfordelingen. Hva blir den ubetingede suksessrate her?

(d) Anta så at senterparametrene μ_1 og μ_2 er ukjente, og estimeres som $\bar{X}_1$ og $\bar{X}_2$ basert på n observerte vektorer fra klasse 1 og n observerte vektorer fra klasse 2, mens Σ fortsatt er kjent. Finn et eksplisitt uttrykk for pcc(1|data), den betingede suksessrate for klasse 1 gitt treningsdata, når regelen som brukes er plug-in-versjonen av den fra (a).

(e) Finn også et eksplisitt uttrykk for pcc(1), den *ubetingede* suksessrate for klasse 1.

(f) Gi relevante (og korte) tilleggskommentarer.

Oppgave 4

[Som for forrige oppgaves vedkommende kan du forsåvidt nøye deg med å gi en løsning for det en-dimensjonale tilfellet, hvis det multivariable blir for vanskelig, og hvis du finner det praktisk kan du gjerne først skrive ned løsninger for dette enklere tilfellet og så gå til det generelle d-dimensjonale tilfellet.]

Vi skal se på en enkel modell for å utnytte 'spatial kontinuitet' i diskriminantanalyse. En prøvedrilling i forbindelse med oljeleting gir muligheten for å observere diverse målinger $X_i = (X_{i,1}, \ldots, X_{i,d})'$ for hvert 'dyp' i, for eksempel hver 25. centimeter. Man vil gjerne kunne gjenkjenne den underliggende geologi ut fra disse indirekte målingene, med andre ord klassifisere X_i til den korrekte geologiske klasse C_i.

For enkelthets skyld skal vi her forutsette at det bare er to slike klasser, si $C_i \in \{1, 2\}$, at de er a priori like sannsynlige, og at X_i enda en gang er multinormal (μ_1, Σ) hvis $C_i = 1$ og multinormal (μ_2, Σ) hvis $C_i = 2$. Og inntil videre er disse klassebeskrivelsesparametrene kjente.

(a) Den enkle og nonkontekstuelle diskriminantregelen er som i 3(a): Dersom

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

Hva er denne regelens suksessrate, hvis avstanden mellom klassene δ er lik 1, 2, 4, 6, 8?

(b) Geologene tenker seg at C_i-prosessen har en rimelig 'romlig kontinuitet', i den forstand at geologien i naboposisjoner oftere er lik enn ulik. Hvis homogene tripler (C_{i-1}, C_i, C_{i+1}) er hyppige kan man midle målinger for å få mindre støy. Dette motiverer denne kontekstuelle regelen: Hvis

$$G(X_i) = (\mu_2 - \mu_1)'\Sigma^{-1}(\bar{X}_i - \bar{\mu}) \begin{cases} > 0, & \text{så påstå at } C_i \text{ er klasse 2,} \\ \leq 0, & \text{så påstå at } C_i \text{ er klasse 1.} \end{cases}$$

Her er $\bar{X}_i = \frac{1}{3}(X_{i-1} + X_i + X_{i+1})$. Vi skal anta at X_j-ene er betinget uavhengige gitt underliggende geologisk prosess. Hvilken fordeling har $\bar{X}_i$, for gitte underliggende klasser (C_{i-1}, C_i, C_{i+1})?

(c) Finn suksessraten $\text{pcc}_{1,1,1}(1)$ for klasse 1 for denne regelen, for den fordelaktige situasjon der det er gitt at (C_{i-1}, C_i, C_{i+1}) er et homogent trippel (ikke bare er $C_i = 1$, men de to naboene er også lik 1). Kommenter svaret.

(d) Og finn et uttrykk for suksessraten $\text{pcc}_{2,1,2}(1)$ for klasse 1 for denne regelen, for den minst fordelaktige situasjon der det er gitt at $C_{i-1} = 2 = C_{i+1}$ mens $C_i = 1$.

(e) Anta som en illustrasjon at

$$\begin{aligned} \Pr\{C_{i-1} = 1, C_{i+1} = 1 | C_i = 1\} &= 0.80, \\ \Pr\{C_{i-1} = 1, C_{i+1} = 2 | C_i = 1\} &= 0.09, \\ \Pr\{C_{i-1} = 2, C_{i+1} = 1 | C_i = 1\} &= 0.09, \\ \Pr\{C_{i-1} = 2, C_{i+1} = 2 | C_i = 1\} &= 0.02. \end{aligned}$$

Undersøk kort om den kontekstuelle metoden vinner over den nonkontekstuelle.

(f) Spekuler kort over andre muligheter for å lage kontekstuelle klassifikasjonsregler. Tenk deg at oljeselskapet har treningsdata av formen $(C_1, X_1), (C_2, X_2), \ldots, (C_n, X_n)$ for rimelig stor n.

Oppgave 5

Vedlagt er et utsnitt av et større datasett som handler om automatisk statistisk gjenkjenning av håndskrevne tall. To variable (X, Y) registreres automatisk for hver symbolkandidat. 25 slike par er listet opp for symbolet `'0'` og 25 for symbolet `'8'`.

Din oppgave er å motivere og lage en diskriminantmetode som skal prøve å skille mellom `'0'` og `'8'`. Du bør lage den så eksplisitt som mulig; tenk deg at du skal bruke algoritmen din på et par tusen nye objekter etterpå. Og du skal gjerne kjøre metoden på treningssettet.

Det er lov til å prøve mer enn en metode, men bare hvis du har en rimelig motivasjon, og skriv konsist.

Hva slags feilrater ser det ut til at metoden din gir?

[Dataene er hentet fra en større treningsmengde fra et prosjekt på Norsk Regnesentral. Diskrimineringen mellom `'0'` og `'8'` er egentlig meget lettere ved bruk av tilleggsvariable, men poenget her er å kunne illustrere og sammenligne flere metoder i en situasjon der perfekt diskriminering ikke er mulig.]

Signe Arbeidet