

Uppsala universitet
Signaler and System
MG 018 - 471 3229

Tentamen i datoriserad mönsterigenkänning

för civilingenjörsprogrammen i teknisk fysik och informationsteknologi samt för matematiskt naturvetenskapligt program.

Plats och Tid: Post Scriptum 2, tisdag 20 augusti, 2002, 9.00-14.00

Hjälpmedel: Formelsamling på kursens hemsida.

Obs! Svar utan motivering ger noll poäng!

Problem #1 (6p=3p+2p+1p)

Antag att sannolikhetstäthetsfunktionerna för två lika troliga klasser ω_1 och ω_2 ges av

$$p(\mathbf{x}|\omega_1) = \begin{cases} 1/9 & \text{om } 1 \leq x_1 \leq 4 \text{ och } 1 \leq x_2 \leq 4 \\ 0 & \text{annars} \end{cases}$$

och

$$p(\mathbf{x}|\omega_2) = \begin{cases} 1 & \text{om } 2 \leq x_1 \leq 3 \text{ och } 2 \leq x_2 \leq 3 \\ 0 & \text{annars} \end{cases}$$

- Rita beslutsgränser för den optimala klassificerare som minimerar antalet klassificeringsfel (i medeltal).
- Bestäm den optimala klassificerare som minimerar antalet klassificeringsfel (i medeltal).
- Beräkna klassificeringsfelet för den optimala klassificeraren som minimerar antalet klassificeringsfel (i medeltal).

Problem #2 (6p=3p+2p+1p)

- Civ har 40 designexempel och vill bygga en linjär regressionsmodell. Hon bygger med hjälp av OLS två olika modeller, en med hjälp av de 20 första exemplen, en annan med hjälp av de 20 sista. Modellerna blir dock helt olika! Ge en trolig förklaring till detta fenomen och ge förslag på två metoder som kanske kan lösa detta problem. Förklara också med några få meningar varför dessa två metoder troligen är bättre än OLS i detta fall.
- Förklara hur man stegvis designar en linjär klassificerare med hjälp av minsta kvadratmetoden. Antag att de två klasser som man försöker designa en linjär

klassificerare för är linjärt separabla. Varför är det troligen bättre att använda perceptronregeln (istället för minsta kvadratmetoden) i detta fall?

c) Motivera med hjälp av en figur ekvationen för ett hyperplan med normalvektor $\mathbf{n}$ som passerar genom punkten $\mathbf{x}_o$.

Problem #3 (6p=3p+2p+1p)

Antag att N fyrdimensionella vektorer $\mathbf{x}_i, i = 1, 2, 3, \dots$ ska överföras via en smalbandig teleledning. På grund av den begränsad kapaciteten måste varje vektor reduceras till två (reella) tal innan överföring (vi bortser här från problemet med binär representation av reella tal). Med hjälp av ett antal designexempel estimeras kovariansmatrisen för de fyrdimensionella vektorerna och resultatet är egenvektorerna $\mathbf{w}_1 = [0.78, -0.61, 0, 0]^T$, $\mathbf{w}_2 = [0.61, 0.78, 0, 0]^T$, $\mathbf{w}_3 = [0, 0, 1, 0]^T$ och $\mathbf{w}_4 = [0, 0, 0, 1]^T$ med egenvärden $\lambda_1 = 0.44$, $\lambda_2 = 4.56$, $\lambda_3 = 4.0$, $\lambda_4 = 1.0$. OBS! Egenvektorerna är inte ordnade efter storleken på egenvärdena!

a) Antag att den vektorn $\mathbf{x} = [1, 0, 1, 0]^T$ står på tur att överföras. Bestäm de två tal som får representera vektorn $\mathbf{x}$ vid överföringen så att rekonstruktionen på mottagarsidan (där man har tillgång till alla egenvektorer och egenvärden) blir så bra som möjligt (i minsta kvadratmening). Bestäm också hur den rekonstruerade vektorn $\hat{\mathbf{x}}$ kommer att se ut på mottagarsidan. Bestäm till sist också kvadraten på rekonstruktionsfelet: $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ för denna vektor.

b) Förklara med hjälp av storleken på egenvärdena varför det är väntat att rekonstruktionen i deluppgift a) blev så dålig. Förklara också varför resultatet inte skulle bli speciellt bra även om man tillåter att tre istället för två tal får överföras för varje bild.

c) Varför kan man (under antagande om normalfördelning) påstå att det är ganska oväntat att man ska observera vektorn $\mathbf{x} = [1, 0, 1, 0]^T$ dvs att denna vektor nästan kan betraktas som en outlier?

Problem #4 (6p=3p+2p+1p)

Antag att vi har en diskret fördelning där data genereras med samma sannolikhet från fyra olika punkter. Studera de två fallen I: punkterna är (1,1), (2,2), (-1,-1) och (-2,-2) och II: punkterna är (1,-1), (2,2), (-1,1) och (-2,-2).

a) Rita först två figurer där punkterna för respektive (punkt-)fördelning är utritade och bestäm fördelningarnas väntevärden och kovariansmatriser.

b) Bestäm egenvektorer till respektive kovariansmatris och normaliserade egenvektorerna i figuren från deluppgift a). Uttryck sedan de fyra diskreta punkterna i den nya basen som utgörs av egenvektorerna.

c) Plocka bort den komponent som svarar mot minsta egenvärdet. Bestäm felet man gör då man representerar data endast med hjälp av den kvarvarande egenvektorn i fallen I och II.

Problem #5 (6p=2p+2p+2p)

Längs en avlägsen väg har Vägverket satt upp en mätstation som mäter vindhastighet v och vägutemperatur T . Tyvärr har radiolänken blivit feldimensionerad och kan endast överföra hälften av den tänkta datamängden. Istället för att mäta hälften så ofta frågar man om du skulle kunna "sätta ihop någon smart grunka" som kan komprimera datamängden till hälften utan att felet blir alltför stort. Du bestämmer dig för att försöka med PCA och Karhunen-Loeve-transformen. Din kompis Civ har precis gått en kurs i Meteorologi och vet att man kan anta att storheterna är normalfördelade kring ett väntevärde som är dygns- och säsongsberoende. Civ har också skattat normalfördelningarna och kovariansen mellan v och T samt sammanställt en väntevärdestabell åt dig från vägverkets "Meteorologiska data längs avlägsna vägar -98".

$$\begin{aligned}\text{Cov}[v, v] &= r_v = 5, & \text{Cov}[T, T] &= r_T = 2 \\ \text{Cov}[v, T] &= r_{vT} = 2\end{aligned}$$

1. Beräkna K-L-expansionen, d v s visa hur v och T kan uttryckas med hjälp av egenvektorer till kovariansmatrisen.
2. Bestäm transformationerna för komprimering resp rekonstruktion. Hur stort blir det kvadratiske felet (i medel)?
3. Vid ett tillfälle uppmättes vindstyrkan 2.1 m/s och temperaturen 7.1 °C. Väntevärdena var vid denna tidpunkt 2.55 m/s resp 6.2 °C. Visa tydligt hur denna mätning komprimeras och rekonstrueras. Hur stort blir rekonstruktionsfelet (MSE)? Rita en figur!

Problem #6 (6p=2p+2p+2p)

Du skall konstruera en klassificerare som kan skilja på objekt som är lika vanliga och härrör från två olika klasser (c_1 och c_2). För detta ändamål har du tillgång till sensorer som mäter två värden per objekt och du vet att mätsignalerna är normalfördelade.

1. Ange a priori-sannolikheten för respektive klass.
2. Beräkna diskrimantfunktionen för respektive klass.
3. Vilken ekvation beskriver den optimala beslutsytan? Skriv på formen

$$\mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x} + \mathbf{c} = 0 \quad (1)$$

4. Använd följande väntevärden och kovariansmatriser

$$\mu_1 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}, \quad \Sigma = \Sigma_1 = \Sigma_2 = \frac{1}{2} \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix} \quad (2)$$

Beräkna den optimala beslutsytan och ange den explicit i de ingående variablerna (x_1 och x_2 eller x och y). Vilken geometrisk form har den? Rita noggrant en figur där bl a nivåytor och beslutsytor ingår med korrekta skalor. Kovariansmatrisen Σ har följande egenvärden och egenvektorer

$$\lambda_1 = 4, \quad \lambda_2 = 1, \quad \mathbf{v}_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad (3)$$

Problem #7 (6p)

Antag att mönster $\mathbf{x}$ genereras från en Gaussian mixture (Gaussisk mixtur),

$$f(\mathbf{x}|\mathbf{w}) = \sum_{j=1}^J P(c_j) f(\mathbf{x}|c_j, \mathbf{w}_j)$$

- Beskriv en algoritm i form av ett antal punkter eller i form av pseudokod som visar hur man genererar exempel från en sådan fördelning. Du kan anta att du har tillgång till en slumpgenerator och till en funktion $\text{gauss}(\mathbf{m}, \mathbf{C})$ som genererar ett exempel från en normalfördelning med medelvärde $\mathbf{m}$ och kovariansmatris $\mathbf{C}$.
- Vid ett experiment där 10 mönster genererades visade det sig att alla exempel genererades från samma normalfördelning $f(\mathbf{x}|c_j, \mathbf{w}_j)$. Hur stor är sannolikheten för denna händelse?
- Exempel genereras med samma sannolikhet från två normalfördelningar med medelvärden $+1$ och -1 , båda har variansen ett. Beräkna felsannolikheten för en klassificerare som har beslutsgräns vid 0.1. Ledning: Utnyttja tabell för att beräkna aktuell integral.

Problem #8 (6p=3p+3p)

Vid ett klassificeringsproblem visade sig mönstervektorernas komponenter x_i , ($i = 1, 2, \dots, N$) vara helt statistiskt oberoende. Vidare var de binära dvs $x_i = 1$ eller $x_i = 0$. Mönstren tillhörde alltid en av två klasser c_1 och c_2 . a) Visa att en diskriminantfunktion som minimerar antalet felklassningar är

$$g_i(\mathbf{x}) = \sum_{k=1}^N \log P(x_k|c_i) + \log P(c_i)$$

där $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$ och $P(x_k|c_i)$ betecknar den betingade sannolikheten att erhålla komponent x_k då data genereras från klass c_i . $P(c_i)$ är sannolikheten att data genererats från klass c_i .

b) Antag att

$$P(x_k|c_1) = \begin{cases} p_{k1} & \text{då } x_k = 1 \\ q_{k1} & \text{då } x_k = 0 \text{ (dvs } q_{k1} = 1 - p_{k1}) \end{cases}$$

$$P(x_k|c_2) = \begin{cases} p_{k2} & \text{då } x_k = 1 \\ q_{k2} & \text{då } x_k = 0 \text{ (dvs } q_{k2} = 1 - p_{k2}) \end{cases}$$

för $k = 1, 2, \dots, N$. Visa att diskriminantfunktionen $g_i(\mathbf{x})$ kan skrivas som en linjär funktion av komponenterna x_k dvs att det finns en linjär diskriminantfunktion som kan användas för optimal klassificering. Ledning: Uttryck först varje $P(x_k|c)$ på formen $P(x_k|c) = p^{a(x_k)}q^{b(x_k)}$ där $a(x) = a_1x + a_0$ och $b(x) = b_1x + b_0$ är linjära funktioner. Välj koefficienterna a_0, a_1, b_0 och b_1 på lämpligt sätt!