Uppsala universitet
Signaler and System
MG 018 - 471 3229

Tentamen i mönsterigenkänning

Teknisk fysik, 3B, 3E, 3Sy Postscriptum, lördag 19 augusti, 2000, 9.00-15.00

Hjälpmedel: Mathematical Handbook, Mathematics Handbook (BETA), räknare formelsamling utdelad på lektion.

Obs! Svar utan motivering ger noll poäng!

Problem #1 (1+2p)
Antag att x är en stokastisk variabel med väntevärde $m_x = E\{x\}$ och varians $\sigma^2 = E\{(x - m_x)^2\}$. Antag vidare att det finns N realiseringar x_i, $i = 1, 2, \ldots, N$ och att väntevärdet estimeras som

$$\hat{m}_x = \frac{1}{N} \sum_{n=1}^{N} x_n$$

a) Visa att $E\{\hat{m}_x\} = m_x$.
b) Visa att variansen för estimatet kan skrivas

$$E\{(\hat{m}_x - m_x)^2\} = \frac{\sigma^2}{N}$$

Antag att man önskar att estimatets varians skall vara mindre än 0.01. Hur många exempel N behöver man för att detta ska gälla om $\sigma = 0.2$?

Problem #2 (1+3+1+1p=6p)
Antag att två-dimensionella exempel genereras från modellen

$$\mathbf{x} = [x_1 \quad x_2]^T = [a \quad a^2 - 1]^T$$

där a är en normalfördelad stokastisk variabel med väntevärde noll och varians ett. T betecknar transponat så mönstren är (som vanligt) kolumnvektorer.
a) Bestäm väntevärdet för $\mathbf{x}$.
b) Bestäm kovariansmatrisen för $\mathbf{x}$. Ledning: $E\{a^4\}$=3.

c) Avgör med hjälp av kovariansmatrisen möjligheterna att göra en linjär komprimering av mönster genererade från denna modell.
d) Rita en figur som visar hur exempel dragna från modellen fördelar sig i $x_1 x_2$-planet.

Problem #3 (3p)
Ett fundamental problem med Bayesiansk klassificering är att man måste ha tillgång till sannolikhetstäthetsfunktioner och sannolikheterna för varje klass. Eftersom täthetsfunktionerna sällan är kända antar man i praktiken en parametrisk form för den okända fördelningen och parametrarnas värden estimeras med hjälp av exempel i en sk träningsmängd.
a) En standardmetod för parameterestimering är maximumlikelihood-metoden. För t.ex. en normalfördelning är maximum-likelihoodskattningen det aritmetiska medelvärdet. Antag att det finns ett fåtal sk outliers i träningsmängden. Varför är dessa farliga i samband med parameterestimeringen? Exemplifiera med hjälp av dina erfarenheter från datorlaboration.
b) Ge minst en tänkbar förklaring till hur sk outliers uppstår i en träningsmängd.
c) Ge minst ett förslag på hur man kan reducera problemet med outliers.

Problem #4 (2p)
Som alternativ till att minimera medelkvadratfelet kan man vid logistic regression istället försöka minimera den sk korsentropin $E(\mathbf{w})$ som för P exempel definieras som

$$E(\mathbf{w}) = \sum_{p=1}^{P} d_p \log(1/y_p) + (1 - d_p)\log(1/(1 - y_p))$$

Här är som vanligt d_p önskad output för mönster $\mathbf{x}_p$ och y_p är nuvarande output. Bestäm en ekvation för viktsuppdatering (inlärning) som söker ett minimum för detta kriterium med hjälp av gradientmetoden (steepest descent).

Problem #5 (2+2=4p)
Förbehandling av data är en viktig del vid design av ett system för mönsterigenkänning. En typ av förbehandling är att hantera saknade värden (missing data).
a) Presentera minst två rimliga förklaringar till varför enkilda variabelvärden (särdarg) ibland saknas.
b) Presentera minst två metoder för hur saknade värden kan fyllas i.

Problem #6 (2+1+2=5p)
Den (differentiella) entropin $h(\mathbf{y})$ för en N-dimensionell stokastisk variabel $\mathbf{y}$ (en familj av mönstervektorer) är ett mått på hur “utsmetad” sannolikhetstäthetsfunktionen $f_y(\mathbf{y})$ är eller hur osäker man är om var exempel dragna från fördelningen

hamnar i mönsterrummet. Entropin definieras som

$$h(\mathbf{y}) = -\int_{-\infty}^{\infty} f_y(\mathbf{z}) \log[f_y(\mathbf{z})] d\mathbf{z} = -E\{\log[f_y(\mathbf{y})]\}$$

Antag $\mathbf{y} = W^T\mathbf{x}$ där $\mathbf{x}$ normalfördelad med kovariansmatris $C_{xx} = U\Lambda U^T$ och $E\{\mathbf{x}\} = \mathbf{0}$. Kolumnnerna i U är egenvektorer till $C_{xx} = E\{\mathbf{x}\mathbf{x}^T\}$, diagonalelementen i diagonalmatrisen Λ är motsvarande egenvärden-
a) Visa att $h(\mathbf{y}) = \frac{1}{2}\log(|W^T C_{xx} W|) + c$ där c är en konstant som är oberoende av W. Ledning: $E\{\mathbf{y}^T E\{\mathbf{y}\mathbf{y}^T\}\mathbf{y}\}$=N.
b) Visa att om W är kvadratisk och ortonormal dvs $W^T W = I$, så kan entropin kan skrivas $h(\mathbf{y}) = c + \frac{1}{2}\sum_{n=1}^{N}\log[\lambda_n]$ där c är en konstant som inte beror av W.
c) Antag att W endast har $K < N$ kolumner och att dessa kolumner är egenvektorerna till C_{xx} med största egenvärdena. För detta fall, visa att komponenterna i $\mathbf{y}$ är okorrelerade, och beräkna $h(\mathbf{y})$.

Problem #7 (1+1+1=3p)
a) Förklara hur en flerlagerpeceptron med 4 ingångar, 3 gömda noder och 4 utnoder ser ut. Rita detaljerad figur som också beskriver vad som händer inuti noderna.
b) Antag att träningen av en flerlagerperceptron ger (nästan) noll fel på träningsdata. Vad betyder detta geometriskt i termer av hur beslutsgränsen för neuronnätet ligger i förhållande till träningsdata? Rita figur med ett exempel.
Normalt vill man inte att träningsfelet ska bli noll vid träning av en flerlagerperceptron. Varför?
c) Vid närmaste granne-klassificering kan måste man bestämma värdet på en parameter K som definierar hur många grannar som ska betraktas vi klassificeringen. Om man väljer ett alltför litet värde eller ett alltför stort värde på K så blir generaliseringsförmågan dålig. Ge en rimlig förklaring till att dessa båda extremfall troligen inte är optimala val med avseende på generaliseringsförmåga.