Uppsala universitet
Signaler and System
MG 018 - 471 3229

Tentamen i mönsterigenkänning

Teknisk fysik, 3B, 3E, 3Sy Postscriptum, Sal 1,
fredag 10 december, 1999, 9.00-15.00

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

Obs! Svar utan motivering ger noll poäng!

Problem #1 (3p)
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 s_x^2 för estimatet $\hat{m}_x$ kan skrivas

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

c) Antag att man önskar att estimatets standardavvikelse skall vara mindre än eller lika med 0.01. Hur många exempel N behöver man för att detta ska gälla då $\sigma_x = 0.1$?

Problem #2 (3p)
Vid ett kemiskt experiment registrerades okorrelerade intensiteter x_i för ett stort antal frekvenser $f_i, i = 1, 2, \ldots, N$. Varje experiment gav alltså upphov till en mönstervektor (ett spektrum) $\mathbf{x} = [x_1, x_2, \ldots, x_N]$. Varje komponent x_i hade ett unikt väntevärde m_i och en unik varians σ_i^2.

Civ får som uppgift att bygga en klassificerare för dessa mönster i en tillämpning där $N = 20000$.

a) För att enklare kunna klassificera så stora mönster försöker Civ med principalkomponentanalys (PCA). Hjälp Civ att bestämma egenvektorer och egenvärden.
b) Civ noterar att $\sigma_i < 0.01, i = 1001, 1002, ..., 18000$ och att $\sigma_i > 2.5$ för alla andra våglängder. Till hur många dimensioner kan Civ reducera der ursprungliga mönstret med hjälp av principalkomponentanalysen?
c) Civ tillämpar också K-means algoritmen där K=53 och får då ett mycket litet rekonstruktionsfel. Hur kan Civ utnyttja detta resultat för att göra en betydligt bättre kompression av mönstren än med PCA?

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 esimeras 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 klassificeringen? 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 undvika att få med med outliers i träningsdata.

Problem #4 (2+2=4p)
a) Förklara skillnaden mellan parametriska och icke-parameteriska metoder. Är en enlagerperceptron en icke-parametrisk eller en parameterisk metod? Vad gäller för en flerlagerperceptron?
b) Varje metod har sina fördelar och nackdelar. Ge exempel på fördelar/nackdelar med K-närmaste grannemetoden och flerlagerpreceptronen.

Problem #5 (3p)
Visa att om man antar att mönster från två klasser är dragna från två multivariata normalfördelningar så blir den optimala bestutsgränsen (i det allmänna fallet) kvadratisk. Vad krävs för att beslutsgränsen ska vara linjär?

Problem #6 (1+1=2p)
Förbehandling av data är en viktig del vid design av ett system för mönsterigenkänning. En typ av förbehandling är att hantera symboliska variabler.
a) Inom genomanalys studerar man DNA-strängar som är uppbyggda av bokstäverna A,C,G och T. Ett exempel på ett mönster är ATTCGGTGGCCA. Antag att du får i uppgift att bygga en klassificerare som ska sortera DNA-strängar i

olika klasser. Presentera minst ett lämpligt förslag på hur DNA-strängarna kan kodas numeriskt så att de passar de numeriska mönsterigenkänningsmetoder som studerats i kursen. Obs! Det finns ingen uppenbar ordning mellan bokstäverna. *Kodningen måste därför vara sådan att det blir samma euklidiska avstånd mellan alla bokstäverna*. Detta är inte fallet om man tex använder kodningen: A=1, C=2, G=3 och T=4 eller den binära kodningen A=0001, C=0010, G=0100, T=1000. Ledning: Hörn på en kub.
b) Vid mönsterienkänning av uppgifter från patientjournaler är en variabel av intresse representerad som "låg", "medel", och "hög". Presentera en lämplig numerisk kodning av denna variabel. Obs! Kodningen måste ta hänsyn till att t.ex. avståndet är längre mellan "låg" och "hög" än mellan "låg" och "medel".

Problem #7 (3p)
a) Förklara hur K-närmaste grannemetoden fungerar vid klassning och estimering.
b) Hur bestämmer man med hjälp av korsvalidering bestämma bästa värdet på K?
c) Antag att det i K-omgivningen av $\mathbf{x}$ finns K_c exempel från klass c, $\sum_{c=1}^{C} K_c = N_c$. Då gäller $P(\text{klass c}|\mathbf{x}) \approx K_c/N_c$. Om man vill estimera denna sannolikhet, vad är fördelen/nackdelen med att välja K stor/liten?

Problem #8 (2p)
Du får som uppgift att bygga en prediktor som skall prediktera ljud från en bilmotor under drift. Träning sker med ett relativt stort antal exempel som samlas in (kontinuerligt) under körningen.
a) Förklara varför K-närmaste grannemetoden inte är lämplig för denna tillämpning.
b) Då motorn är kall ger den ifrån sig helt andra ljud än när den är varm. Detta kräver att prediktorn är tidsvariabel dvs ändras i takt med att motorns temperatur stiger. Förklara varför en flerlagerperceptron inte verkar lämplig för denna uppgift.
c) Förklara med ledning av dina svar i a) och b) varför ett radialt basfunktionsnät eller linjär regression kan vara intressanta alternativ. Redogör för fördelar och nackdeler dem emellan. Med vilken metod ska det radiala basfunktionsnätet tränas för att vara ett realistiskt alternativ? Ledning: klustring.

Problem #9 (2p)
a) Beskriv minst tre olika typer av olineariteter som används i neuronnät (flerlagerperceptroner och radiala basfunktionsnät).
b) Vilken roll har bias (tröskeln) för en neurons förmåga att klassificera? Ledning: På vilket sätt begränsas beslutsytan om man inte väljer att ha bias?

LYCKA TILL!