Uppsala universitet
Signaler and System
MG 018 - 471 3229

Tentamen i mönsterigenkänning

för civilingenjörsprogrammen i Teknisk fysik (E/Sy/B) och
Informationsteknologi,
Postscriptum, 9 oktober, 2003, 9.00-14.00

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

Problem #1 (3p)
Civ håller på att bygga ett datoriserad mönsterigenkänningssystem som ska sitta i en bilmotor och styra tändningen. Tanken är att systemet ska känna igen olika situationer och då välja bland ett litet antal olika tändningsinställningar. I motorn sitter 10 sensorer utplacerade från vilka systemet varje mikrosekund ska samla in mätvärden och sedan välja en av tändningsinställningarna. Civ har byggt ett prototypsystem baserat på en kombination av PCA och kNN.

a) Varför verkar det klokt av Civ att använda PCA?

b) Varför verkar det *inte* klokt att använda kNN i denna tillämpning?

c) Civ är inte helt nöjd med den prestanda hon får när hon testar på oberoende testdata. Förklara varför det kan bli bättre om hon minskar dimensionen på mönstervektorn från 10 till t.ex. 5 med hjälp av variabelselektion. Förklara också varför en sådan variabelselektion ibland också faktiskt resulterar i en *sämre* prestanda i jämförelse med att använda alla särdrag.

Problem #2 (3p)
Antag att exempel från en klass (class ω_1) är dragna från en normalfördelining med medelvärde noll och varians ett och att data från en annan klass (klass ω_s) är dragna från en normalfördelning med medelvärde ett och varians ett.

a) Bestäm en optimal beslutsregel som minimerar antalet klassificeringsfel för data dragna från dessa fördelningar.

b) Antag att det är dyrare (farligare) att klassa exempel från klass ω_1 som tillhörande klass ω_2 istället för tvärt om. Ge en intuitiv förkaring till hur man ska flytta beslutsgränsen (punkten) för att minska fel av denna typ.

c) Förklara hur man med hjälp av en ROC-kurva enkelt kan välja en kompromiss mellan sannolikheten att göra fel av den ena typen och sannolikheten att göra fel av den andra typen. Ledning: Rita upp och förklara hur en ROC-kurva fungerar.

Problem #3 (3p)
Vid linjär multivariat regression finns ett antal viktiga saker att tänka på.

a) Förklara hur man förbehandlar alla variabler så att de får medelvärde noll och varians ett.

b) Förklara hur man kommer tillbaka till en regressionsmodell uttryckt i originalvariablerna efter att en regressionsmodell $\tilde{y} = \mathbf{a}^t\tilde{\mathbf{x}}$ har designats i termer av variablerna $\tilde{y}$ och $\tilde{x}_i$ som alla har medelvärde noll och varians ett.

c) Vid ridge regression måste man välja värdet på en regulariserings-parameter λ. Varför använder man ridge regression över huvud taget och behövs i så fall ridge regression för ett problem där man ser att det värde på λ som ger bäst prestanda är $\lambda = 0$?

Problem #4 (3p)
a) Förklara i några få steg hur man gör vid PCR (principalkomponentregression).

b) Motivera behovet till metoder som PCR och RR (ridge regression) vid multivariat regression.

c) Härled OLS-lösningen vid multivariat regression.

Problem #5 (3p)
Optimal (minsta kvadrat) regression och optimal (Bayesiansk) klassning är viktiga grundbegrepp i samband med datoriserad mönsterigenkänning.

a) Dessa optimala lösningar bygger på antagandet att underliggande statistiska fördelningar är kända. Om fördelningarna inte är kända kan man alltid approximera fördelningarna med hjälp av t.ex. histogram eller antagandet att fördelningarna är Gaussiska. Varför finns det då någon anledning att titta på andra metoder t.ex. linjär regression och linjära klassificerare? Ledning: Förklara varför de "optimala" metoderna inte alltid vinner över t.ex. linjära metoder.

b) Antag att vi känner till att täthetsfunktionen $p(y|x)$ är normalfördelad med medelvärde m_x och varians σ_x^2. Skriv ut definitionen av $E\{y|x\}$ i termer av en integral och bestäm värdet på $E\{y|x\}$.

c) Vid generell olinjär regression är problemet att hitta en funktion $g(x)$ av en variabel x sådan att funktionalen

$$J(g) = \frac{1}{N} \sum_{n=1}^{N} (y_n - g(x_n))^2$$

minimeras där N är antalet designexempel. Förklara varför detta problem inte är välformulerat ("ill-posed" på engelska) och hur man normalt sett formulerar om problemed med hjälp av en straffterm. Ledning: Du behöver inte beskriva detaljerna i strafftermen.

Problem #6 (1+2=3p)
Civ har byggt en klassificerare och påstår att den har en error rate på (nära) 10%. Detta har hon kommit fram till genom att testa klassificeraren med 20 dyrbara testexempel och då endast observerat två felklassingar.

a) Med hjälp av Bayes sats kan man konstatera att HPD-regionen ("highest probability density region") för detta fall är $[2\%, 28\%]$. Hur skulle du förklara för Civ vad detta betyder?

b) Istället för att utföra en enda test på detta sätt skulle Civ kunna utföra korsvalidering. Förklara hur LOO-metoden fungerar och beskriv minst en fördel och en nackdel med denna metod i jämförelse med holdout-metoden.

Problem #7 (3p)
Design av linjära klassificerare kan ske med hjälp av flera olika metoder och kriterier.

a) Förklara grundtankarna bakom Fishers linjära diskriminant och varför den inte bara är lämplig för visualisering utan också för klassificering av mönster.

b) Förklara hur det går till att designa en klassificerare med hjälp av minstrakvadratkriteret.

c) Då målet är att minimera antalet fel på designdata vid design kan kriteriefunktionen skrivas

$$J(\mathbf{v}) = \sum_{\mathbf{y_n} \in Y} (-\mathbf{v}^T \mathbf{y}_n)$$

där mängden Y definieras som $Y = \{\mathbf{y}_n | \mathbf{v}^T \mathbf{y}_n < 0\}$ dvs Y innehåller alla exempel som blir felklassade då viktsvektorn är lika med $\mathbf{v}$. Förklara hur den optimala viktsvektorn $\mathbf{v}$ som man får fram vid minimeringen av J sedan ska användas för klassning av ett okänt exempel $\mathbf{x}_j$. Ledning. Förklara hur mönstret $\mathbf{x}_j$ har transformerats till vektorn $\mathbf{y}_j$ i kriteriefunktionen J ovan och varför man vet att $\mathbf{x}_j$ blivit felklassad om olikheten $\mathbf{v}^T \mathbf{y}_j < 0$ är uppfylld.

Lycka till!