Uppsala universitet
Signaler and System
MG 018 - 471 3229

Tentamen i datoriserad mönsterigenkänning

för civilingenjörsprogrammen i Teknisk fysik (E/Sy/B) och
Informationsteknologi, Polacksbacken, 14 oktober, 2005, 15.00-20.00

Hjälpmedel: Mathematical Handbook, Mathematics Handbook (BETA)

Problem #1 (1+1+1+2+2+1=8p)
Civ vill bygga ett mikroskop för automatisk klassificering av levande celler baserat på form och storlek. Civ har tillgång till ett bildbehandlingsprogram som kan identiera cellernas konturer.

a) Koordinaterna för de identifierade konturerna är inte rotationsinvarianta. Civ använder därför sk Fourierdeskriptorer för att konvertera koordinaterna till en särdragsvektor (feature vector) bestående av rotationsinvarianta särdrag. Varför måste Civ göra denna konvertering? Varför skulle prestandan bli mycket dålig om Civ byggde en klassificerare direkt med koordinaterna som särdragsvektor?

b) Civ gör en enkel form av särdragsselektion (feature selection) där hon hela tiden ökar på särdragsvektorn med ett nytt rotationsinvariant särdrag och estimerar resulterande prestanda med hjälp av 10 oberoende testexempel. Det visar sig att de bästa resultaten erhåller Civ för 10 särdrag då den resulterande klassificeraren gör 2 fel vid test med de 10 testexemplen. Då Civ istället använder endast 2 särdrag blir resultatet 3 fel vid test. Motivera varför Civ bör välja den klassificerare som baserar sig på 2 särdrag trots att den ger ett större antal testfel.

c) Civerth tycker att Civ gör fel som inte utnyttjar de 10 testexemplen bättre. Civerth föreslår att Civ ska uppskatta prestanda genom att göra design med alla exempel och sedan också testa med alla designexemplen (apparent error). Förklara varför Civerths förslag är både bra och dåligt, ett halvt svar ger tyvärr noll poäng.

d) Istället för att estimera prestanda med hjälp av tio oberoende testexempel som i b) ovan, ge ett förslag på ett alternativ sätt att uppskatta klassificerarens prestanda. Ge inte bara ett namn på en metod utan förklara också kort hur ditt

föreslagna alternativ fungerar. (2p)

e) Civ testar sin klassificerare på 20 helt nya exempel och noterar att klassificeraren gör fel på 4 av dem. Civerth testar Civs klassificerare på 2000 helt nya exempel och noterar att klassificeraren gör fel på 200 av dem. Både Civ och Civerth rapporterar att sannolikheten att göra fel är 20%. Förklara varför Civs estimat är mindre pålitligt och hur man med hjälp av Bayesiansk inferens (Bayes sats) kan räkna fram osäkerhetsmått för både Civ och Civerths prestandaestimat. (2p)

f) Civ skaffar sig till sist en mycket stor påse med testexempel och använder dessa för att rita upp ROC-kurvor för tre olika klassificerare som heter A, B och C. Civ noterar att A alltid är sämst medan B är bättre än C för falsklarm på intervallet $[10\%, 30\%]$. Rita ett exempel på tänkbara ROC-kurvor för de tre klassificerarna som motsvarar denna situation.

Problem #2 (1+1+1+2=5p)

a) Förklara, vilken är den huvudsakliga skillnaden mellan den beslutsgräns man får efter träning med perceptronalgoritmen och efter träning av en supportvektormaskin (support vector machine)? Varför är den beslutsgräns som supportvektormaskinen producerar oftast mer attraktiv? Ledning: Marginal

b) Trots att linjära klassificerare endast kan producera beslutsgränser i form av hyperplan presterar de ofta lika bra som tex kvadratiska klassificerare och kNN när antalet träningsexempel är litet. Varför är det så?

c) Förklara kort grundtanken bakom Fishers linjära klassificerare. Ledning: Alla exempel projiceras på en linje, hur vill man att exemplen ska fördela sig på linjen?

d) Det visar sig att om man vid multivariat linjär regression med minsta kvadratmetoden OLS använder y=+1 för exempel från klass 1 och y=-1 för exempel från klass 2 så blir OLS lösningen $\mathbf{w}_{OLS} = (X^T X)^{-1} X^T \mathbf{y}$ parallell med viktsvektorn i Fishers linjära diskriminantfunktion. Man kan alltså bestämma Fishers klassificerare (Fishers linjära diskriminant) med hjälp av OLS istället för med hjälp av spridningsmatriserna som används i den ursprungliga formuleringen! Med ledning av det du vet om problem med OLS-metoden då variablerna är korrelerade och metoder för att lösa det, ge ett förslag på hur man skulle kunna konstruera en modifierad version av Fishers klassificerare som fungerar bra även då variablerna i mönstervektor är starkt korrelerade. (2p)

Problem #3 (1+1+1=3p)
Antag att exempel från en klass (class ω_1) är dragna från en triangelformad fördelning på intervallet $[0, 3]$ där fördelningens (triangelns) centrum ligger i punkten 1.5 och fördelningen är symmetrisk. Antag vidare att exempel från en annan klass (class ω_2) är dragna likformigt på intervallet $[2, 5]$ (intervallen överlappar alltså). Antag att sannolikheten att observera ett exempel från ω_1 och ω_1 är lika dvs 50%.

a) Bestäm en optimal beslutsregel som minimerar (förväntade) 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 förklaring till hur man ska flytta beslutsgränsen (punkten) för att minska fel av denna typ.

c) Modellbaserad Bayesiansk klassificering är elegant men lider av två stora problem som gör att andra metoder som tex kNN och supportvektorklassificering är intressanta alternativ. Förklara kort minst ett av dessa två problem.

Problem #4 (1+1+2+1=5p)
a) Civ har komprimerat ett N-dimensionellt mönster $\mathbf{x}$ till ett M-dimensionellt mönster $\mathbf{t}$ med hjälp av egenvektorer $\mathbf{e}_i$, $i = 1, 2, ...M$ och medelvärdesvektorn $\mathbf{m}$. Hur ska Civ göra för att rekontruera $\mathbf{x}$ så bra som möjligt (i minsta kvadratmening)? Ledning: Skriv ned och förklara formeln för rekonstruktionen.

b) Vid analys av en 10×10-dimensionell kovariansmatris C noterar Civ att alla egenvärden är noll utom två. Förklara vad detta betyder geometriskt och vilka möjligheter det öppnar för visualisering av exempel dragna från den underliggande 10-dimensionella fördelningen.

c) Vid hierarkisk klustring måste man först definiera ett mått på likhet (avstånd) mellan mönster och sedan också ett mått på likhet (avstånd) mellan grupper av mönster. Ge två exempel på vanliga mått som används för att mäta likhet mellan grupper av mönster i detta sammanhang. (2p)

d) Förklara hur man kan kombinera icke-hierarkisk klustring med PCA för att komprimera data på ett effektivt sätt.
Lycka till!