Uppsala universitet, Signaler and System
MG 018 - 471 3229, 611 4241

Tentamen i datoriserad mönsterigenkänning

för civilingenjörsprogrammen i Teknisk fysik (E/Sy/B) och Informationsteknologi,
Magistern sal A+B, fredag 15 december 2006, 8.00-13.00

Hjälpmedel: Mathematical Handbook, Mathematics Handbook (BETA), 5-sidig utskriven formelsamling "Collection of Formulas". Istället för den utskrivna formelsamlingen är det tillåtet att ha med sig ett eget A4 papper med valfria egna formler. Obs! Svar utan motiveringar ger noll poäng!

Problem #1 (5p)

a) Algorimen KMEANS för icke-hierarkist klustring bygger på att försöka minimera ett felkriterium. Vilket kriterium är det som KMEANS försöker optimera? Förklara också varför detta kriterium är vettigt?

b) Förklara hur första stegen ser ut vid hierarkisk klustring och förklara därmed behovet av två likhetsmått, ett för avstånd mellan individer och ett för avstånd mellan grupper.

c) Ge exempel på två av de vanligast måtten som används för att mäta avstånd mellan grupper vid hierarkisk klustring.

d) Kalle utför hierarkisk klustring av 1000 exempel i en databas där alla exempel har 10 särdrag och där medelvärdet av alla 1000 exempel ligger i origo. Kalle vet att det finns 5 mycket väl separerade kluster med nästan ingen spridning alls inom respektive kluster. Rita en skiss på ett dendrogram som är ett exempel på hur resultatet av klustringen skulle kunna se ut. Tips: Du kan självklart inte rita ut alla 1000 löv i trädet, gör bara en enkel skiss.

e) Kalle utför nu PCA på exemplen som beskriva i deluppgift d). Hur många av egenvärdena kommer att vara i stort sett lika med noll? Glöm inte att motivera ditt svar.

Problem #2 (5p)

a) Med hjälp av en stor påse med testexempel analyserar Civ hur en klassificerare beter sig då antalet träningsexempel varieras. Rita en typisk "learning curve" som beskriver hur prestandan för klassificeraren bör se ut som funktion av antalet träningsexempel.

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 principen för hur Fishers linjära diskriminant fungerar dvs vilken typ av särdrag man försöker skapa med hjälp av denna metod.

d) Förklara hur klassificering går till med hjälp av ett binärt beslutsträd då ett nytt mönster $\mathbf{x} = (x_1, x_2)$ observeras. Ledning: Välj själv ett exempel på ett beslutsträd inklusive beslutströsklar och rita upp hur trädets beslutsregioner ser ut. Förklara sedan hur beslutsprocessen går till.

e) Förklara varför beslutsgränserna för ett binärt beslutsträd alltid är parallella med koordinataxlarna.

Problem #3 (1+2+1=4p)
a) Antag att exempel från en klass ω_1 är dragna från en likformig fördelning på regionen $[0, 1] \times [0, 1]$ i planet och att exempel från en annan klass ω_2 är dragna likformigt i regionen $[0, 2] \times [0, 2]$. Denna andra region innesluter alltså den första. Antag vidare att sannolikheten att observera ett exempel från ω_1 är 50% och sannolikheten att observera ett exempel från ω_2 är 50%. Bestäm en optimal beslutsregel som minimerar (förväntade) antalet klassificeringsfel för data dragna från dessa fördelningar.

b) Hur ligger besutsgränsen då sannolikheten för att observera ett exempel från ω_1 minskar till 19% (nitton procent)?

c) Antag att kostnaden för att klassificera exempel från ω_1 fel är extremt mycket större än att göra fel på exempel från ω_2. Under samma antagande som i deluppgift b) bestäm optimal besutsgräns i detta fall.

Problem #4 (2p)
a) Skriv matlabkod som genererar 10000 stycken normalfördelade fem-dimensionella vektorer i ett underrum som spänns upp av kolumnerna i en på förhand given 5×3 matris $\mathbf{B}$.

b) Förklara varför man med hjälp av PCA kan använda de 10000 exemplen från deluppgift a) för att rekonstruera det rum som kolumnerna i matrisen $\mathbf{B}$ spänner upp.

Problem #5 (3p)
a) Förklara hur K-faldig korsvalidering går till och varför denna metod utnyttjar tillgängliga exempel på ett effektivt sätt i jämförelse med en vanlig "holdout" test.

b) Kalle utför 10-faldig CV och får resultatet att sannolikheten att göra en felklassning är 15%. Kalle gör sedan en slutdesign med hjälp av alla tillgängliga exempel och konstaterar efter test med 10000 exempel att prestandan är 35%. Ge minst en förklaring till vad som kan vara orsaken till den stora skillnaden.

c) Förklara hur man beräknar ett Bayesianskt konfidensintervall efter att ha gjort k_t fel vid test med N_t testexempel.

Problem #6 (2p)
Kalle studerar ett tvåklassproblem där varje exempel består av 10 särdrag. Det visar sig simultana fördelningen för särdrag 6 och 7 dragna från klass ω_1 består av en summa (mixtur) av två normalfördelningar, båda med enhetsmatisen som kovariansmatris. Den ena av dess har sitt centrum i punkten $(0, 10)$, den andra har sitt centrum i punkten $(10, 0)$. Ett exempel från en sådan mixtur-fördelning genereras genom att först dra ett slumptal som bestämmer vilken av de två normalfördelningarna som ska väljas och sedan dras exemplet från den utvalda normalfördeningen.

a) Rita en grov skiss som visar nivåytorna för den simulatan fördelningen för särdrag 6 och 7 för exempel dragna från klass ω_1. Det visar sig simultana fördelningen för särdrag 6 och 7 dragna från klass ω_2 också består av en summa av två normalfördelningar, även dessa med enhetsmatisen som kovariansmatris. Den ena normalfördelningen har sitt centrum i punkten $(0, 0)$, den andra har sitt centrum i punkten $(10, 10)$. Rita in nivåytor även för denna fördelning i figuren.

b) Förklara varför en särdragsselektionsalgoritm som studerar ett särdrag i taget aldrig skulle välja särdrag 6 och 7.

Lycka till!