

Tentamen i datoriserad mönsterigenkänning

för civilingenjörsprogrammen i Teknisk fysik (E/Sy/B) och
Informationsteknologi samt Naturvetarprogrammet
Polacksbackens skrivsal, 8 oktober, 2004, 13.00-18.00

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

Problem #1 (2+1p)

Antag att en-dimensionella exempel från en klass ω_1 är dragna från en likformig fördelning på intervallet $[0,3]$ och att exempel från en annan klass ω_2 är dragna likformigt på intervallet $[2,5]$ (intervallen överlappar). Antag vidare att sannolikheten att observera ett exempel från ω_1 är 80% och sannolikheten att observera ett exempel från ω_2 är 20%.

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

b) Förklara på principnivå (gärna med en figur) hur ROC kurvorna för en kNN klassificerare och en linjär klassificerare ser ut för ett problem där kNN klassificeraren endast är bättre än den linjära i ett fall där det är acceptabelt med ett falsklarm som är mindre än 10%.

Problem #2 (3p)

a) Vid igenkänning av ord från en inspelad mikrofonsignal är det viktigt att göra lämplig särdragsextraktion. En metod för detta är att spara ett antal prototypsignaler från varje klass och sedan använda en algoritm som heter dynamic time warping för att i princip räkna ut likheten mellan dessa prototypsignaler och de signaler som används för träning och testning. Detta spektrum av likheter kan sedan fungera som särdragsvektor vid design av en klassificerare. Ge minst en rimlig förklaring till varför denna typ av särdragsvektor är bättre att använda för klassificering än vektorer bestående av de inspelade mikrofonsignalerna helt

obehandlade.

b) Förklara varför det är viktigt att identifiera så kallade outliers vid design av ett mönsterigenkänningssystem. Ge ett konkret exempel på vad en outlier kan orsaka vid design av en klassificerare eller regressionsmodell.

c) Beräkna kovariansen mellan två stokastiska variabler x och y då den simultana täthetsfunktionen $p(x, y)$ är ett $(p(x, y) = 1)$ på en enhetskvadrat (kvadrat med sidlängd ett) centrerad i origo och noll $(p(x, y) = 0)$ utanför. Ledning: Utgå från definitionen av kovarians.

Problem #3 (3p)

Civ har byggt en klassificerare för två klasser och testar den sedan med två oberoende testexempel ($N=2$) och gör ett ($k=1$) fel.

a) Vad är punktskattningen för klassificerarens prestanda?

b) Civ överväger att använda k -fold CV för prestandaskattning. Förklara minst en väsentlig fördel respektive nackdel med CV i förhållande till att använda holdout-metoden.

c) Härled mha Bayes sats ett uttryck för osäkerheten kring en klassificerares verkliga prestanda efter en test med två testexempel som resulterar i ett fel. Ledning: Osäkerheten ska beskrivas i form av en täthetsfunktion $p(e|k = 1, N = 2)$ med arean ett under sig.

Problem #4 (3p)

Perceptronkriteriet för en linjär klassificerare kan 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 $\mathbf{v}$.

a) Motivera och förklara denna formel, vad är tex $\mathbf{v}$ och hur ska den användas vid klassificering av ett okänt exempel $\mathbf{x}_n$? Vilken är relationen mellan $\mathbf{x}_n$ och motsvarande vektor $\mathbf{y}_n$ i formeln?

b) Förklara skillnaden mellan felkorrigering inlärning (som vid Perceptroninlärning) och indirekt inlärning (som vid design av Bayesiansk klassificerare).

c) Ett alternativ till linjära diskriminantfunktioner är kvadratiske diskriminantfunktioner, $g_i(\mathbf{x}) = \mathbf{x}^T Q_i \mathbf{x} + \mathbf{q}_i^T \mathbf{x} + q_i$.

Som vanligt väljs klass j om $g_j > g_i$, $i \neq j$. För ett två-klassproblem, bestäm ett uttryck för klassificerarens beslutsgräns och förklara hur dessa beslutsgränser skiljer sig från de gränser man får då diskriminantfunktionerna är linjära.

Problem #5 (3p)

a) Civ försöker bygga en linjär regressionsmodell med hjälp av OLS. Hon delar först de exempel hon har i två halvor och bygger två olika regressionsmodeller med hjälp av de två uppdelade datamängderna. Civ upptäcker då att hon får helt olika regressions-koefficienter i de två modellerna och blir orolig. Civerth kallas in och föreslår att Civ ska använda PLS och nu fungerar det mycket bättre! Ge en rimlig förklaring till varför det fungerar bra med PLS men inte med OLS.

b) Ett alternativ till RR är den sk LASSO-metoden. Den bygger på att hitta en regressionsvektor $\mathbf{w}$ som minimerar kriteriefunktionen $J_{LASSO}(\mathbf{w}) = \sum_{n=1}^N (y_n - \mathbf{w}^T \mathbf{x}_n)^2 + \alpha \sum_{i=1}^d |w_i|$. Här betyder $|w_i|$ absolutbeloppet av elementet w_i i vektorn $\mathbf{w}$, N är antalet tränings exempel, d är antalet särdrag och α är en regulariseringskonstant som motsvarar den vid RR. Skriv ned kriteriefunktionen för RR på samma form som denna kriteriefunktion J_{LASSO} och identifiera på vilket/vilka sätt den skiljer sig från J_{LASSO} . Bestäm också uttrycket för den partiella derivatan $\frac{\partial J_{LASSO}}{\partial w_k}$. Ledning: $f(a) = |a|$ betyder att $f(a) = a$ om $a > 0$ och $f(a) = -a$ då $a < 0$.

c) Civ bygger en RR modell med hjälp av mycket få exempel och bestämmer utan problem ett bästa värde på regulariseringskonstanten med hjälp av CV. Tyvärr visar sig regressionsmodellen trots detta inte alls fungerar bra på de 1000 testexempel som Civs chef sedan använder. Försök att ge en rimlig förklaring till vad som kan ha inträffat. Ledning: överanpassning.

Problem #6 (3p)

Diskutera följande klassificerare med avseende på deras relativa behov av minne, tränings tid och responstid. Glöm inte att motivera dina svar genom att kort för var och en av klassificerarna svara på följande frågor: Vad måste lagras efter träningen? Vad är det som gör träningen snabb/långsam? Vad måste beräknas

vid respons? Klassificerarna är: 1) kNN. 2) Linjär Gaussisk klassificerare. 3) Linjär klassificerare tränad med perceptron-algoritmen.

Problem #7 (3p)

Civerth vill skapa ett automatiskt system för infärgning av 1000 digitalt lagrade röntgenbilder föreställande en viss kroppsdel (tex lårbenet) hos 1000 patienter. Blodkärl, senor, muskler etc ska få olika färger för att underlätta medicinsk tolkning och för att kunna följa förändringar mellan provtagningar. Förklara med hjälp av din erfarenhet från liknande problem vid färläggning av kartbilder hur Civerth ska gå tillväga för att träna en klassificerare att lösa denna uppgift.

Lycka till!