Pattern Recognition
Exam on 2008-02-04

NO OPEN BOOK! GEEN OPEN BOEK! - It is not allowed to use the course book(s) or any other (printed, written or electronic) material during the exam.
Give sufficient explanations to demonstrate how you come to a given solution or answer!
The 'weight' of each problem is specified below by a number of points, e.g. (20 p).

1. (20 p) Minimum error classification. Missing features.
Consider a two-dimensional, three-category pattern classification problem with priors $P(\omega_1) = 0.5$, $P(\omega_2) = 0.25$, $P(\omega_3) = 0.25$. We define the 'square distribution' $S(\boldsymbol{\mu},a)$ to be uniform inside a square of size $a \times a$ (i.e. side length a) centered on $\boldsymbol{\mu}$, and elsewhere 0. The sides of the square are parallel to the coordinate axes. The class-conditional probabilities for the three categories ω_1, ω_2, and ω_3 are such square distributions $S(\boldsymbol{\mu}_i,a_i)$, i = 1, 2, 3, with the following parameters:

ω_1: $\boldsymbol{\mu}_1 = (0, 0)$, $a_1 = 3$; ω_2: $\boldsymbol{\mu}_2 = (-1, 1)$, $a_2 = 1$; ω_3: $\boldsymbol{\mu}_3 = (1, -1)$, $a_3 = 2$.

a) **(6 p)** Classify the points (1,1), (0.5,-0.5) and (-0.7,1) with minimum probability of error.
b) **(14 p)** Classify with minimum probability of error the patterns (*,1) and (1,*), where * denotes a missing feature.

2. (20 p) Binary decision trees. Consider the following multi-set S of fruits, represented as four-feature patterns in a ten-category problem. Each pattern is defined by four features (colour, size, shape, texture) which can take the following values:
colour: y(ellow), g(reen), r(ed), b(lue), o(range);
size: xs (extra small), s(mall), m(edium), l(arge), xl (extra large);
shape: r(ound), e(lipsoidal), n(arrow), rcv (round with concavity);
texture: s(mooth), c(itrus) .
Each pattern is labeled by a category label lemon, apple, banana, orange, melon, water melon, peach, grapes, blue berry or mango. The labeled patterns in the multi-set S are:
S = { labeled as lemon: (y,m,r,c), (g,m.r,c), (y,m,e,c);
labeled as apple: (r,m,rcv,s), (g,m,rcv,s), (y,m,rcv,s);
labeled as banana: (y,m,n,s), (g,m.n,s), (y,l,n,s);
labeled as orange: (o,m,r,c), (o,m,r,c);
labeled as melon: (y,l,r,s), (y,l,e,s), (g,l,e,s);
labeled as water melon: (g,xl,r,s), (g,l,r,s);
labeled as peach: (y,m,rcv,s), (r,m,rcv,s);
labeled as grapes: (b,s,r,s), (g,s,r,s), (y,s,r,s);
labeled as blue berry: (b,xs,r,s), (b,xs,r,s);
labeled as mango: (g,m,e,s), (y,m,e,s)
}

a) Compute the misclassification impurity of S. ($i(S) = 1\text{-}max_j P(\omega_j)$)
b) Split S in two multi-subsets L and R using the following rule and compute the impurity drop achieved by this split: Q1: "Put a pattern in L if (*size* = m), otherwise put it in R."
c) Split S in two multi-subsets L and R using the following rule and compute the impurity drop achieved by this split.: Q2: "Put a pattern in L if (*colour* = y OR o OR r), else put it in R."

d) Which of the two rules Q1 and Q2 would you use for building a decision tree? Why?
e) Take the subset L obtained with the optimal rule from part d above and split it into two subsets LL and LR using a rule of your choice that concerns the shape or the texture feature. Compute the impurity drop achieved by that split.

3. (20 p) Learning vector quantization (LVQ). K-means algorithm. Describe the basic LVQ algorithm (LVQ1) and the k-means algorithm. What are the similarities and differences between these two algorithms.

4. (10 p) Receiver operating characteristics. What do you understand by a receiver operating characteristics (ROC)? To which class of problems does it apply? What is the common property of points that lie on the same ROC curve?

5. (10 p) UPC and natural patterns. What are the main differences between the Universal Product Code (UPC) and feature vectors extracted from natural objects?

6. (10 p) Iris pattern recognition. Assume that you are given a set of 100 000 binary feature vectors, each of which is a binary code of the iris pattern of a person. The set contains 100 iris codes of each of 1000 persons. Describe how you would use this data to design an authentication system based on statistical decision theory.

7. (10 p) Eigenfaces. What is an eigenface and how is this concept used in face recognition?