Gjalt

Pattern Recognition
Exam on 2009-01-26

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. (10 p) Naïve Bayes rule. Give and explain naïve Bayes rule. How would you use this rule to design a spam filter? Give an example.

2. (30 p) Bayesian decision boundaries for normal distributions. The following two sets of feature vectors originate from two bivariate normal distributions that represent two different classes:

Class w_1: S1 = {(3,3), (7,3), (3,7), (7,7)}
Class w_2: S2 = {(6,6), (10,6), (6,10), (10,10)}.

a) **(10 p)** Using these feature vectors and the maximum likelihood method, estimate the parameters of the two distributions.
b) **(10 p)** Find the analytical form of the optimal Bayesian decision boundary between the two classes, assuming equal prior probabilities. Give an approximate drawing of this boundary together with the positions of the means of the distributions and the elipses (or circles) that consist of points which are at Mahalanobis distance of 1 from the means. Put class labels on the regions in which the 2D feature space is divided by the decision boundary.
c) **(5 p)** Using the obtained decision boundary, classify the following feature vectors: (7,5), (8,8), (5,6), (6,8).
d) **(5 p)** Classify the same feature vectors using **nearest neighbour classification** and Euclidian distance and compare the results of the two classification methods.

3. (15 p) Binary decision trees. This problem conserns the construction of a binary decision tree for three categories from the following two-dimensional data:

S = {Patterns with label w_1: (0,2), (1,7), (3,4);
Patterns with label w_2: (5,6), (6,1), (8,8);
Patterns with label w_3: (5,9), (7,3), (8,10)}

a) **(1 p)** Compute the misclassification impurity of S.
b) **(4 p)** Split S in two subsets L and R using the following rule and compute the misclassification impurity drop achieved by this split: Q_S: "Put a pattern in L if $x_1 > 3.5$, otherwise put it in R."
c) **(8 p)** Continue to grow your tree fully using as criterion the misclassification impurity drop. If two candidate splits are equally good, prefer the one based on x_1 (rather than x_2). Show the final tree and all queries.
d) **(2 p)** Use your tree to classify the point (8,3).

4. (15 p) Cross-validation. Consider a data set that includes 1000 128-dimensional feature vectors that come from four different categories. Using this data set, you want to construct a k-nearest neighbour classifier.
a) **(5 p)** Describe how you can select the value of the parameter k using cross-validation.

b) **(5 p)** What is overfitting and how can you detect and prevent it using cross-validation?
b) **(5 p)** Which types of cross-validation can you use and what are the advantages and disadvantages of each type?

5. (10 p) Hierarchical clustering. Construct two cluster dendrograms for the one-dimensional data $S = \{1, 3, 6, 10, 16\}$
a) **(5 p)** using the distance measure $d_{max}(S_i,S_j) = \max_{x \in S_i,\, x' \in S_j} |x-x'|$,
b) **(5 p)** using the distance measure $d_{min}(S_i,S_j) = \min_{x \in S_i,\, x' \in S_j} |x-x'|$.

6. (10 p) Describe the basic LVQ algorithm (LVQ1) and the k-means algorithm. What are the similarities and differences between these two algorithms?