In The Name of God, The Compassionate, The Merciful

Name: ....

Student ID#: ....

Statistical Pattern Recognition (CE-725)
Department of Computer Engineering
Sharif University of Technology

Midterm Exam - Spring 2011
(120 minutes - 110 points)

1. **Overall Concepts: True or False (30 points)**

   For each of the following parts, specify that the given statement is true or false. In the case of true, describe a brief explanation for why that is true, otherwise, propose a counter example.

   a. .... In n-fold cross-validation, each data point belongs to exactly one test fold, so the test folds are independent. Then the error estimates of the separate folds also independent. So, given that the data in the test folds i and j are independent, the error estimates on test folds i and j, also independent.

   b. .... There is an a priori good choice of n for n-fold cross-validation.

   c. .... Cross-validation can reveal overfitting.

   d. .... With k-fold cross-validation, larger k is always better.

   e. .... Applying feature selection to reduce the feature dimension decreases the probability of overfitting.

   f. .... A classifier trained on less training data is less likely to overfit.

   g. .... Given m data points, the training error converges to the true error as $m \to \infty$.

h. ………… Any optimally designed classifier must use a feature vector with at least as many elements as the number of possible decision categories.

i. ………… In thPe intersection point of ROC curve and the line which connects (TP=0, FP=1) point to (TP=1, FP=0) point, sensitivity is equal to specificity.

j. ………… Changing the priors changes the ROC curve.

2. Confusion Matrix (10 points)

In a medical application domain, suppose we build a classifier for patient screening (True means patient has cancer). Suppose that the confusion matrix is from testing the classifier on some test data. Which of the following situations would you like your classifier to have and which of them not (write down ✗ or ✓ in the blank space)?

- ………… FN >> FP
- ………… FN × TP << FP × TN
- ………… Specificity → 1
- ………… Sensitivity → 1
- ………… Specificity < Sensitivity

3. Feature Extraction (15 points)

Given the following 2-d data for a two class classification problem:

ω_1 = [(1, 1), (1, 2), (2, 4), (2, 1), (3, 1), (3, 3)]

ω_2 = [(2, 2), (3, 2), (3, 4), (5, 1), (5, 4), (6, 5)]

Determine the direction of the optimal projection line in a single dimension using FLD method.

In The Name of God, The Compassionate, The Merciful

4. Probabilistic methods (30 points)

a) We have K classes $C_1, C_2, \ldots, C_k$ where each class C_i is uniformly distributed over $-(2^{i-2}) < x < 2^{i-2}$.

a1) Find the likelihood ratio test classifier.

a2) Find P(error) of the designed classifier.

a3) What is $\lim_{k\to\infty} P(\text{error})$?

b) In a particular binary hypothesis testing application, the conditional densities for a scalar feature x given class w_1 and w_2 are:

$$P(x \mid w_1) = k_1 \exp(-x^2/20)$$
$$P(x \mid w_2) = k_2 \exp(-(x-6)^2/12)$$

b1) Find k_1 and k_2 (Avoid explicit calculation of k_1 and k_2).

b2) Assume that the prior probabilities of the two classes are equal, and that the cost for choosing correctly is zero. If the costs for choosing incorrectly are $C_{12}=\sqrt{3}$ and $C_{21}=\sqrt{5}$ (where C_{ij} corresponds to predicting class i when it belongs to class j), Find the decision boundary, which minimizes the Bayesian risk.

In The Name of God, The Compassionate, The Merciful

5. Linear Discriminant Functions (15 points)

Suppose we have the following two-dimensional points for categories ω1:{[0,2], [2,0]} and {[3,1], [1,3]} for class ω2 . Construct a linear discriminant function for classifying these samples by pseudo-inverse matrix method.

6. Bonus Problem (10 points)

Consider the data set $\{x_1, x_2, \dots, x_N\}$ where each x_i resides in R^n. Using the orthonormal n*k matrix C (where k<n) the new data set $D = \{C^T x_1, \dots, C^T x_N\}$ is generated. Find the relation between the principal direction and principal values of the two data sets.

Good Luck!