In The Name of God, The Compassionate, The Merciful

Name: ..**Problems' keys**..

Student ID#: .....

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

Midterm Exam - Spring 2012
(100 minutes – 100 points)

1) Basic Concepts (30 points)

a) True or false questions: For each of the following parts, specify that the given statement is true or false. In the case of true, provide a brief explanation, otherwise, propose a counter example.

i- **False** Assume zero-one costs. The minimum probability of error decision boundary always passes through the point where the two classes' likelihood functions are equal.

Explanation: this statement is true, iff the priors are equal.

ii- **False** Assume d=2 dimensions and the two classes have covariance matrices equal to the Identity matrix. The decision boundary that minimizes the probability of error will always pass through the region bounded by the two means of classes. (It will always cross the line connecting their means.)

Explanation: It depend to P(w1)/p(w2). Refer to HW2.

iii- **True** Assume d=50 dimensions and the two classes have covariance matrices that are equal to the identity matrix. The optimal Fisher linear projection for separating the classes is to project onto a line connecting the means of the two class densities.

Explanation: Since the Fisher direction is $w = S_w^{-1}(\mu_1 - \mu_2)$, and S_w is $(n_1+n_2)I$, then the given statement is true.

b) Short answer questions:Please indicate for each of the following actions whether it increases or decreases overfitting.

i- **Decreases** For k-NN change k from 1 to 5.
ii- **Decreases** Apply feature selection to reduce the feature dimension from 1000 to 20.
iii- **increases** Reduce the training data set from 2000 to 1000.

c) Suppose that A^TA is the covariance matrix of a few number of data points in a high dimensional space. How we can calculate the Eigen values and Eigen vectors of this covariance matrix, having matrix A, and the Eigen values and the Eigen vectors of the AA^T (without calculating A^TA)?

Sol:

Suppose that the Eigen vectors v_i of AA^T, such that

$$AA^T v_i = \lambda_i v_i$$

Pre-multiplying both sides by A^T, we have

$$A^T AA^T v_i = \lambda_i A^T v_i$$

From which we see the Eigen values and Eigen vectors of A^TA are λ_i and $A^T v_i$, respectively.

2) Feature Reduction (35 points)

a) Consider two features x_1 and x_2, and true label y in a 2-class classification problem. The covariance matrix of these three random variables is given as the following:

	$\mathbf{x_1}$	$\mathbf{x_2}$	**y**
$\mathbf{x_1}$	a	d	e
$\mathbf{x_2}$	d	b	f
y	e	f	c

For each of the following parts, specify that the given statement is true or false?

i- **True** When d>>0 (d is very large number), someone may ignore either of the two features, without losing any performance.

ii- **True** When d<<0 (d is very small number), someone may ignore either of the two features, without losing any performance.

iii- **False** When d=0, someone may ignore either of the two features, without losing any performance.

iv- **True** When e=0, someone may ignore one of the two features, without losing any performance.

v- **True** When f=0, someone may ignore one of the two features, without losing any performance.

b) Assume we have 100 samples equally in two Classes (A & B) in 2-D space with the following means and covariances:

$$\mu_A = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \Sigma_A = \begin{bmatrix} 4 & -3 \\ -3 & 8 \end{bmatrix} \qquad \mu_B = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \quad \Sigma_B = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

Find the first principal component of all data points.

Note: $S_{A\cup B} = S_A + S_B + \frac{n_1 n_2}{n_1 + n_2}(\mu_A - \mu_B)(\mu_A - \mu_B)^T$

Sol:

$$S_A = \begin{bmatrix} 200 & -150 \\ -150 & 400 \end{bmatrix}, S_B = \begin{bmatrix} 200 & 100 \\ 100 & 200 \end{bmatrix}$$

$$S_{A\cup B} = S_A + S_B + \frac{n_1 n_2}{n_1 + n_2}(\mu_A - \mu_B)(\mu_A - \mu_B)^T$$

$$\Rightarrow S_{A\cup B} = \begin{bmatrix} 200 & -150 \\ -150 & 400 \end{bmatrix} + \begin{bmatrix} 200 & 100 \\ 100 & 200 \end{bmatrix} + \frac{50*50}{50+50}\begin{bmatrix} 2 \\ 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \end{bmatrix} = \begin{bmatrix} 500 & 0 \\ 0 & 625 \end{bmatrix}$$

$$\Rightarrow \Sigma_{A\cup B}=\begin{bmatrix}5 & 0\\ 0 & 6.25\end{bmatrix}\Rightarrow\begin{cases}\lambda_1=6.25,\ v_1=\begin{pmatrix}0 & 1\end{pmatrix}^T\\ \lambda_2=5,\ v_2=\begin{pmatrix}1 & 0\end{pmatrix}^T\end{cases}$$

c) Given the following 2-d data points for two classes:

$$w_1=\{(1,1),(1,2),(2,1),(2,4),(3,1),(3,3)\}$$

$$w_2=\{(2,2),(3,4),(4,2),(5,1),(5,4),(5,5)\}$$

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

Sol:

Let w be the direction of the projection line, then the Fisher linear discriminant method finds that the best w as $w=S_w^{-1}(\mu_1-\mu_2)$

$$\mu_1=\begin{bmatrix}2\\2\end{bmatrix},\mu_2=\begin{bmatrix}4\\3\end{bmatrix},$$

$$x-\mu_1=\begin{bmatrix}-1 & -1 & 0 & 0 & 1 & 1\\ -1 & 0 & -1 & 2 & -1 & 1\end{bmatrix},\ x-m_2=\begin{bmatrix}-2 & -1 & 0 & 1 & 1 & 1\\ -1 & 1 & -1 & -2 & 1 & 2\end{bmatrix}$$

Therefore

$$S_1=\begin{bmatrix}4 & 1\\ 1 & 8\end{bmatrix},\ S_2=\begin{bmatrix}8 & 2\\ 2 & 12\end{bmatrix}$$

And then

$$S_w=S_1+S_2=\begin{bmatrix}12 & 3\\ 3 & 20\end{bmatrix}\Rightarrow S_w^{-1}=\frac{1}{231}\begin{bmatrix}20 & -3\\ -3 & 12\end{bmatrix}$$

Finally we have

$$w=S_w^{-1}(\mu_2-\mu_1)=\frac{1}{231}\begin{bmatrix}20 & -3\\ -3 & 12\end{bmatrix}\begin{bmatrix}2\\1\end{bmatrix}=\frac{1}{231}\begin{bmatrix}37\\6\end{bmatrix}$$

3) Probabilistic Classification (15 points)

a) Consider a two-class classification problem, with the equal prior probabilities of two classes. Suppose that the likelihoods of two classes are given as the following:

$$P(x\mid C_1)=\frac{1}{\sqrt{2\pi}\left|\Sigma_1\right|}e^{-\frac{1}{2}x^t\Sigma_1^{-1}x};\Sigma_1=\begin{bmatrix}\sigma_{11}^2 & 0\\ 0 & \sigma_{12}^2\end{bmatrix},\quad P(x\mid C_2)=\frac{1}{\sqrt{2\pi}\left|\Sigma_2\right|}e^{-\frac{1}{2}x^t\Sigma_2^{-1}x};\Sigma_2=\begin{bmatrix}\sigma_{21}^2 & 0\\ 0 & \sigma_{22}^2\end{bmatrix}$$

In which of the following conditions, the boundary of two classes is an ellipse, if we use the minimum error rate classifier?

i- $\sigma_{21}^2<\sigma_{11}^2$ *and* $\sigma_{22}^2>\sigma_{12}^2$

ii- $\sigma_{21}^2>\sigma_{11}^2$ *and* $\sigma_{22}^2<\sigma_{12}^2$

iii- $\sigma_{21}^2>\sigma_{11}^2$ *and* $\sigma_{22}^2>\sigma_{12}^2$

iv- The boundary is always linear.

Sol:

The Gaussians' cross sections are two origin-centered non-skewed ellipses (according to the means and the covariance matrices). The boundary will also be elliptical, if one of those elliptical cross sections be completely within the another one. Since the variances are proportional to diagonals of the ellipses, then the option (iii) is the only correct answer.

Or, in the following equations, the resulting boundary will be an ellipse if K_1, K_2, $K > 0$. The only option that results these conditions is option (iii).

$$P(C_1 \mid x) = P(C_2 \mid x)$$

$$P(x \mid C_1)P(C_1) = P(x \mid C_2)P(C_2)$$

$$\frac{1}{2}\frac{1}{\sqrt{2\pi|\Sigma_1|}}e^{-\frac{1}{2}x^t\Sigma_1^{-1}x} = \frac{1}{2}\frac{1}{\sqrt{2\pi|\Sigma_2|}}e^{-\frac{1}{2}x^t\Sigma_2^{-1}x}$$

$$\frac{1}{\sqrt{|\Sigma_1|}}e^{-\frac{1}{2}x^t\Sigma_1^{-1}x} = \frac{1}{\sqrt{|\Sigma_2|}}e^{-\frac{1}{2}x^t\Sigma_2^{-1}x}$$

$$-\ln\sqrt{|\Sigma_1|} - \frac{1}{2}x^t\Sigma_1^{-1}x = -\ln\sqrt{|\Sigma_2|} - \frac{1}{2}x^t\Sigma_2^{-1}x$$

$$-\ln\sqrt{(\sigma_{11}\sigma_{12})^2} - \frac{1}{2}\left(\frac{x_1^2}{\sigma_{11}^2} + \frac{x_2^2}{\sigma_{12}^2}\right) = -\ln\sqrt{(\sigma_{21}\sigma_{22})^2} - \frac{1}{2}\left(\frac{x_1^2}{\sigma_{21}^2} + \frac{x_2^2}{\sigma_{22}^2}\right)$$

$$\frac{x_1^2}{\sigma_{11}^2} + \frac{x_2^2}{\sigma_{12}^2} - \frac{x_1^2}{\sigma_{21}^2} - \frac{x_2^2}{\sigma_{22}^2} = \underbrace{\ln\frac{(\sigma_{21}\sigma_{22})^2}{(\sigma_{11}\sigma_{12})^2}}_{K}$$

$$\underbrace{x_1^2\left(\frac{1}{\sigma_{11}^2} - \frac{1}{\sigma_{21}^2}\right)}_{K_1} + \underbrace{x_2^2\left(\frac{1}{\sigma_{12}^2} - \frac{1}{\sigma_{22}^2}\right)}_{K_2} = K$$

$$\frac{x_1^2}{K_2} + \frac{x_2^2}{K_1} = \frac{K}{K_1K_2}$$

b) Consider a two-class classification problem, which the prior probabilities of two classes are equal. The classifier input is a feature vector $X = (X_1, X_2)^T$ with two elements that are non-negative and statistically independent of each other.

The likelihoods of features depending the classes are given in the following:

$$P(x_k \mid C_i) = \begin{cases} \lambda_{ik}e^{-\lambda_{ik}x_k}, & 0 \le x_k \\ 0 \end{cases}, \qquad C_1\ Params: \begin{cases} \lambda_{11} = 1 \\ \lambda_{12} = 2 \end{cases}, C_2\ Params: \begin{cases} \lambda_{21} = 2 \\ \lambda_{22} = 1 \end{cases}$$

b1. Design an optimal classifier that can guess the states of nature with minimum error probability, and simplify the classifier to show that it is possible to make optimal decision using a single linear discriminant function of the type $g(x_1, x_2) = ax_1 + bx_2 + c$, with a threshold mechanism.

Sol:

As both source alternatives are equally probable, we use the maximum likelihood decision rule. As the two feature elements are independent, the feature-vector density is just the product of the density functions for the feature elements. We can use a single discriminant function

$$g(x) = \ln P(x \mid C_1) - \ln P(x \mid C_2) =$$
$$= \ln\lambda_{11} - \lambda_{11}x_1 + \ln\lambda_{12} - \lambda_{12}x_2 - \ln\lambda_{21} + \lambda_{21}x_1 - \ln\lambda_{22} + \lambda_{22}x_2$$
$$= \ln\frac{\lambda_{11}\lambda_{12}}{\lambda_{21}\lambda_{22}} + (\lambda_{21} - \lambda_{11})x_1 + (\lambda_{22} - \lambda_{12})x_2 = x_1 - x_2$$

Thus, the classifier should use the decision rule

$$r(x_1, x_2) = \begin{cases} C_1, & x_1 > x_2 \\ C_2, & \text{otherwise} \end{cases}$$

b2.What is the probability of correct decision, using the optimal classifier?

Sol:

Considering features independence, and the Bayesian boundary (x_1=x_2), we have

$$P_{correct} = P(C_1)\int_{R_1} P(x \mid C_1)dx + P(C_2)\int_{R_2} P(x \mid C_2)dx$$
$$= \frac{1}{2}\int_0^\infty \int_0^{x_1} 2e^{-x_1 - 2x_2} dx_2 dx_1 + \frac{1}{2}\int_0^\infty \int_{x_1}^\infty 2e^{-2x_1 - x_2} dx_2 dx_1$$
$$= \frac{1}{2}\int_0^\infty \left[-e^{-x_1 - 2x_2}\right]_0^{x_1} dx_1 + \frac{1}{2}\int_0^\infty \left[-2e^{-2x_1 - x_2}\right]_{x_1}^\infty dx_1$$
$$= \frac{1}{2}\int_0^\infty -e^{-3x_1} + e^{-x_1} dx_1 + \frac{1}{2}\int_0^\infty 2e^{-3x_1} dx_1$$
$$= \frac{1}{2}\left[\frac{1}{3}e^{-3x_1} - e^{-x_1}\right]_0^\infty + \frac{1}{2}\left[\frac{-2}{3}e^{-3x_1}\right]_0^\infty$$
$$= \frac{1}{2}\frac{2}{3} + \frac{1}{2}\frac{2}{3} = \frac{2}{3}$$

4) Discriminant Functions (10 points)

Let the components of the vector $x=[x_1, \ldots, x_d]^T$ be binary-valued (0 or 1), and let $P(C_j)$ be the prior probability for the state of nature C_j where j=1,...,c. We define:

$$p_{ij} = P(x_i = 1 \mid C_j) \qquad i = 1,\ldots,d;\ j = 1,\ldots,c$$

with the components of x_i being statistically independent for all x.
Show that the minimum probability of error is achieved by the following decision rule:

"Decide C_k if $g_k(x) > g_j(x)$ for all j and k", where

$$g_j(x) = \sum_{i=1}^d x_i \ln\frac{p_{ij}}{1 - p_{ij}} + \sum_{i=1}^d \ln(1 - p_{ij}) + \ln P(C_j)$$

Sol:

Consider the following discriminant function

$$g_j(x) = \ln\left[P(x \mid C_j)P(C_j)\right] = \ln P(x \mid C_j) + \ln P(C_j)$$

The components of x are statistically independent for all x in C_j, then we can write the density as a product

$$P(x \mid C_j) = \prod_{i=1}^d P(x_i \mid C_j) = \prod_{i=1}^d p_{ij}^{x_i}(1 - p_{ij})^{1 - x_i}$$

Then we have the discriminant function

$$g_j(x) = \sum_{i=1}^d \left[x_i \ln p_{ij} + (1 - x_i)\ln(1 - p_{ij})\right] + \ln P(C_j)$$
$$= \sum_{i=1}^d x_i \ln\frac{p_{ij}}{1 - p_{ij}} + \sum_{i=1}^d \ln(1 - p_{ij}) + \ln P(C_j)$$

5) Non-parametric Modeling (10 points)

Four categories are represented by their sample means

$$m_1=[-1\ -1],\ m_2=[1\ 1],\ m_3=[-1\ 1] \text{ and } m_4=[1\ -1].$$

Give $p(x|c_i)$ and $P(c_i)$ whose minimum error rate classification leads to the same decision regions resulting from the nearest neighbor rule. You need to specify all the parameters. (Hint: Think about normal distributions).

Sol:

The equal priors, and the Gaussians with the means same as the given means, and the same covariances equal to the Identity (I).

Good Luck!