Midterm Answers

Istanbul Technical University- Fall 2007
Pattern Recognition and Analysis (BBL514E)
Total worth: 25% of your grade.
Date: Wednesday, November 5, 2007.
Time: 120 mins
Closed books and notes. Please write as neat and clear as possible. Good luck!

1. [20 points] *What is (use at most three sentences per question):*
a) *Reject region:* For a classification problem with more than two classes, reject region is the region where we can not decide on which class a data point belongs to. We can also define the reject region for a binary classification problem as the places in input space where the discriminant function values are too close to each other and we would rather have an expert arrive at a decision.

b) *Bayes' rule:* P(A|B) = P(B|A)P(A)/P(B)
We use it to compute the posterior probability of a class (C) given and input (x) using the the likelihood of x (p(x|C)), prior of the class (P(C)) and evidence for the input (p(x)). i.e.

$$P(\mathrm{C_i} \mid \mathbf{x}) = \frac{P(\mathrm{C_i})\, p(\mathbf{x} \mid \mathrm{C_i})}{p(\mathbf{x})} = \frac{P(\mathrm{C_i})\, p(\mathbf{x} \mid \mathrm{C_i})}{\sum_j p(\mathbf{x} \mid \mathrm{C_j}) P(\mathrm{C_j})}$$

c) *Naïve Bayes Classifier :* A classifier which assumes that all input dimensions are independent from each other, hence the probability of any given input vector can be written as the multiplication of each input component alone: $p(x_1,..,x_d) = p(x_1)\ldots p(x_d)$.

d) *multivariate classification:* A classification problem with more than one dimensional inputs.

e) the relationship between bias, variance and mean square error of an estimator:
Given:
Unknown parameter θ
Estimator $d_i = d\,(X_i)$ on sample X_i
Bias: $b_\theta(d) = E\,[d] - \theta$
Variance: $E\,[(d - E\,[d])^2]$
mse= $r\,(d,\theta) = E\,[(d-\theta)^2]$

Mean square error (expected values are over X):
$r\,(d,\theta) = E\,[(d-\theta)^2] = E[(d-E[d]+ E[d]-\theta)^2]$
$= E[(d-E[d])^2+(E[d]-\theta)^2+2\,(d-E[d])(E[d]-\theta)]$
$= E\,[(d-E\,[d])^2] + E[(E[d]-\theta)^2]+2\,E[(d-E[d])(E[d]-\theta)]$
$= E\,[(d-E\,[d])^2] + E[(E[d]-\theta)^2]+2\,(E[d]-E[d])(E[d]-\theta)$

$= E[(d - E[d])^2] + (E[d] - \theta)^2$
$=$ Variance+Bias^2

2. [20 points]

a) *What is the difference between PCA (Principal Component Analysis) and Backward Feature Selection. Assume that you are given inputs* $X = \{\underline{x}^t\}_{t=1}^N$ *where each* $\underline{x}^t \in R^d$.

PCA projects inputs into a lower dimensional space, i.e. it is a feature projection technique. Each dimension in the lower dimensional space is a linear combination of the original inputs. The weights for the linear combination for the first dimension is the eigenvector corresponding to the largest eigenvalue, the second set of weights is the eigenvector corresponding to the next largest eigenvalue,…etc. PCA tries to project the inputs according to the directions in which there is maximum variance.

Backward feature selection is a feature selection method. It first trains d classifiers, leaving out a single feature at a time. It chooses the feature whose validation error is the minimum as the feature to leave out. This process is repeated until enough number of features or validation error is reached.

b) *Given a multivariate binary classification problem and assumption of normally distributed d dimensional inputs, what are the most complex and least complex classifiers that you could produce? Explain in detail your assumptions to arrive at those classifiers and the number of parameters needed to be estimated from training data*.

The least complex classifier that can be obtained is the least mean classifier, which is a linear classifier. It just computes the distance between the input vector and the class means and selects the class whose mean is closest. This corresponds to assuming that both class covariance matrices are the same and can be written as $\sigma^2 I$ where I is the dxd identity matrix. The number parameters that need to be estimated is 2d + 1, d parameters for each of the mean vectors and 1 parameter for σ^2.

The most complex classifier is a quadratic classifier. It corresponds to assuming that both classes have arbitrary covariance matrices. The number of parameters that need to be estimated is 2d + d(d+1). Again d parameters for each of the mean vectors and 0.5 d(d+1) parameters for each of the class covariance matrices.

3) [30 points] *Assume a two-class problem with equal a priori class probabilities and Gaussian class-conditional densities as follows:*

$$p(x \mid w_1) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}a & c\\c & b\end{bmatrix}\right) \quad \textit{and} \quad p(x \mid w_2) = N\left(\begin{bmatrix}d\\e\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$$

where $ab - c^2 = 1$.

a) *Find the equation of the decision boundary between these two classes in terms of the given parameters, after choosing a logarithmic discriminant function.*

Using the given class means and covariance matrices:

$$p(\underline{\mathbf{x}} \mid w_i) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)^T \Sigma^{-1}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)\right]$$

$$p(\underline{\mathbf{x}} \mid w_1) = \frac{1}{(2\pi)^{d/2}1^{1/2}} \exp\left[-\frac{1}{2}\underline{\mathbf{x}}^T \begin{bmatrix} b & -c \\ -c & a \end{bmatrix} \mathbf{x}\right]$$

$$p(\underline{\mathbf{x}} \mid w_2) = \frac{1}{(2\pi)^{d/2}1^{1/2}} \exp\left[-\frac{1}{2}\underline{\left(\mathbf{x} - \begin{bmatrix} d \\ e \end{bmatrix}\right)^T}\left(\mathbf{x} - \begin{bmatrix} d \\ e \end{bmatrix}\right)\right]$$

The discriminant function using log likelihood is:

$g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$

For both classes, $P(w_1) = P(w_2) = 0.5$

The discriminant function can be formulated by:

$$g_1(x) = g_2(x)$$

$$\ln p(x \mid w_1) + \ln P(w_1) = \ln p(x \mid w_2) + \ln P(w_2)$$

$$\ln p(x \mid w_1) = \ln p(x \mid w_2)$$

$$\underline{\mathbf{x}}^T \begin{bmatrix} b & -c \\ -c & a \end{bmatrix} \mathbf{x} = \underline{\left(\mathbf{x} - \begin{bmatrix} d \\ e \end{bmatrix}\right)^T}\left(\mathbf{x} - \begin{bmatrix} d \\ e \end{bmatrix}\right)$$

$$\underline{\mathbf{x}}^T \left(\begin{bmatrix} b & -c \\ -c & a \end{bmatrix} - I\right)\mathbf{x} + 2\underline{\mathbf{x}}^T \begin{bmatrix} d \\ e \end{bmatrix} - \begin{bmatrix} d \\ e \end{bmatrix}^T \begin{bmatrix} d \\ e \end{bmatrix} = 0$$

b) *Determine the constraints on the values of a, b, c, d and e, such that the resulting discriminant function results in a linear decision boundary.*

For the linear decision boundary we need to get rid of the quadratic part, therefore:
The discriminant function can be formulated by:

$$\begin{bmatrix} b & -c \\ -c & a \end{bmatrix} - I = \begin{bmatrix} b & -c \\ -c & a \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$a = 1, b = 1, c = 0$

When d=e=0, the two classes have the same input distribution and hence are not separable. Therefore, we need either d≠0 or e≠0.

***Hint1**: If* $\underline{\mathbf{x}} \sim \mathsf{N}_d\left(\underline{\boldsymbol{\mu}}, \Sigma\right)$*, then the pdf for x is given by:*

$$p\left(\underline{\mathbf{x}}\right) = \frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right)^T \Sigma^{-1}\left(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}\right)\right]$$

***Hint2**:*

For a 2×2 matrix,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

the matrix inverse is

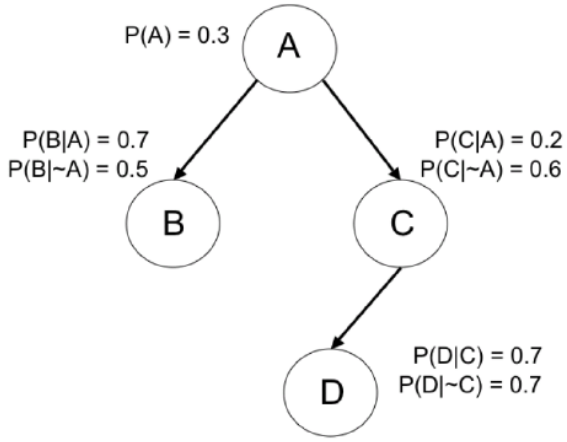
$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

***4) [30 points]** For the Bayesian network shown below, compute the following:*

a) P(A,B,C,D)
=P(D|C)P(C|A)P(B|A)P(A)
= 0.7 * 0.2 * 0.7 * 0.3 = 0.0294

b) P(A|B)
= P(B,A)/P(B)
=P(B|A)P(A)/(P(B|A)P(A)+P(B|~A)P(~A))
=0.7*0.3/(0.7*0.3+0.5*0.7)
=0.21/0.56
=0.375

c) P(C|B)
=P(C,B)/P(B)
=P(B)P(C)/P(B)
=P(C)
=P(C|A)P(A)+P(C|~A)P(~A)
=0.2*0.3 + 0.6*0.7
= 0.48


P(A) = 0.3
A
P(B|A) = 0.7
P(B|~A) = 0.5
P(C|A) = 0.2
P(C|~A) = 0.6
B
C
P(D|C) = 0.7
P(D|~C) = 0.7
D