Midterm Answers

Istanbul Technical University- Fall 2006
Pattern Recognition and Analysis (BBL514E)
Machine Learning (BLG527E)

Total worth: 25% of your grade.
Date: Wednesday, November 8, 2006.
Time: 120 mins
Good luck!

1.
1a) [15 pts] What is the VC (Vapnik-Chervonenkis) dimension of a hypothesis class? How and why is it used?
1b) [10 pts] What is k-fold cross-validation? How and why is it used?
----------Answer 1----------
1a) VC dimension of a hypothesis class is the maximum number of data points N that could be separated by a classifier from the hypothesis class for all 2^N different labelings of the data points. There has to be a classifier that could separate each labeling. The position of the data points can be chosen beforehand to make sure that all labelings could be separated.
VC dimension of a hypothesis class is used in VC inequality. The VC inequality is an upper bound on the test error of a classifier trained on N data points. The upper bound consists of the training error of the classifier, plus an additive term that depends on the VC dimension, number of training examples used and the probability that the inequality holds.

1b) k-fold cross validation is a technique that tries to compute an approximation of the test error on unseen test points. The whole training data is partitioned into k sets.

For i=1..k
    Train a classifier using all k partitions except the ith one.
    testError[i] = Test of the classifier using the ith partition.

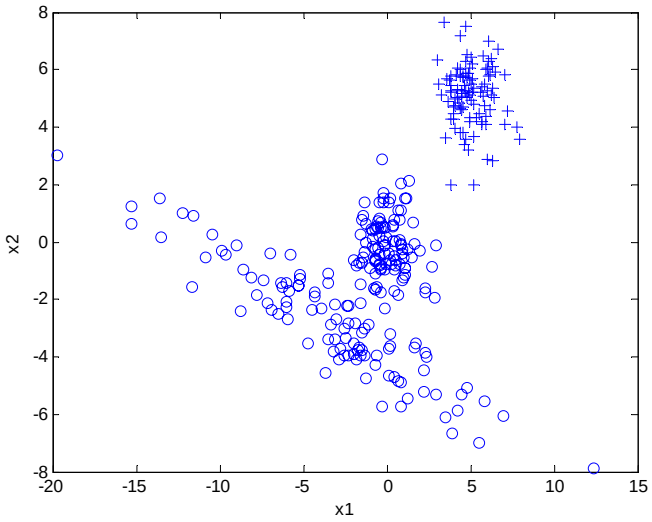
Report average of testError[] as the k-fold cross validation error.

k-fold cross validation is used to compute the best classifier parameters, such as the best degree of the polynomial for a polynomial classifier, for example.
----------Answer 1----------

2.

2a) [10 pts] Compare and contrast clustering and a dimensionality reduction technique (for example, principal component analysis) in terms of how they transform a certain set of unlabeled data set $X=\{\underline{x}^t\}_{t=1}^N$ where each $\underline{x}^t \in R^d$.

2b) [15 pts] Given the following set of labeled data with two classes (+, o), discuss which input preprocessing techniques you would use. Would you do your input pre-processing on all data or data on a specific class? Assume that your final classifier is a hyperplane, give a sketch of how you would construct your classifier.



----------Answer 2----------

2a) Dimensionality reduction reduces the dimensionality of inputs, hence our training set becomes:

$X=\{\underline{x}'^t\}_{t=1}^N$ where $\underline{x}'^t \in R^{d'}$ and $d' < d$.

On the other hand clustering groups all N data points into k clusters, reduces all N data points into k cluster centers, hence the data set becomes:

$X=\{\underline{x}'^t\}_{t=1}^k$ where $\underline{x}'^t \in R^d$ and $k < N$.

Clustering is useful if there are groups of data in d dimensional space, dimensionality reduction is useful if there are correlations between input dimensionalities.

2b) I would preprocess data for class o only. Data for class o seems to already have two clusters in them.

i) I would first cluster the data for class o into two clusters using mixture of two Gaussians and the EM algorithm. EM algorithm is needed because the cluster on the lower side of class o is clearly not spherical.

ii) Then I would do dimensionality reduction on the ellipse shaped cluster, because there seems to be correlation between input dimensionalities.

In order to produce the classifier, I would use multivariate classification and I would assume that the data for class o is a mixture of two Gaussians. I would compute the parameters of the Gaussians using the sample data.

----------Answer 2----------

3)

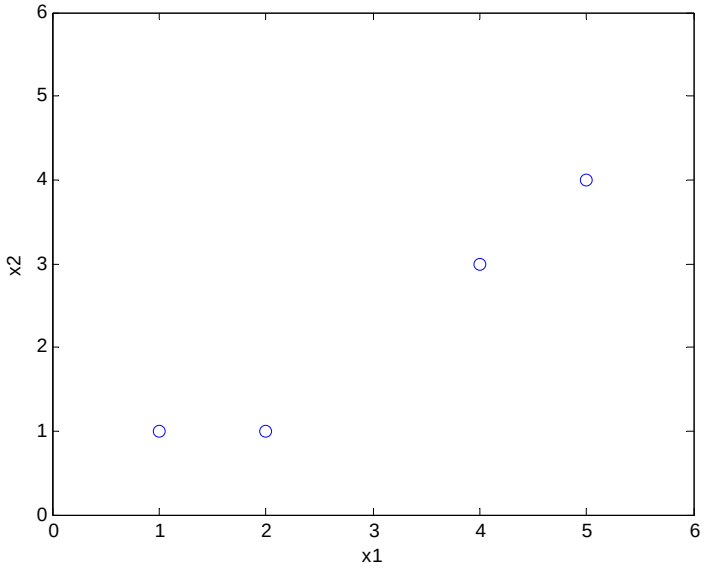
3a) [15 pts] Cluster the data shown below

$$X = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 5 \\ 4 \end{bmatrix} \right\}$$

using single link agglomerative hierarchical clustering and city block distance (sum of the absolute value of differences per dimension) as the distance measure.

Show each step of the clustering clearly.

What are the coordinates of the cluster centers (means) if you want to obtain two clusters?



3b) [10 pts] Compare k-means and hierarchical clustering in terms of running time and performance? Suggest ways to improve performance of k-means clustering.

----------Answer 3----------

3a)

Assume the following names for inputs: $x1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, x2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, x3 = \begin{bmatrix} 4 \\ 3 \end{bmatrix}, x4 = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$

Distance matrix between inputs is:

	$x1$	$x2$	$x3$	$x4$
$x1$	0	1	5	7
$x2$	1	0	4	6
$x3$	5	4	0	2
$x4$	7	6	2	0

G1 = {x1, x2}

G2 = {x3, x4}

G3 = {G1, G2}

Cluster center (mean) for G1 = $\begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$

Cluster center (mean) for G2 = $\begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}$

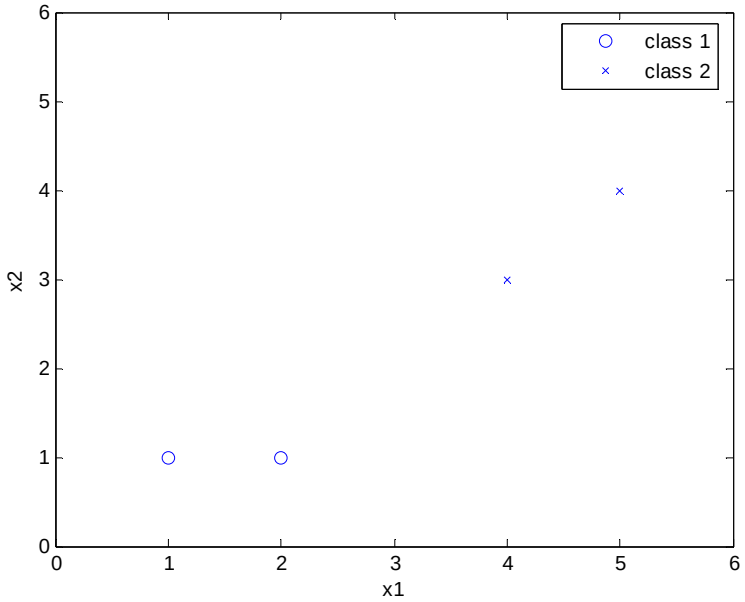
3b) kmeans has complexity O(n)*no of iterations whereas hierarchical clustering has complexity $O(n^2)$. Hence k-means is faster. k-means is highly affected by the initial choice of centers. Both methods are trying to find the optimum selection of cluster membership among a set of k^N possibilities, hence they could find local optimums as opposed to the global optimum.

k-means' performance can be improved by careful selection of initial centers and re-running the algorithm a number of times and then choosing the best clustering.

----------Answer 3----------

4) Consider the labeled data points given as follows:

$$X = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 4 \\ 3 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 5 \\ 4 \end{bmatrix}, 2 \right) \right\}$$



Assuming that inputs are normally distributed with class covariance matrices as follows:

$$S_1 = S_2 = s^2 I = s^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

4a) [20 pts] Compute the discriminant functions for both classes, $g_1(\underline{x})$ and $g_2(\underline{x})$.

4b) [5 pts] Compute and draw the discriminant function that separates the two classes.

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

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

Hint2: Use the log likelihood for the discriminant function.

Hint3: $(1.5)^2 = 2.25$, $(2.5)^2 = 6.25$, $(3.5)^2 = 12.25$, $(4.5)^2 = 20.25$

----------Answer 4----------

The discriminant function using log likelihood is:

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

For both classes 1 and 2, $P(w_i) = 2/4 = 0.5$

Using the simplifying assumption about the covariance matrices, the likelihood reduces to:

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

$$= \frac{1}{(2\pi)s} \exp\left[-\frac{1}{2s^2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)^T (\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)\right]$$

s2 for the first class is 0.25 and it is 0.5 for the second class. The average of them is 0.37.

Class means are $\boldsymbol{\mu}_1 = \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}$ and $\boldsymbol{\mu}_2 = \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}$

Plugging them into simplified formula above, we find log likelihood for each class as:

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

$$= \ln \frac{1}{(2\pi)s} \exp\left[-\frac{1}{2s^2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)^T (\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)\right] + \ln 0.5$$

$$= -\frac{1}{2s^2}(\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i)^T (\underline{\mathbf{x}} - \underline{\boldsymbol{\mu}}_i) + \ln \frac{1}{(2\pi)s} + \ln 0.5$$

In particular:

$$g_1(x) = -\frac{1}{2s^2}\left\|\underline{\mathbf{x}} - \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}\right\|^2 + \ln \frac{1}{(2\pi)s} + \ln 0.5$$

$$g_2(x) = -\frac{1}{2s^2}\left\|\underline{\mathbf{x}} - \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}\right\|^2 + \ln \frac{1}{(2\pi)s} + \ln 0.5$$

Where $\left\|\underline{\mathbf{x}} - \underline{\mu}\right\|^2 = (\underline{\mathbf{x}} - \underline{\mu})^T (\underline{\mathbf{x}} - \underline{\mu})$.

2b) The discriminant function that separates the two classes can be computed by equating the discriminant for both classes:

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

$$-\frac{1}{2s^2}\left\|\underline{\mathbf{x}} - \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}\right\|^2 + \ln \frac{1}{(2\pi)s} + \ln 0.5 = -\frac{1}{2s^2}\left\|\underline{\mathbf{x}} - \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}\right\|^2 + \ln \frac{1}{(2\pi)s} + \ln 0.5$$

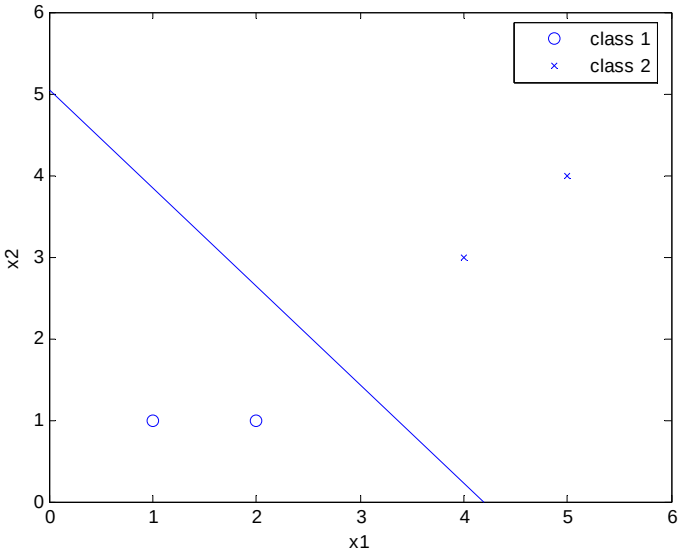
$$\left\|\underline{\mathbf{x}} - \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}\right\|^2 = \left\|\underline{\mathbf{x}} - \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}\right\|^2$$

$$-2\begin{bmatrix} 1.5 \\ 1 \end{bmatrix}^T x + \begin{bmatrix} 1.5 \\ 1 \end{bmatrix}^T \begin{bmatrix} 1.5 \\ 1 \end{bmatrix} = -2\begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}^T x + \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}^T \begin{bmatrix} 4.5 \\ 3.5 \end{bmatrix}$$

$$\begin{bmatrix} 6 \\ 5 \end{bmatrix}^T x - 25.25 = 0$$

$$6x_1 + 5x_2 - 25.25 = 0$$

We can draw this line easily by finding the intercept points:
When x1=0 , x2=25.25/5=5.05
When x2=0, x1=25.25/6=4.2



----------Answer 4----------