Modelling and visualization of high-dimensional data

Date: Monday 17th January 2011
Time: 09:45 - 11:45

Answer ALL Questions in Sections A and B and one question from Section C

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text.
SECTION A (Answer ALL questions)

(a) Draw schematic diagrams for two 2-D data sets of two classes such that (i) for the first data set, PCA and LDA find the same direction and (ii) for the second data set, PCA and LDA find totally different directions.

(4 marks)

(b) MDS can provide a visual representation of the pattern of proximities among a set of objects. You are given the information of air mileage among 100 different cities in the world. Justify how you can apply an MDS technique to produce a 2-D map that reflects their intrinsic geographic relationship among 100 cities. It is essential to give main steps in your solution.

(6 marks)

(c) Face recognition is an important task in biometrics. The use of salient features of a facial image often significantly improves the performance of a face recognition system in terms of both recognition accuracy and computational efficiency. You are asked to apply two dimensionality reduction techniques to generate two types of facial features (representations) that tend to capture salient facial features and further discuss their advantages and disadvantages. It is essential to give details on how to extract such features with your chosen techniques.

(10 marks)

[PTO]
SECTION C (Answer ONE question only from this section)

C1.
In Multiple Discriminant Analysis (MDA) involving $n$ $d$-dimensional data points and $C$ classes, expressions for Within-Class Scatter ($S_W$), Between-Class Scatter ($S_B$) and Total Scatter ($S_T$) are given by

$$S_W = \sum_{i=1}^{C} S_i$$
$$S_i = \sum_{x \in D_i} (x - m_i)(x - m_i)^T$$
$$S_B = \sum_{i=1}^{C} n_i(m - m_i)(m - m_i)^T$$
$$S_T = \sum_x (x - m)(x - m)^T$$
$$m = \frac{1}{n} \sum_{x \in D} x$$
$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

where $D_i$ is the set of $n_i$ samples belonging to class $\omega_i$, $\sum_{i=1}^{C} n_i = n$ and $n > d$.

(a) Prove that the Total Scatter ($S_T$) can be expressed as sum of $S_W$ and $S_B$. What is the interpretation for this equality? (8 marks)

(b) Extend what you have learned for binary discriminant analysis to give the criterion to be minimized in multiple discriminant analysis. (8 marks)

(c) Based on the criterion given in (b), what is a solution to the transformation matrix in multiple discriminant analysis? (4 marks)

C2.

(a) Define the linear transformation: $z = W^T(x - m)$, where $k$ columns of $W$ are the top $k$ principal components of a given data set, $\{x\}$, and $m$ is its mean of this data set. Prove that the covariance matrix of the projected data set, $\{z\}$, is diagonal. (8 marks)

(b) Give those criteria to be minimized in locally linear embedding (LLE) and explain why minimizing such criteria is viewed as “think globally, fit locally”. (8 marks)

(c) When the label information is available for a given data set, figure out an idea to incorporate such information into the LLE for supervised learning. (4 marks)

END OF EXAMINATION