CS 391D	Data Mining: A Mathematical Perspective	Fall 2009

Homework 2

Instructor: Inderjit Dhillon — Date Due: October 8, 2009

Keywords: *Probability, Principal Component Analysis, Classification*

Use Matlab for problem 2. Turn in your code along with the results in hard copy only. Note that the assignment is due IN CLASS.

1. (5 points) Suppose candy comes in three different flavors: *apple*, *cherry* and *orange*, which is sold in very large bags. There are known to be 3 kinds of bags but indistinguishable from outside: Bag type 1 (b_1) contains 30% *apple*, 40% *cherry* and 30% *orange*; bag type 2 (b_2) contains 50% *apple* and 50% *orange*; bag type 3 (b_3) contains 40% *apple*, 30% *cherry* and 30% *orange*. Let's assume that the manufacturers advertise a prior over the bag types: $p(b_1) = 0.2$, $p(b_2) = 0.2$, $p(b_3) = 0.6$. Any candy is selected with equal probability from any type of bag.

   (a) (1 point) What is the probability that you select first an *orange* candy when you randomly open a bag?

   (b) (2 points) If we observe that the selected candy is in fact an *orange*, what is the probability that it came from the bag type 2?

   (c) (2 points) What is the probability that the second candy selected from the same bag that you opened is an *orange* given the first one is an *orange*?

2. (6 points) This exercise will go through the procedure of Principal Component Analysis (PCA) and Fisher's Linear Discriminant Analysis (LDA). There are two datasets for this problem, which are available at `http://www.cs.utexas.edu/~wtang/cs391d/hw2data.tar.gz`. Both datasets contain three classes of instances and each class consists of 100 instances.

   (a) (2 points) Implement PCA and LDA in Matlab.

   (b) (2 points) For dataset 1, use PCA and LDA to project instances into two-dimensional subspace, respectively. Show the projected instances in separate plots. Identify the labels of instances in your plots. (Hint: You can use Matlab "`text`" command to notate each instance.)

   (c) (2 points) Repeat step (b) for dataset 2. What do you observe in your plots?

3. (4 points) In a class lecture, we showed that by modeling each class as a multivariate Gaussian with the same covariance matrix Σ and applying Bayes theorem, the decision surface for a 2-class problem is linear and given by

$$(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \Sigma^{-1} \boldsymbol{x} + c = 0,$$

   for some constant c.

   When $\Sigma = I$, the separating hyperplane $\boldsymbol{w}^T\boldsymbol{x} + c = 0$ has $\boldsymbol{w} = \boldsymbol{m}_2 - \boldsymbol{m}_1$, i.e., the normal to the separating hyperplane is parallel to $\boldsymbol{m}_2 - \boldsymbol{m}_1$.

   (a) (2 points) Show that, given a non-singular Σ, the normal to the separating hyperplane $\boldsymbol{w}^T\boldsymbol{x} + c = 0$ cannot be exactly perpendicular to $\boldsymbol{m}_2 - \boldsymbol{m}_1$, i.e., $\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) \neq 0$.

   (b) (2 points) Construct an example where the normal to the separating hyperplane is almost perpendicular to $\boldsymbol{m}_2 - \boldsymbol{m}_1$, i.e., $\boldsymbol{w}^T(\boldsymbol{m}_2 - \boldsymbol{m}_1) \approx 0$. Show a picture to visualize your example. Assume that $\boldsymbol{m}_1$ is well separated from $\boldsymbol{m}_2$ (say $\|\boldsymbol{m}_2 - \boldsymbol{m}_1\| = O(1)$).