CS 688 – Spring 2016

Homework 2 – Due March 4

Professor: Carlotta Domeniconi

Problem 1 Solve exercises 4.5 and 4.8 of the textbook.

Problem 2 Consider three classes C_1, C_2, and C_3 each having a bivariate normal distribution with identity covariance matrices and means $(-2,0)^T$, $(0,0)^T$, and $(0,2)^T$, respectively. Show that the classifier that minimizes the misclassification error has decision boundaries which are piecewise linear. Assume equal class prior probabilities. Draw the decision boundary.

Now define a class, A, as a mixture of C_1 and C_3 as follows:

$$p(\mathbf{x}|A) = 0.5p(\mathbf{x}|C_1) + 0.5p(\mathbf{x}|C_3)$$

and a class B as bivariate normal with identity covariance matrix and mean $(a,b)^T$, for some a and b. What is the equation of the Bayes' decision boundary (i.e., the optimal decision boundary that minimizes the misclassification error for classes A and B)? Assume equal class prior probabilities. Under what conditions is it piecewise linear?

Problem 3 Show that for $K > 2$ classes, under the assumption of Gaussian class-conditional densities with equal covariance matrices, the posterior probability $p(C_k|\mathbf{x}_n)$ can be written as follows:

$$p(C_k|\mathbf{x}_n) = \frac{e^{\mathbf{w}_k^T\mathbf{x}+w_{k0}}}{\sum_j e^{\mathbf{w}_j^T\mathbf{x}+w_{j0}}}$$

where $\mathbf{w}_k = \Sigma^{-1}\boldsymbol{\mu}_k$ and $w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T\Sigma^{-1}\boldsymbol{\mu}_k + \ln p(C_k)$.

Problem 4 Implement the learning algorithm for the Perceptron discussed in class (on-line version). You can use your favorite programming language to implement the algorithm.

Download the data (from the class webpage) to be used for training and testing of the perceptron. The folder contains ten training sets and one test set. (The file names for the training sets are: *set1.train*, ..., *set10.train*; the file name for the test set is *set.test*.) Each training set contains 40 two-dimensional points, each associated with the corresponding class label (either 0 or 1). The test set contains 2000 points. Points in the test set are also associated with the corresponding class label (to compute misclassification errors). For all given sets, points of one class are drawn from the same underlying distribution. The structure of each file is as follows:

x_1 y_1 $label_1$
x_2 y_2 $label_2$
...

$x_i \quad y_i \quad label_i$

Use each training set to train the perceptron, then use the test data to compute the misclassification error. Initialize the weights randomly before each training phase. This process will give you ten trained perceptrons and 10 error rates. Plot each training set and corresponding decision boundary. Also plot each learned decision boundary (separately) along with the test set. Use your favorite program to do the scatterplots (e.g., Matlab, Gnuplot, Excel, etc.). For each training set, report the number of times (iterations) you provided the sequence as input to the perceptron during training.

Comment your results. What can we say about the training data sets? Does the solution found by the perceptron change for different training data? How about the misclassification error? Motivate your findings.

Instructions to submit Turn in a hard copy of your solutions by Friday, March 4 at 5PM. Slide it under the door of my office. Turn in the source code of Problem 4 electronically as an attachment by email to carlotta@cs.gmu.edu. Make sure your code compiles and runs properly. The source code is also due by Friday, March 4 at 5PM. No extension will be granted.