EEL4930/EEL5840 Fall 2016 – Homework 4
Bayes Classifier and Linear Discriminant Analysis

October 20, 2016

Due: October 27, 2016, 11:59 PM

Instructions

For this homework, please show any plots and tables. Label your plots' axes and include plot titles. As well, state all assumptions that you made. Do not include code. You should mention whether you programmed the solutions yourself or if you downloaded a package online and from which website.

Remember that commenting your results is very important. It is expected of you to systematically discuss your results. If no explanation is given, your grade will be penalized.

Your homework submission must cite any references used, including articles, books, code, websites, and personal communications). All solutions must be written in your own words, and you should program the algorithms yourself. If you do work with others, you must list the people you worked with. Submit your solutions as a single PDF file to the course website at http://elearning.ufl.edu/.

If you have any questions, then address them to:

- Catia Silva (TA) – catiaspsilva@ufl.edu
- Isaac Sledge (TA) – isledge@ufl.edu

Problems

In this homework, you will be implementing two different types of classifiers to distinguish between species of rock crabs of genus Leptograpsus. The feature data are provided on the eLearning website.

The dataset is composed of 200 samples of different crab specimens. There are seven features that were captured for each specimen. These features include anatomical properties: the front lip, rear width, length, width and depth of the grab. There are two features that characterize the gender of the grab. Your goal is to discriminate between the two species of crab using these provided features. The species of crab are given as binary labels in the first column of the provided dataset. The remaining seven columns of data are the real-valued features. You will be considering two classifiers: the Bayes classifier and a classifier found from linear discriminant analysis.

Complete the following tasks:

1) (5 points) Implement the Bayes classifier, under the assumption that your likelihood model $p(x|j)$ is unimodal Gaussian and that the prior probabilities $p(j)$ are dictated by the number of samples $n_j \in \mathbb{R}$ that you have for each class. This classifier is given by the following discriminant function for each class j

$$g_j(x) = -\frac{1}{2}(x-\mu_j)^\top \Sigma_j^{-1}(x-\mu_j) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log(\det(\Sigma_i)) + \log(p(j)).$$

Here, $x \in \mathbb{R}^7$ is the feature observation, while $j \in \mathbb{R}$ is the class. For this classifier, we assume that each class j can have an arbitrary mean $\mu_j \in \mathbb{R}^7$ and an arbitrary covariance matrix $\mathbb{R}^{7\times 7}$. Both of these quantities are to be estimated from the observations in each class j as follows:

$$\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^j. \qquad \Sigma_i = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^\top,$$

where $n_j \in \mathbb{R}$ is the number of observations in class j and $x_i^j \in \mathbb{R}^7$ is the ith sample from class j. Break apart the observations into training and testing sets. Use the first 70% of the data for training (first 140 samples) and the remaining 30% of the data for testing (remaining 60 samples). Provide a class confusion matrix. For each pair of features, also provide a plot of the class decision boundaries. Lastly, record the amount of time, in seconds, needed for you to classify the entire dataset (200 samples).

2) (5 points) Implement a classifier based on linear discriminant analysis. This is a particular case of the Bayes classifier. For this type of classifier, the discriminant function is given by

$$g(x):\ (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)\cdot x > \frac{1}{2}(-\mu_0^\top \Sigma_0^{-1}\mu_0 + \mu_1 \Sigma_1^{-1}\mu_1).$$

Here, $\mu_0, \mu_1 \in \mathbb{R}^7$ are the class means for the two classes and $\Sigma_0, \Sigma_1 \in \mathbb{R}^{7\times 7}$ are the class covariances for the two classes. These variables are to be estimated as follows for each class

$$\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^j. \qquad \Sigma_i = \frac{1}{n_j}\sum_{i=1}^{n_j}(x_i^j - \mu_j)(x_i^j - \mu_j)^\top.$$

Break apart the observations into training and testing sets. Use the first 70% of the data for training (first 140 samples) and the remaining 30% of the data for testing (remaining 60 samples). Provide a class confusion matrix. For each pair of features, also provide a plot of the class decision boundaries. Lastly, record the amount of time, in seconds, needed for you to classify the entire dataset (200 samples).