Homework 5

October 23, 2017

Due: October 30, 2017, 11:59 PM EST

Instructions

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 must program the algorithms yourself. If you do work with others, you must list the people you worked with. Submit your solutions as a PDF to the E-Learning at UF (http://elearning.ufl.edu/).

Your programs must be written in either MATLAB or Python. The relevant code to the problem should be in the PDF you turn in. If a problem involves programming, then the code should be shown as part of the solution to that problem. If you solve any problems by hand just digitize that page and submit it (make sure the problem is labeled).

If you have any questions address them to:

- Catia Silva (TA) – catiaspsilva@ufl.edu
- Sheng Zou (TA) – shengzou@ufl.edu

Problems - 10 points

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 e-Learning 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 crab. There are two additional features that characterize the gender of the crab. Your goal is to discriminate between the two species of crab using these provided features (total of 7). 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 $n_j \in \mathcal{R}$ that you have for each class. This classifier is given by the following discriminant function for each class j.

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

Here, $x \in \mathbb{R}^d$, $d = 7$, 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 $R^{7\times 7}$. Both of these quantities are to be estimated from the observations in each class j as follows:

$$\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_i^j \quad (2)$$

and

$$\Sigma_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(x_i^j - \mu_j\right)\left(x_i^j - \mu_j\right)^T \quad (3)$$

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 the training set, the testing set, and the entire dataset. As well, 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

$$g(x) : \left(\frac{\Sigma_0 + \Sigma_1}{2}\right)^{-1} (\mu_1 - \mu_0) \cdot x > \frac{1}{4}\left(\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1 + T\right) \tag{4}$$

or, alternatively,

$$g(x) : (\Sigma_0 + \Sigma_1)^{-1} (\mu_1 - \mu_0) \cdot x > \frac{1}{4}\left(\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_1^{-1} \mu_1 + T\right) \tag{5}$$

Here, $T \in \mathbb{R}$ is a potentially non-zero threshold value that you should choose, in some fashion, on your own. The variables $\mu_0, \mu_1 \in \mathbb{R}^7$ 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

$$\mu_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_i^j \tag{6}$$

and

$$\Sigma_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \left(x_i^j - \mu_j\right)\left(x_i^j - \mu_j\right)^T \tag{7}$$

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 the training set, the testing set, and the entire dataset for the best value of T that you found. As well, record the amount of time, in seconds, needed for you to classify the entire dataset (200 samples).