Homework 7

November 9, 2017

Due: November 21, 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

Materials and Methods

This is the last homework to be assigned in this course and will be worth 10% of your final grade.

In this homework, you will be implementing feature selection methods and the Expectation-Maximization (EM) algorithm. Code the algorithms on your own and include your code with the homework solution. You will be testing your implementations with two different data sets: "WQDataSet_HW7.zip" for problem 1 and "GMDataSet_HW7.txt" for problem 3.

- "WQDataSet_HW7.zip" is a wine quality data set of two different types of wine (red and white wine). This folder contains two files: "WhiteWine_HW7.txt" and "RedWine_HW7.txt". The data set is composed of 4898 white wine examples ("WhiteWine_HW7.txt") and 1599 red wine examples ("RedWine_HW7.txt"). There are 13 real-valued features that were captured for each wine type. These features include: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality scored between 0 and 10, and rate. Your goal is to discriminate the two types of wine using the most informative features out of these 13 features.
- "GMDataSet_HW7.txt" is a data set containing a high-dimensional mixture of Gaussian distributions. Your goal is to determine the mean and (diagonal) covariance associated with each of the mixture components.

Problem 1 (10 points)

For this problem you are given a high-dimensional Gaussian mixture data set "WQDataSet_HW7.zip". Implement **Forward Feature Selection (FFS)** and **Backward Feature Selection (BFS)** methods using the **Fisher discriminant ratio** to select the features to discriminate the two types of wine.

(1) How many features should you retain based on the FFS and BFS approaches?

(2) Does classification performance using the **Bayes Classifier** have best performance using the feature set you determined using FFS and BFS? Why or why not? How did you determine this?

Problem 2 (10 points)

For this problem you are given a high-dimensional Gaussian mixture data set "GMDataSet_HW7.txt". Implement the **EM algorithm** to determine the mixture proportion, mean and (diagonal) covariance associated with each of the mixture components.

(i) How many mixture components are found in the data? How did you determine this?

(ii) What did you estimate for the mixture proportions, means and covariances associated with each mixture?