EE655000 Machine Learning HW1

TA: Hsueh-Tse Lin, Weng-Tai Su, Hsin-Chieh Wang

ranayukirin1991@gmail.com
wengtai2008@hotmail.com
lucas85062055@gmail.com

Deadline: 4/20 (Mon)

Grading Policy:

1. In the handwriting assignment, you need to provide detailed derivations. Partial points will be credited when a wrong answer is accompanied by correct reasoning. Please hand in the handwriting assignment in class on 4/20. If online teaching is on, please hand in handwriting assignment to EECS 804.

2. In the programing assignment, the code, test data and report should be compressed into a **ZIP** file and upload to iLMS website. Also, please write a Readme file to explain how to run your code and discuss characteristics in your report. The report format is not limited.

3. The programming language that can be used on this assignment includes Python and MATLAB. Built-in machine learning libraries or functions (like sklearn.linear_model) are NOT allowed to use except for PCA.

4. Discussions are encouraged, **but plagiarism is strictly prohibited.**

Part 1. Handwriting homework assignment

You can find the corresponding problems from the textbook.

1. Exercise 2.26 (30 points)

2. Exercise 3.6 (40 points)

3. Exercise 3.11 (30 points)

Part 2. Computer assignment

This dataset [1] is the result of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

That is, there are 3 types of wines and 13 different features of each instance. In this problem, you will implement the Maximum A Posteriori probability (MAP) of the classifier for 54 instances with their features.

The dataset is provided in wine.csv. There are total 178 instances in wine.csv. The first column is the label (1, 2, 3) of type and other columns are the detailed values of each feature.

Information of each feature:

1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

Assume that all the features are independent and the distribution of them is Gaussian distribution.

1. (30 points)

   To split the train and test data from the provided wine.csv. It is necessary to know how to read and write the csv file. Thus, you need to randomly split 18 instances from each type as testing dataset and totally 54 instances from the whole dataset. Then save the training dataset as train.csv and testing dataset as test.csv.

(124 instances for training and 54 instances for testing.)

2. (50 points)

   To evaluate the posterior probabilities, you need to learn likelihood functions and prior distribution from the training dataset. Then, you should calculate the error rate or accuracy rate of the MAP detector by comparing to the label of each instance in the test data. Note that, the accuracy rate will be different depends on the random result of splitting data, but it should exceed 90% overall.

   (Please add corresponding comments in your code to describe how you obtain the posterior probability.)

3. (20 points)

   Please discussion characteristics and plot the visualized result of testing data in your report.

   (You can directly use built-in PCA function to get visualized result.)

[1] https://archive.ics.uci.edu/ml/datasets/Wine