EE655000 Machine Learning HW1

TA: 賴清元 y8574317@gmail.com
陳冠宇 seed0123456@gmail.com
林辰安 strike860930@gmail.com

Deadline: 2022/04/07 (THU.)

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 4/7.
2. In the programming assignment, the code, test data and report should be compressed into a ZIP file and uploaded to eeclass 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. You are required to finish this homework with Python 3. Moreover, built-in machine learning libraries or functions (like sklearn.linear_model) are NOT allowed to use.
   You can use dimension reduction functions such as *sklearn.decomposition.PCA* for better visualization in discussion.
4. Discussions are encouraged, but plagiarism is strictly prohibited (changing variable names, etc.). You can use any open source with clearly mentioned in your report. If there is any plagiarism, you will get 0 in this homework.

Part 1. Handwriting homework assignment:

Text book exercise from

Bishop - Pattern Recognition And Machine Learning - Springer 2006

Exercise 1.7 (10%)

Exercise 2.26 (10%)

Exercise 2.27 (10%)

Exercise 3.6 (10%)

Exercise 3.11 (10%)

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 a 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. Non Flavonoid 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. (10%) 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. (25%) To evaluate the posterior probabilities, you need to learn likelihood functions and prior distribution from the training dataset. Then, you should calculate the 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 depending 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. (5%) Please plot the visualized result of testing data in your report. (You can directly use the built-in PCA function to get visualized result.)
4. (10%) Please discuss the effect of prior distribution on the posterior probabilities in your report.

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