CSCI 3320: Fundamental of Machine Learning **2018-2019 Term 2**

Programming Assignment 1

Instructor: Prof. John C.S. Lui Due: 23:59 on Sun. Mar. 10th, 2019

1 Introduction

In Programming Assignment 1, you are required to do the following:

- Write a Python program with pandas (or any other packages) to process four input files.
- **Implement your own classifier** using the parametric methods we discussed in class and please do not use any learner from scikit-learn.

1.1 File Descriptions

To start, you need to download the `asgn1.zip` file from the course website. In `asgn1.zip`, we provide the following files for you:

- `input_1.csv`: contains the training and testing data for Problem 1.
- `input_2.csv`: contains the training and testing data for Problem 2.
- `input_3.csv`: contains the training and testing data for Problem 3.
- `input_4.csv`: contains the training and testing data for Problem 4.

Note: The details will be discussed in each problem.

2 Problem 1(25%)

In this programming exercise, you are asked to do classification via the parametric method we learnt in the lecture.

You need to read in a csv file, input_1.csv. The attributes of this file are: feature_value and class #. The feature values are outcomes from a Bernoulli distribution. In other words, the feature values will be either 0 or 1. These feature values came from two classes ($C = 1$ and $C = 2$). Use the first 80% of the inputs as training data and the remaining 20% for testing the accuracy of your prediction.

To accomplish this task, you have to perform the "parametric estimation" of p_i for class C_i, where $i = 1, 2$. While p_i is the probability of having an outcome 1 for class i.

You need to perform the following:

- Based on the input training data, compute the priors of C_1 and C_2.
- Perform the parametric estimation on the input training data for p_1 and p_2.
- Use the prior of C_i and the probability mass function of p_i to define discriminant functions $g_i()$ for $i = 1, 2$.
- Perform the testing of your classification using the two discriminant functions.
- Output the confusion matrix and save it in the **report.pdf** file.
- Output the (1) accuracy, (2) precision, (3) recall, (4) f1 score for each class as well as the average f1 score for the classification task and save them in the **report.pdf** file.
- Save your python script and name it as **p1.py**.

3 Problem 2(25%)

In this programming exercise, you continue to do classification using the parametric method.

You need to read in a csv file, input_2.csv. The attributes of this file are: feature_value and class #. The feature values are outcomes from a Gaussian distribution. In other words, the feature values will be some real numbers. These feature values came from two classes ($C = 1$ and $C = 2$). Use the first 80% of the inputs as training data and the remaining 20% for testing the accuracy of your prediction.

To accomplish this task, you have to perform the "parametric estimation" of m_i for class σ_i^2, where $i = 1, 2$ and m_i and σ_i^2 are the estimated mean and variance for class i.

You need to perform the following:

- Based on the input training data, compute the priors of C_1 and C_2.
- Perform the parametric estimation on the input training data for m_i and σ_1^2 for $i = 1, 2$.
- Use the prior of C_i and the probability density function of Gaussian distribution to define two discriminant functions $g_i()$ for $i = 1, 2$.
- Perform the testing of your classification using the two discriminant functions.
- Output the confusion matrix and save them in the **report.pdf**.
- Output the (1) accuracy, (2) precision, (3) recall, (4) f1 score for each class as well as the average f1 score for the classification task and save them in the **report.pdf** file.
- Save your python script and name it as **p2.py**

4 Problem 3(25%)

In this programming exercise, you continue to do classification using the parametric method.

You need to read in a csv file, input_3.csv. The attributes of this file are: feature_value and class #. The feature values are outcomes from a Gaussian distribution. In other words, the feature values will be some real numbers. These feature values came from four classes ($C = 1$, $C = 2$, $C = 3$, $C = 4$). Use the first 80% of the inputs as training data and the remaining 20% for testing the accuracy of your prediction.

To accomplish this task, you have to perform the "parametric estimation" of m_i for class σ_i^2, where $i \in \{1, 2, 3, 4\}$ and m_i and σ_i^2 are the estimated mean and variance for class i.

You need to perform the following:

- Based on the input training data, compute the priors of C_i where $i \in \{1, 2, 3, 4\}$.
- Perform the parametric estimation on the input training data for m_i and σ_1^2 for $i = 1, 2, 3, 4$.
- Use the prior of C_i and the probability density function of Gaussian distribution to define four discriminant functions $g_i()$ for $i = 1, 2, 3, 4$.
- Perform the testing of your classification using the four discriminant functions.
- Output the confusion matrix and save it in the **report.pdf** file.
- Output the (1) accuracy, (2) precision, (3) recall, (4) f1 score for each class as well as the average f1 score for the classification task and save them in **report.pdf** file.
- Save your python script and name it as **p3.py**

5 Problem 4(25%)

In this programming exercise, you continue to do multi-features classification using the parametric method.

You need to read in a csv file, input_4.csv. The attributes of this file are: feature_value_1, feature_value_2 and class #. The first feature values are some real numbers from a Gaussian distribution while the second feature values are outcomes from a Bernoulli distribution. These feature values came from two classes ($C = 1$ and $C = 2$). Use the first 80% of the inputs as training data and the remaining 20% for testing the accuracy of your prediction.

To accomplish this task, you have to perform the "parametric estimation" of p_i, m_i and σ_i^2 for class i where $i = 1, 2$, and p_i, m_i and σ_i^2 are the probability of having a 1 for class i, estimated mean and estimate variance for class i respectively.

You need to perform the following:

- Based on the input training data, compute the priors of C_1 and C_2.
- Perform the parametric estimation on the input training data for p_i, m_i and σ_1^2 for $i = 1, 2$.
- Use the prior of C_i, the probability mass function of Bernoulli and the probability density function of Gaussian distribution to define two discriminant functions $g_i()$ for $i = 1, 2$.
- Perform the testing of your classification using the two discriminant functions.
- Output the confusion matrix and save it in the **report.pdf** file.
- Output the (1) accuracy, (2) precision, (3) recall, (4) f1 score for each class as well as the average f1 score for the classification task and save them in in the **report.pdf** file.
- Save your python script and name it as **p4.py**

6 Submission

Instructions for the submission are as follows. **Please follow them carefully.**

1. Make sure you have answered all questions in your report.
2. Test all your Python scripts before submission. Any script that has syntax error will not be marked. Also we recommend you to use Python 3 and Linux environment because we will run your scripts with such settings.
3. Zip all Python script files, i.e., the `*.py` files in `asgn1.zip` (Please do not change the filenames of the scripts.) and your report (`report.pdf`) into a single zipped file named `<student-id>_asgn1.zip`, where `<student-id>` should be replaced with your own student ID. e.g., `1155012345_asgn1.zip`
4. Submit the zipped file `<student-id>_asgn1.zip` to CUHK Blackboard System `https://blackboard.cuhk.edu.hk` no later than 23:59 on Sun. Mar. 10th, 2019.