CSCI 3320: Fundamental of Machine Learning **2019-2020 Term 2**

Programming Assignment 0

Instructor: Prof. John C.S. Lui Due: 23:59 on Sunday, Mar. 8th, 2020

1 Introduction

In Programming Assignment 0, you are required to do the following:

- Write a Python program with pandas (or any other packages) to process three 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.

Note: The details will be discussed in each problem.

2 Problem 1(30%)

In this programming exercise, you are asked to do classification via the parametric method we learned 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 three classes ($C = 1$, $C = 2$ and $C = 3$). 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 $i \in \{1, 2, 3\}$. While p_i is the probability of having an outcome 1 for class i.

You need to answer the following questions and save the answer to **report.pdf**:

1. Based on the input training data, what are the priors of C_1, C_2 and C_3?

2. What are the estimated p_1, p_2 and p_3, based on your parametric estimation on the input training data?

3. After defining the discriminant functions $g_i()$ for $i \in \{1, 2, 3\}$, which are based on your previous answers, please perform the testing of your classification using the discriminant functions. What is your confusion matrix?

4. What are the **accuracy**, **precision**, **recall** and **f1 score** for each class, as well as the average f1 score for the classification task?

5. Save your python script and name it as **p1.py**.

3 Problem 2(30%)

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 three classes ($C = 1$, $C = 2$ and $C = 3$). 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 and σ_i^2 for class i, where $i \in \{1, 2, 3\}$. m_i and σ_i^2 are the estimated mean and variance for class i.

You need to answer the following questions and save the answer to **report.pdf**:

1. Based on the input training data, what are the priors of C_1, C_2 and C_3?

2. What are the estimated m_i and σ_i^2, for $i \in \{1, 2, 3\}$?

3. After defining the discriminant functions $g_i()$ for $i \in \{1, 2, 3\}$, which are based on your previous answers, please perform the testing of your classification using the discriminant functions. What is your confusion matrix?

4. What are the **accuracy**, **precision**, **recall** and **f1 score** for each class, as well as the average f1 score for the classification task?

5. Save your python script and name it as **p2.py**.

4 Problem 3(40%)

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

You need to read in a csv file, input_3.csv. The attributes of this file are: feature_value_1, feature_value_2 and class #. The first feature values are outcomes from a Bernoulli distribution while the second feature values are some real numbers from a Gaussian distribution. These feature values came from two classes ($C = 1$, $C = 2$ and $C = 3$). 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 $i \in \{1, 2, 3\}$. p_i, m_i and σ_i^2 are the probability of having a 1 for Bernoulli distribution in class i, estimated mean and estimate variance for Gaussian distribution in class i respectively.

You need to answer the following questions and save the answer to **report.pdf**:

1. Based on the input training data, what are the priors of C_1, C_2 and C_3?
2. What are the estimated p_i, m_i and σ_i^2, for $i \in \{1, 2, 3\}$?
3. After defining the discriminant functions $g_i()$ for $i \in \{1, 2, 3\}$, which are based on your previous answers, please perform the testing of your classification using the two discriminant functions. What is your confusion matrix?
4. What are the **accuracy**, **precision**, **recall** and **f1 score** for each class, as well as the average f1 score for the classification task?
5. Save your python script and name it as **p3.py**.

5 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, and your report in `<student-id>_asgn1.zip`, where `<student-id>` should be replaced with your own student ID,

   e.g., `1155012345_asgn1.zip` (Please do not change the filenames of the scripts.)

4. Submit the zipped file `<student-id>_asgn1.zip` to CUHK Blackboard System `https://blackboard.cuhk.edu.hk` no later than 23:59 on Sunday, Mar. 8th, 2020.