CS 3753 & 5163 Data Science Summer 2020

Homework 5 (100 points)

Submission:

1. submit a single python script (abc123_hw#.ipynb or abc123_hw#.py) through blackboard. All the results are outputted from your Python code.
2. You should have the instruction of running your code at the beginning of your code. It should run successfully either in the basic Python3 environment or in Jupyter Notebook.
3. Do not compress your files and make sure all your files are in the same folder.
4. The late submission will lose 15% points. Your code should run successfully. There is a limit of half points max if the code cannot run.
5. You can submit your homework 3 times before the deadline.

Questions

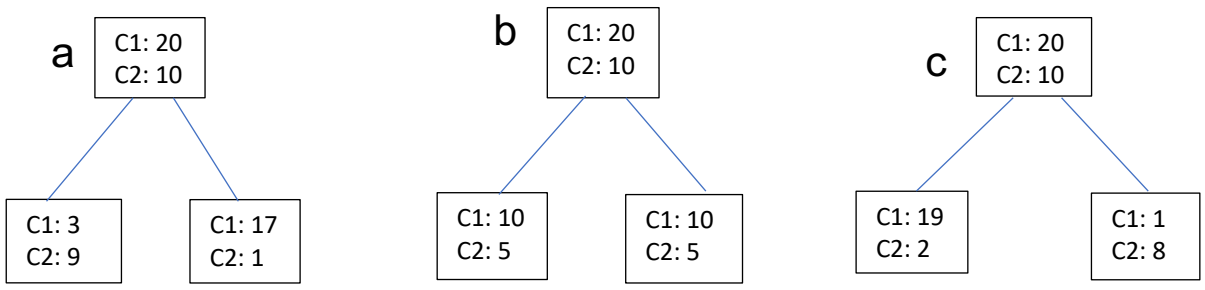
1. (20 points) We do the linear regression on three points (0.5, 1), (2, 2.5), and (3, 3). Please calculate the SSEs of the four linear regression. Which is the best linear regression using SSE? Output all steps through your python code.
   (a) y = x + 0.5
   (b) y = x + 1
   (c) y = 0.8*x + 0.3
   (d) y = 0.8*x + 0.7

2. (20 points) What is the problem solved by Lasso and Ridge regression? What is the major difference between the two regression? Please discuss the advantages and disadvantages of them.

3. (30 pints) Decision Tree (30 points)
   There are various ways to decide on the metric to choose the variable on which splitting for a node is done. Different algorithms deploy different metrices to decide which variable splits the dataset best.

   Let's say we have a sample of 30 records. There are two classes C1 and C2. We have three possible splits a, b, and c (see figure below). The number of records in each class is shown in every node.

   a. Write a Python code to measure the node impurity using Gini Index, Entropy Gain, and Misclassification Error, respectively.
   b. Evaluate the quality of the three splitting and report the best one.
   c. The code should print out the quality of each splitting and your best choice. You also can print out any other information freely.
   d. Finally, print out your conclusion about whether all three methods have the same best choice.



4. (30 points) KNN: this section applies the KNN algorithm to the Iris flowers dataset. The first step is to load the dataset in “iris.csv” and convert the loaded data to numbers that we can use with the mean and standard deviation calculations. For this we will use the helper function load_csv() to load the file, str_column_to_float() to convert string numbers to floats and str_column_to_int() to convert the class column to integer values.

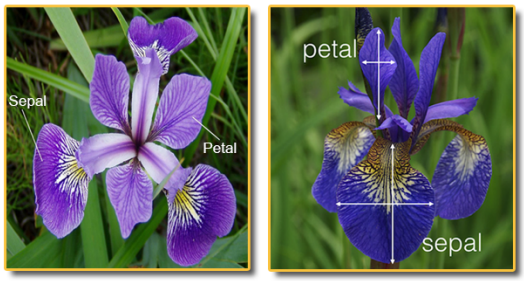
   We will evaluate the algorithm using k-fold cross-validation with 5 folds. This means that 150/5=30 records will be in each fold. We will use the helper functions evaluate_algorithm() to evaluate the algorithm with cross-validation and accuracy_metric() to calculate the accuracy of predictions.

   A new function named k_nearest_neighbors() was developed to manage the application of the KNN algorithm, first learning the statistics from a training dataset and using them to make predictions for a test dataset.

   Download the dataset and save it into your current working directory with the filename “iris.csv“. The Iris Flower Dataset involves predicting the flower species given measurements of iris flowers.

   It is a multiclass classification problem. The number of observations for each class is balanced. There are 150 observations with 4 input variables and 1 output variable. The variable names are as follows:

   a. Sepal length in cm.
   b. Sepal width in cm.
   c. Petal length in cm.
   d. Petal width in cm.
   e. Class



The mean accuracy is around 97%