Assignment 3

CMPT310: Artificial Intelligence Survey

Due date: March 20, 2022 11:59PM

Academic Dishonesty: We will be checking your code against other submissions in the class for logical redundancy. If you copy someone else's code and submit it with minor changes, we will know. These cheat detectors are quite hard to fool, so please don't try. We trust you all to submit your own work only; please don't let us down. If you do, we will pursue the strongest consequences available to us.

Getting Help: You are not alone! If you find yourself stuck on something, contact the course staff for help. Office hours and the discussion forum are there for your support; please use them. If you can't make our office hours, let us know and we will schedule more. We want these projects to be rewarding and instructional, not frustrating and demoralizing. But, we don't know when or how to help unless you ask.

Discussion: Please be careful not to post spoilers.

Question 1 (10 points)

(A) Explain about Random Forest in machine learning. (2 points)

(B) Use the below code to create a random classification dataset including the training and test data. Write a code snippet to train a Random Forest model on the training data, i.e. `X_train`. Hint: You can use `RandomForestClassifier` from `Scikit-learn` package. (2 points)

```
from sklearn.datasets import make_classification
from sklearn.model_selection import train_test_split
# Do not change anything in this code
X, y = make_classification(
    n_samples=1000,
    n_features=12,
```

```
    n_informative=4,
    n_redundant=0,
    n_repeated=0,
    n_classes=2,
    random_state=5,
    shuffle=False)
X_train, X_test, y_train, y_test =
        train_test_split(X, y, stratify=y, random_state=43)
```

(C) The trained random forest using `RandomForestClassifier` function contains some values, called feature importance. Explain what the feature importance is and plot the values as a bar chart. Write the code used to get the feature importance and the code to generate the bar plot. (4 points)

(D) Apply the trained RF model on test data, i.e. `X_test`, and report the accuracy of the model. Hint: Use `score` function. (2 points)

Question 2 (20 points)

Assume that we have the below data points where each point belongs to one of two classes: + and *o*.

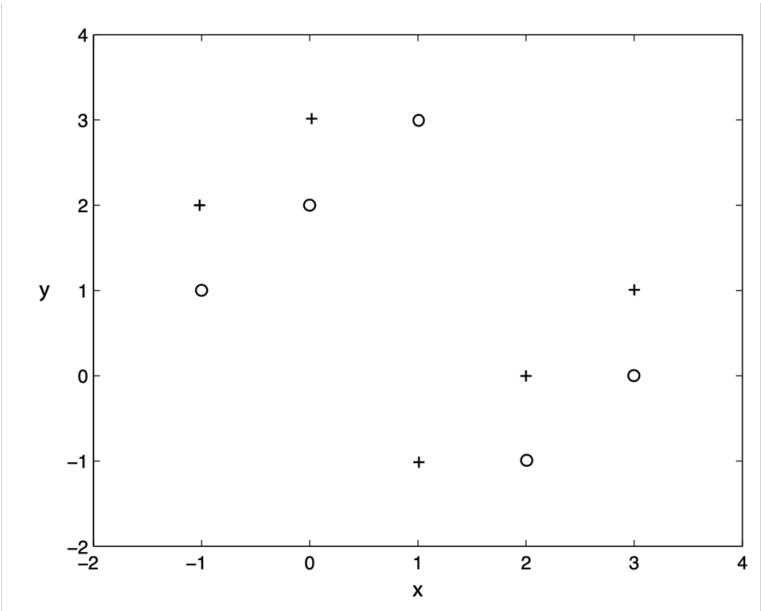

Figure 1: Data points.

(A) Write a Python code to classify the data points using the K-NN algorithm con-

sidering *Euclidean distance* as its metric. You are NOT allowed to use python packages other than numpy in this question. Please write a clean code and leave some comments. (15 points)

(B) Which of the following values of k leads to the minimum leave-one-out cross validation error: 1, 3, 5, 7 or 9? What is the error for that k? Error is measured as $\frac{number\ of\ incorrectly\ classified\ data\ points}{total\ number\ of\ data\ points}$. (5 points)

Question 3 (7 points)

Consider the following data set comprised of three binary input attributes (A1, A2, and A3) and one binary output:

Example	A_1	A_2	A_3	Output y
$\mathbf{x}_1$	1	0	0	0
$\mathbf{x}_2$	1	0	1	0
$\mathbf{x}_3$	0	1	0	0
$\mathbf{x}_4$	1	1	1	1
$\mathbf{x}_5$	1	1	0	1

Figure 2: Data points.

Create a decision tree for these data. Show the computations made to determine the attribute to split at each node. Please note that you need to use $log2()$ in your computations.

Question 4 (3 points)

A regression analysis relating response value (Y) to variable (X) produced the following fitted question: $\hat{y} = 18 - 0.4x$.

(A) What is the fitted value of the response variable corresponding to $x = 7$? (1 point)

(B) What is the residual corresponding to the data point with $x = 3$ and $y = 20$? (2 points)

Submission

In order to submit your assignment, please upload it as a PDF file to Canvas.