陽明交大 DME Machine Learning Principles – Midterm 2022.11.11 09:00–10:30 (90 min.)

Major/Year______________ Student ID# _________________ Name _______________________

Q1. 監督學習數據和無監督學習數據之間最明顯的區別是什麼？ (5%)

What is the most obvious distinction between the “Supervised Learning” and “Unsupervised Learning” in their respectively collected “data”?

Q2. “*把 webpage 作為一個輸入, M.L. 回答 this webpage 是什麼語言*” 屬於哪種學習? (5%)

□ (1) 監督式學習 (Supervised) (2) 無監督式學習 (Unsupervised) (3) 都可能 (4) 都不是

“*With a webpage as an input, M.L. returns what language used in the webpage*” is in the category of (1) Supervised Learning (2) Unsupervised Learning (3) Both possible (4) Neither

Q3. 我們將收集到的數據分為訓練數據集和測試數據集。

We divide the collected data into training datasets and test datasets.

(a) 請解釋 “overfitting” 的含義 (Please explain the meaning of *overfitting*) (5%)

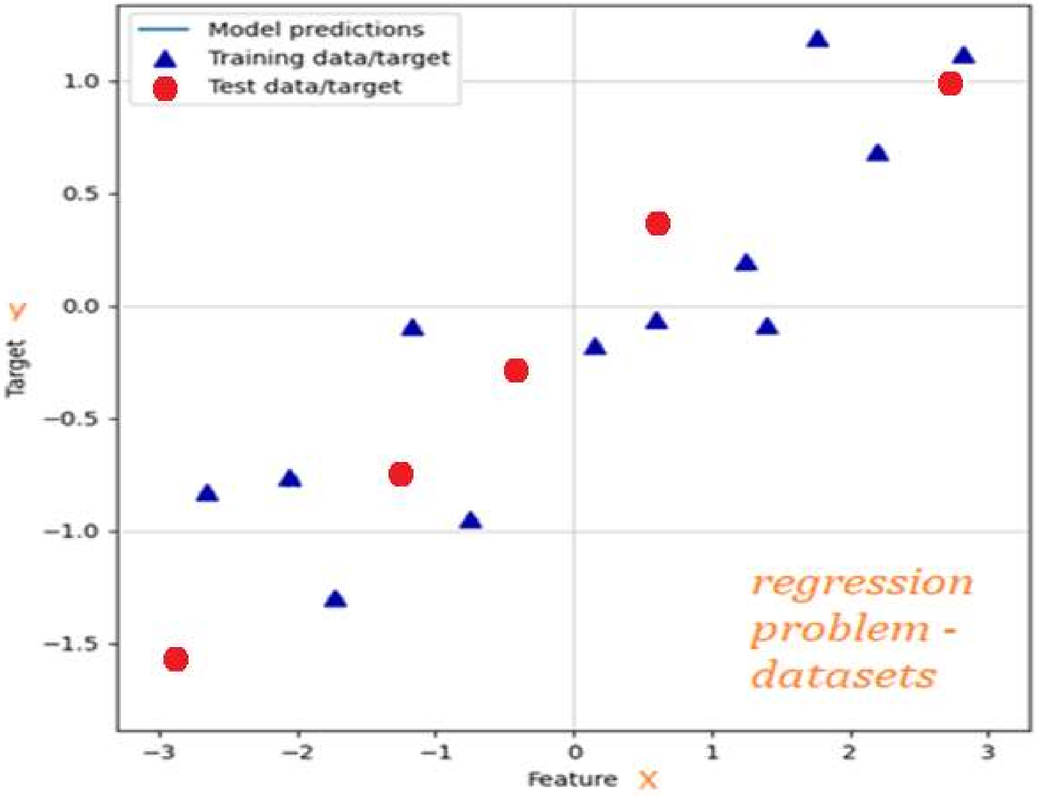
(b) 請解釋 “underfitting” 的含義 (Please explain the meaning of *underfitting*) (5%)

(c) 為什麼 “overfitting”不好 (why is *overfitting* not desired) (5%)

Q4. 為什麼我們在機器學習模型訓練中進行“交叉驗證”？(5%)

Why do we conduct the “*cross validation*” in model training at Machine Learning?

Q5. Given training data (upward blue triangle) below, draw the ***decision boundary line*** (i.e. model prediction line) for this *regression* model if we use KNN method with only 1 nearest neighbor. 給定下面的訓練數據（向上藍色三角形）, 如果我們用 KNN = 1 方法，畫出回歸模型的*決策邊界線*（即模型預測線）。 (5%)

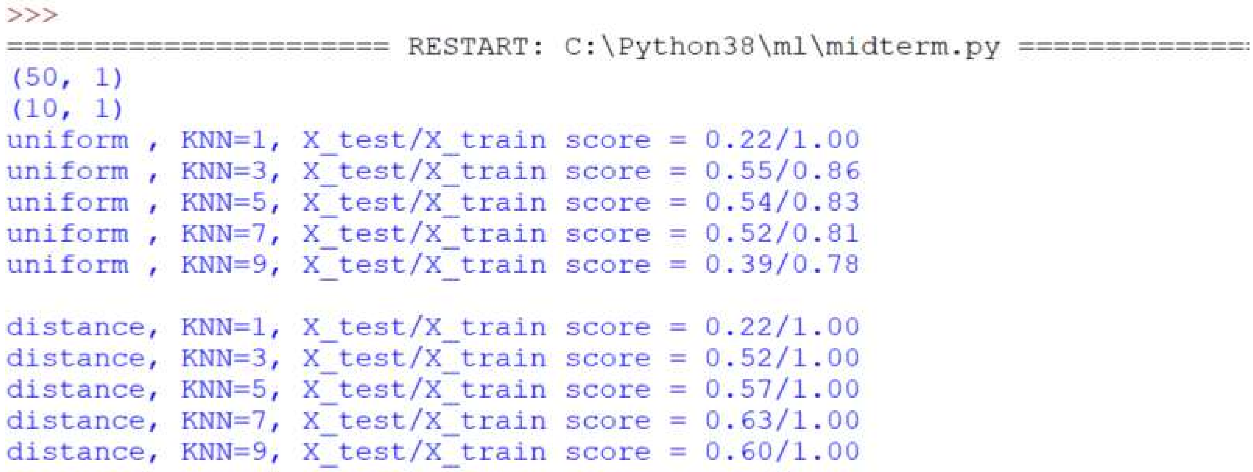


Q6. Using KNN (K-Nearest Neighbors) algorithm with **60** pairs of data for a regression model, you are to write a Python program to read the data from the E3 file “*wave60_dataset.txt*”, **randomly** split them into the test datasets (**10** datasets) and training datasets (**50** datasets), and calculate the test dataset and training dataset scores with the following KNN parameters.

使用 KNN (K-Nearest Neighbors) 算法和 60 組數據作回歸模型，您將寫一個 Python 程序從“wave60_dataset.txt” 中讀取數據，**隨機**將它們分為測試數據集（10 組）和訓練數據集（50 組），並使用以下 KNN 參數計算測試數據集和訓練數據集的分數。 (65%)

n_neighbors = 1, 3, 5, 7, 9
weights = ‘uniform’ ‘distance’

In the *Python Shell Window*, print array shape, the test and training dataset scores in the way as shown below. 在 *Python Shell Window* 中，以如下的方式印出測試數據集和訓練數據集的 shape, 分數。



```
>>>
====================== RESTART: C:\Python38\ml\midterm.py =============
(50, 1)
(10, 1)
uniform , KNN=1, X_test/X_train score = 0.22/1.00
uniform , KNN=3, X_test/X_train score = 0.55/0.86
uniform , KNN=5, X_test/X_train score = 0.54/0.83
uniform , KNN=7, X_test/X_train score = 0.52/0.81
uniform , KNN=9, X_test/X_train score = 0.39/0.78

distance, KNN=1, X_test/X_train score = 0.22/1.00
distance, KNN=3, X_test/X_train score = 0.52/1.00
distance, KNN=5, X_test/X_train score = 0.57/1.00
distance, KNN=7, X_test/X_train score = 0.63/1.00
distance, KNN=9, X_test/X_train score = 0.60/1.00
```

Homework Python programs or sample code in E3 supposedly read similar text data from text data files. Just a reminder – for the regression model from the KNN module, it is “KNeighborsRegressor”.
家庭作業中的 Python 程序或 E3 中的示範例子代碼, 應有類似的文本數據讀取。
提醒一下: 對於回歸模型，KNN 模塊使用的是 “KNeighborsRegressor” 。

(65% - *reading 10%, using loop (for or while) 15 %, kNN 15%, weights 15%, printing format 10%*)