Introduction to Artificial Intelligence - Homework 4

NE6114011 人工智慧所碩一 楊雲翔

1 K-nearest-neighbors linear regression

實現方法

存在訓練資料(X, Y)，其中 X=$(X_1, X_2, \dots, X_n)$，Y$(Y_1, Y_2, \dots, Y_n)$，$X_1 = (x_1, x_2, \dots x_p)$，n 表資料數，p 表資料的維度，以 data2.npz 為例，n=1000, p=2。此時，若要預測一筆新樣本$X' = (x_1, x_2, \dots, x_p)$的值$y'$，步驟如下：

1. 與訓練資料 X 中的每一筆資料都計算一次歐幾里得距離，並尋找距離最近的 K 個樣本
2. 使用距離最近的 K 個樣本進行線性回歸，求解$W = (X^T X)^{-1} X^T y$ (這裡的 X 表 K 個樣本的 x，Y 表 K 個樣本的 y)，獲得線性方程式的參數$W = (w_0, w_1, \dots w_p)$
3. 將樣本X'代入，獲得$y' = W[1\ X'] = w_0 + w_1 x'_1 + w_2 x'_2 + \cdots w_p x'_p$

實驗結果

將訓練資料(X, Y)以 9:1 拆分為訓練集與測試集，並且以 KNN linear regression 進行測試集的預測，並計算 MSE loss。

- 圖 1: 設置不同 K(neighbors 數)時的 MSE loss 變化，
- 圖 2: 根據訓練資料(X, y)，依照 X 的上下界，在中間均勻取點，產生多筆新資料進行 KNN linear regression(K = 20)預測的結果，藍點為原始資料，紅點為新資料

1. Data1.npz:

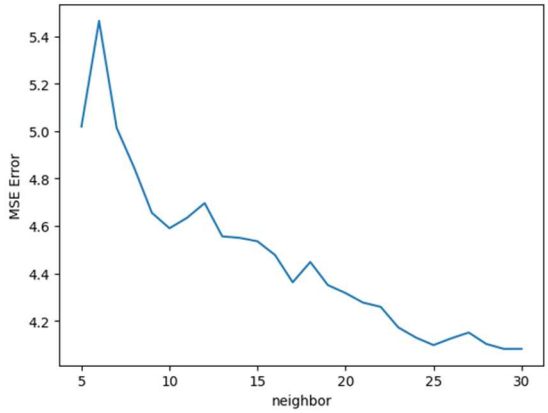


圖1

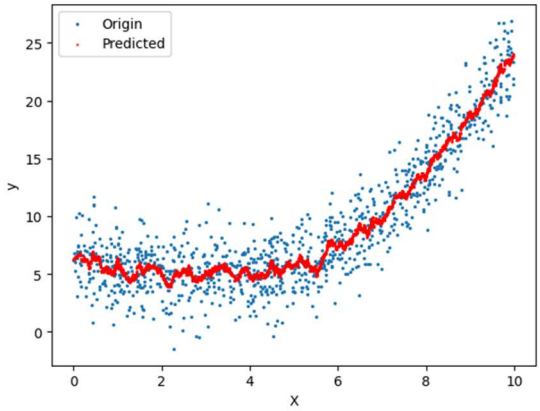


圖2

2. Data2.npz:

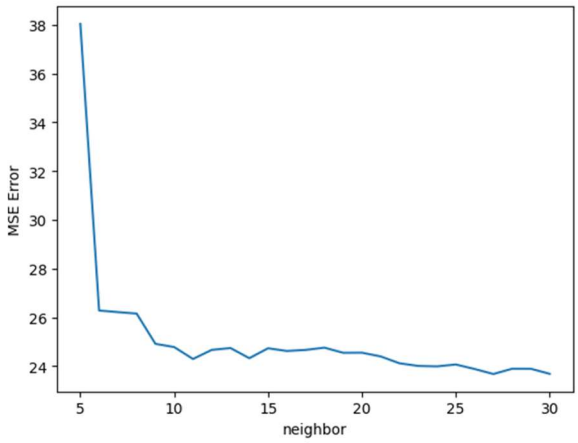


圖1

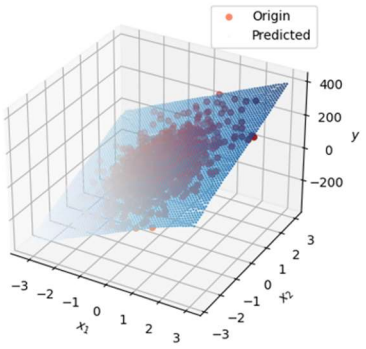


圖2

2 Locally weighted regression

實現方法

存在訓練資料(X, Y)，其中 X=$(X_1, X_2, \dots, X_n)$，Y$(Y_1, Y_2, \dots, Y_n)$，$X_1 = (x_1, x_2, \dots x_p)$，n 表資料數，p 表資料的維度，以 data2.npz 為例，n=1000, p=2。此時，若要預測一筆新樣本$X' = (x_1, x_2, \dots, x_p)$的值$y'$，步驟如下：

1. 與訓練資料 X 中的每一筆資料都計算一次歐幾里得距離，將計算出的距離代入 quadratic kernel，獲得 weight $w = \max(0, 1 - (2|dist|/k)^2)$，k 為 kernel width
2. 使用該 weight w 代入方程式$\theta = (X^T W X)^{-1} X^T W y$(相當於解最佳化式子$w^* = argmin_w \sum_j w_j (y_j - w \cdot x_j)^2$)，並求解
3. 計算$(X^T W X)^{-1} X^T W y$獲得一條線性方程式的權重θ，$\theta = (\theta_0, \theta_1, \dots \theta_p)$
4. 將樣本X'代入，獲得$y' = \theta[1\ X'] = \theta_0 + \theta_1 x'_1 + \theta_2 x'_2 + \cdots \theta_p x'_p$

實驗結果

將訓練資料(X, Y)以 9:1 拆分為訓練集與測試集，並且以 Locally weighted regression 進行測試集的預測，並計算 MSE loss。

- 圖 1: 設置不同 K(kernel width)時的 MSE loss 變化，
- 圖 2: 根據訓練資料(X, y)，依照 X 的上下界，在中間均勻取點，產生多筆新資料進行 Locally weighted regression (K = 10)預測的結果，藍點為原始資料，紅點為新資料

1. Data1.npz:

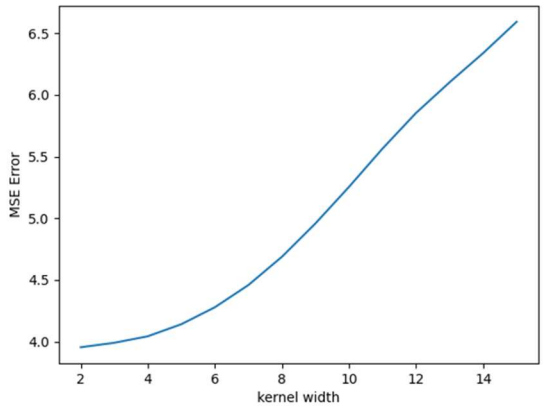


圖1

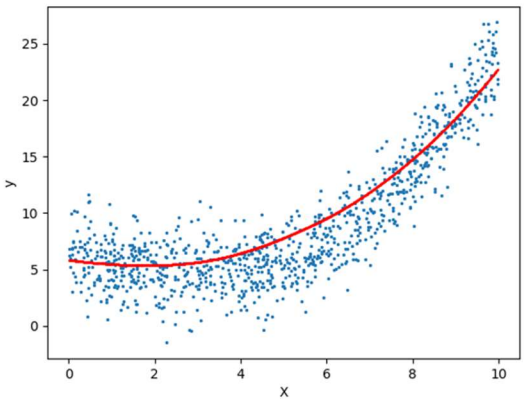


圖2

2. Data2.npz:

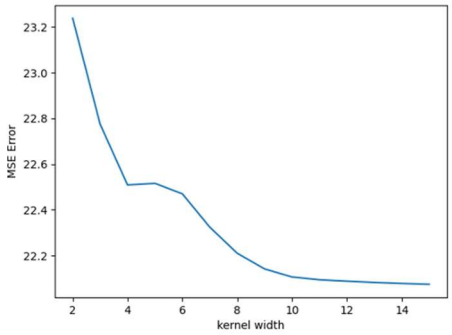


圖1

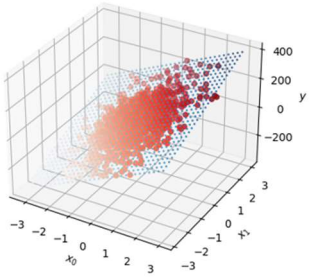


圖2

3 Other Method: K-nearest-neighbors regression

實現方法

存在訓練資料(X, Y)，其中 X=$(X_1, X_2, \dots, X_n)$，Y$(Y_1, Y_2, \dots, Y_n)$，$X_1 = (x_1, x_2, \dots x_p)$，n 表資料數，p 表資料的維度，以 data2.npz 為例，n=1000, p=2。此時，若要預測一筆新樣本$X' = (x_1, x_2, \dots, x_p)$的值$y'$，步驟如下：

1. 與訓練資料 X 中的每一筆資料都計算一次歐幾里得距離，並尋找距離最近的 K 個樣本
2. 使用距離最近的 K 個樣本之 y 計算平均 $1/k \sum_j K_j$(這裡的 K 只含距離最近的 K 個樣本)
3. 此平均值即為樣本X'對應的y'

實驗結果

將訓練資料(X, Y)以 9:1 拆分為訓練集與測試集，並且以 KNN regression 進行測試集的預測，並計算 MSE loss。

- 圖 1: 設置不同 K(neighbors 數)時的 MSE loss 變化，
- 圖 2: 根據訓練資料(X, y)，依照 X 的上下界，在中間均勻取點，產生多筆新資料進行 KNN regression(K = 20)預測的結果，藍點為原始資料，紅點為新資料

1. Data1.npz:

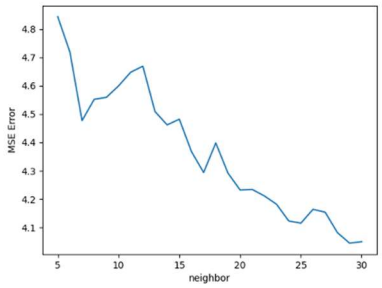


圖1

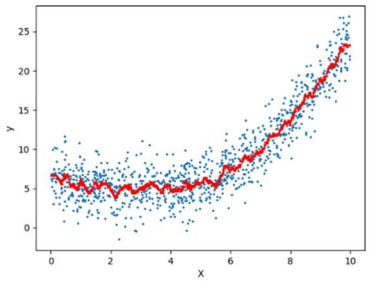


圖2

2. Data2.npz:

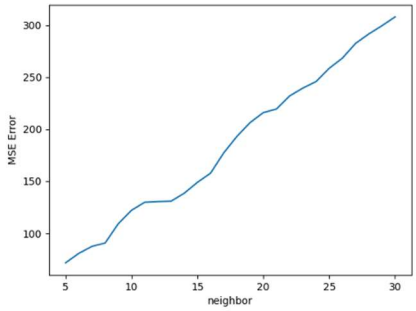


圖1

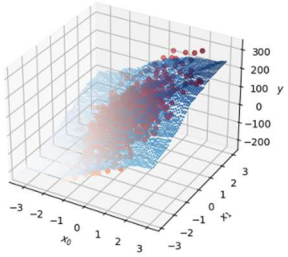
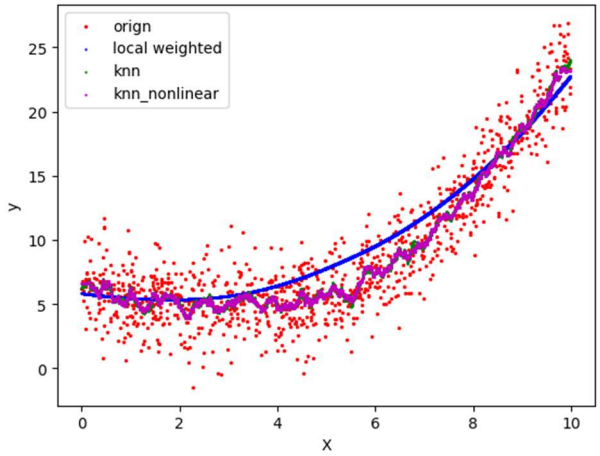


圖2

4 比較

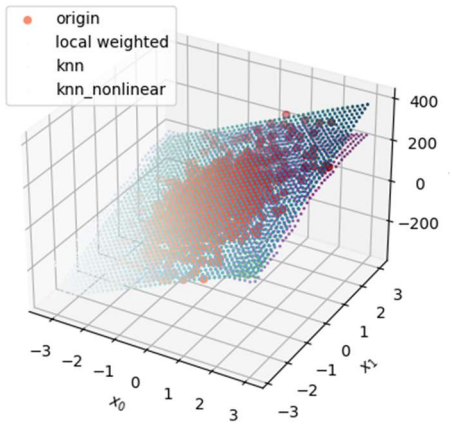
- Data1.npz:

  下圖是將結果放在同一圖上的結果，可以很明顯地看到 KNN k=20(綠點)與 KNN nonlinear(紫點)的曲線大致相同，且同為不連續的曲線，而 local weighted(藍點)則是連續的曲線



- Data2.npz:

  下圖是將結果放在同一圖上的結果，在二維的例子中，可以很明顯地發現 KNN(綠點)與 KNN nonlinear(紫點)的平面不太一致，反而 local weighted(藍點)與 KNN 綠點)的平面較為接近，這個實驗結果也跟上面 loss 的結果大致相同，經由觀察上面的 MSE loss 中，可以發現 KNN regression 在二維的例子(data2.npz)中 MSE Loss 很明顯地較 KNN linear regression 與 local weighted regression 差上許多


origin
local weighted
knn
knn_nonlinear
400
200
0
−200
−3
−2
−1
0
1
2
3
x_0
x_1