Homewort4 Report

資訊系F74084070何勁廷

K-nearest-neighbors linear regression ：

實作內容：

1. **fit:**

   這裡將1000筆資料分成700訓練集、300測試集。將訓練資料集的x和y輸入我建立的class中。

2. **calculate distance:**

   計算訓練集中每筆資料對應測試集中每筆資料的euclidean distance，並存在一個table當中。

3. **search k's closest points:**

   透過先前建立的table，去尋找距離測試資料最近的k的點的index。

4. **get neighbors's value：**

   透過先前獲得的index來取得該筆資料的x, y值。

5. **linear regression：**

   這裡直接引用上課教的方法去尋找這k個點的linear regression方程式係數。

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

   透過這個方式可以取得一個closed form的解。

6. **get predict value：**

   透過先前取得的方程式係數，直接導入測試資料的x值來取得y值。

實驗內容：

根據不同neighbors觀察RMSE(root mean square error)數值，並且利用三維散點圖來觀察數值分佈以及是否outlier：

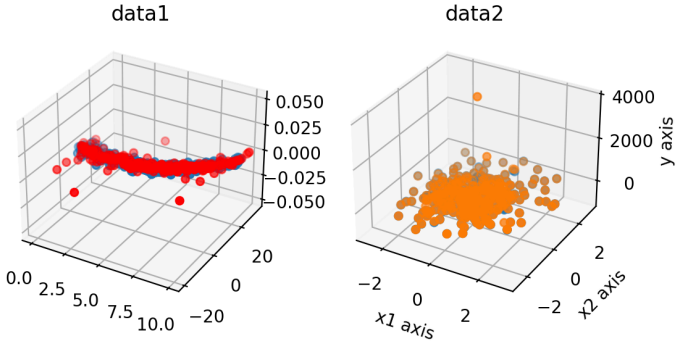
1. n_neighbors = 3

   RMSE for data1 :

   4.001747368190441

   RMSE for data2 :

   247.56412752880073



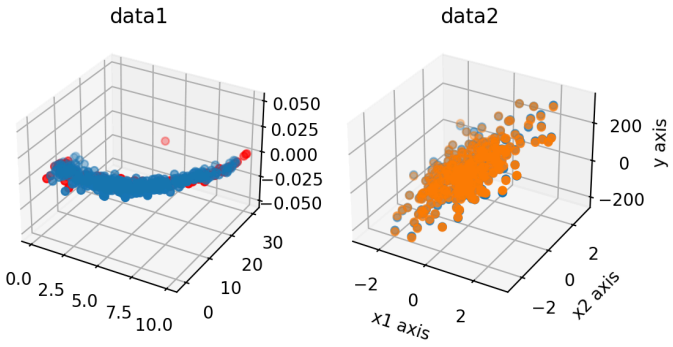
2. n_neighbors = 5

   RMSE for data1 :

   2.7690329226651755

   RMSE for data2 :

   6.234766865329543



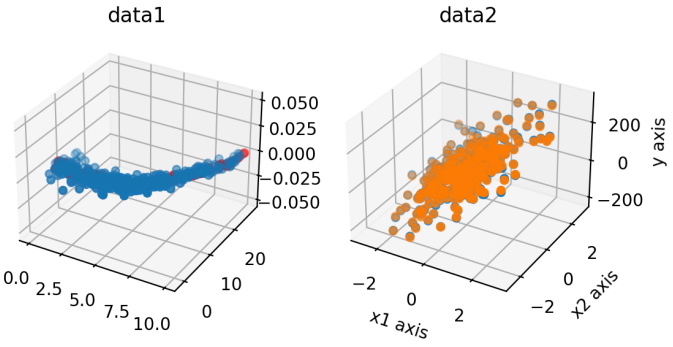
3. n_neighbors = 7:

   RMSE for data1 :

   2.291138079150532

   RMSE for data2 :

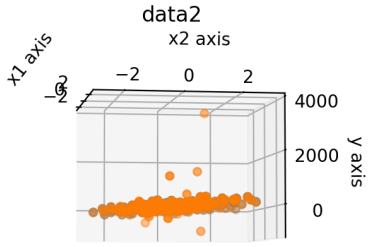
   5.628319217915673



實驗分析：

從上述三種neighbors數量的實驗來看，對data1而言結果最好的是neighbors=7時，RMSE數值大約等於2.3。並且可以發現data1在neighbors越小時RMSE就越大，並且從三維散點圖來看可以發現outlier也越多。

針對data2來看，其趨勢和data1相似，同樣是neighbors越多，RMSE就越小，outlier也越少。特別在neighbors=3時，data2出現了好幾個極大outlier，使得RMSE非常大(如下圖)



左圖為neighbors=3時，data2透過我建立的knn model所產出的資料三維散點圖。

觀察左圖可以發現，雖然RMSE有200多，但其實大部分資料的預測都是在合理範圍內。但有幾個極大outlier，數值將近4000導致RMSE整個失真。

Locally weighted regression :

實作內容：

1. fit:

   這裡將1000筆資料分成700訓練集、300測試集。將訓練資料集的x和y以及gaussian公式中的tau輸入我建立的class中。

2. gaussian:

   透過gaussian distribution來計算兩點之前的weight

   這裡我會透過修改tau參數來做實驗，tau值影響的是權重值得變化速率。

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

(1) 如果 $|x^{(i)} - x| \approx 0$，則 $w^{(i)} \approx 1$。

(2) 如果 $|x^{(i)} - x| \approx +\infty$，則 $w^{(i)} \approx 0$。

3. get weights:

   利用上述的gaussian計算每筆測試資料和每筆訓練資料之間的weights。

4. get predict value:

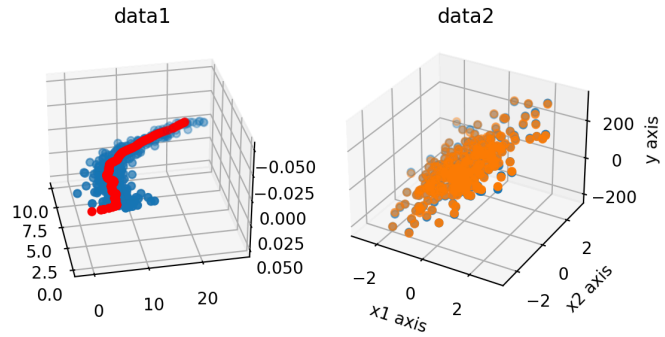
   取得weights後，尋找linear regression係數，使用上課教的closed form方法。

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

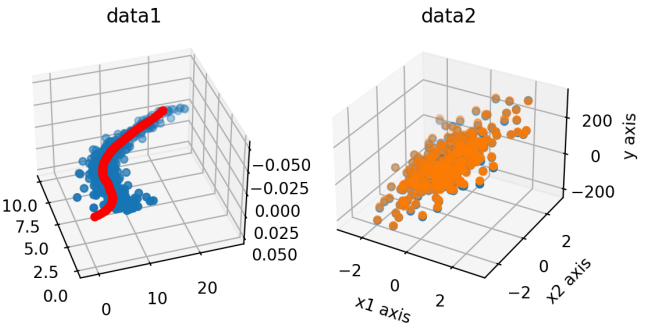
實驗內容：

根據不同tau值來觀察RMSE(root mean square error)，以及三維散點圖。

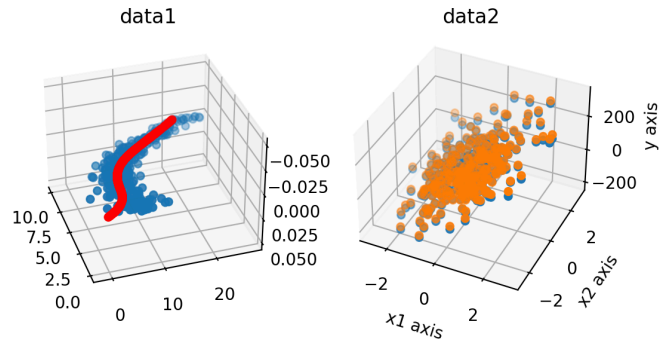
1. tau = 0.2
RMSE for data1:
2.254509216172766
RMSE for data2:
6.656347819414298



1. tau = 0.5
RMSE for data1:
2.4391670514487083
RMSE for data2:
9.955798611056327



2. tau = 1
RMSE for data1:
2.6908451324203644
RMSE for data2:
15.435739108506603



實驗分析：

根據實驗內容來看，tau值越大RMSE就越大，tau越小RMSE就越小，效果越好。這個現象在data2更為明顯。值得一提的是，這個方法得到的結果較不會出現outlier，觀察上圖當tau = 1時，data2的RMSE大約為15，但是資料趨勢大致合理，不像knn會出現非常極端的outlier。

Other method (K-nearest-neighbors nonlinear regression):

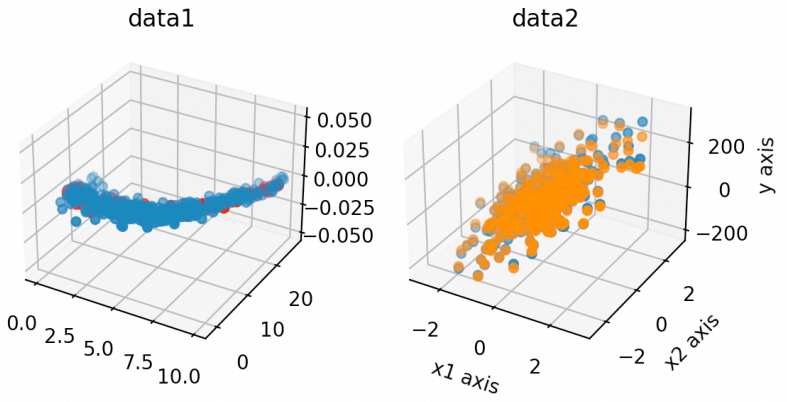
實作內容：

1. **fit:**
   這裡將1000筆資料分成700訓練集、300測試集。將訓練資料集的x和y輸入我建立的class中。
2. **calculate distance:**
   計算訓練集中每筆資料對應測試集中每筆資料的euclidean distance，並存在一個table當中。
3. **search k's closest points:**
   透過先前建立的table，去尋找距離測試資料最近的k的點的index。
4. **get neighbors's value：**
   透過先前獲得的index來取得該筆資料的x, y值。
5. **get predict value:**
   直接將取得的neighbors的y值取平均，即為predict value

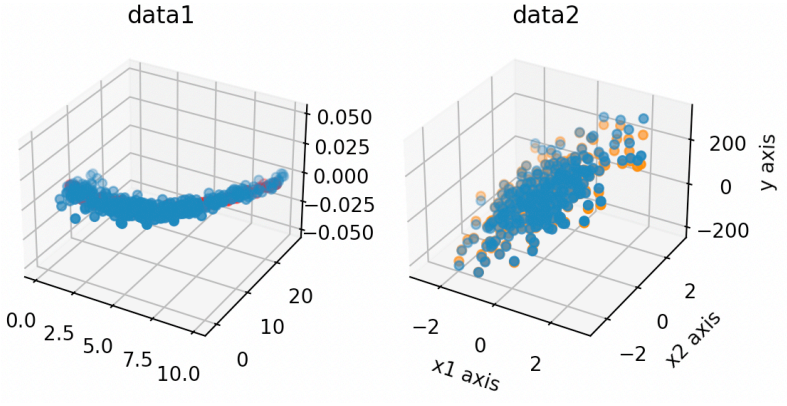
實驗內容：

根據不同neighbors觀察RMSE(root mean square error)數值，並且利用三維散點圖來觀察數值分佈以及是否outlier：

n_neighbors = 3
RMSE for data1 :
2.441147924594815
RMSE for data2 :
9.773157149100998



n_neighbors = 5
RMSE for data1 :
2.245816160111502
RMSE for data2 :
10.243182837413109



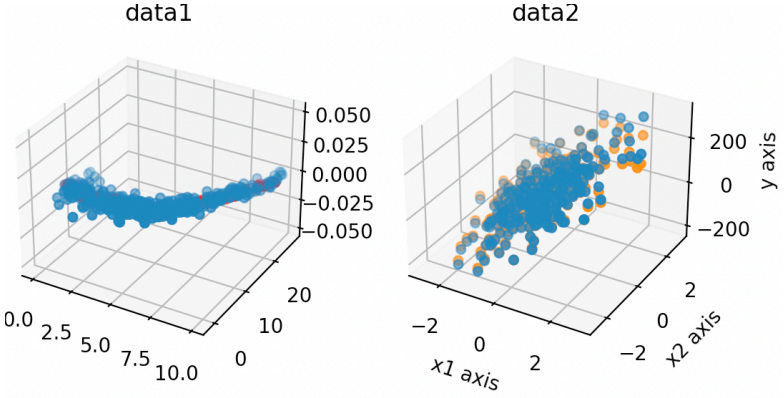
n_neighbors = 7:

RMSE for data1 :

2.233236783853117

RMSE for data2 :

11.019391335400584



實驗分析：

觀察實驗內容可以發現，neighbors的數量變化對RMSE的影響不大。並且針對data1，neighbors越多，RMSE越小。針對data2，neighbors越多，RMSE越大。觀察三維散點圖，資料趨勢平緩，沒有outlier出現。

總結：

從上述三個model的實驗數據觀察可以分成以下幾點結論。

一、Knn linear regression出現outlier的原因：

當knn linear regerssion在neighbors數量小時，會出現極端outlier，其原因主要是因為取的鄰居數量太少，導致在做linear regression時得到的線性方程式失真。這個問題在neighbors數量增加後就解決了，因此可以應證這個推論。

二、Locally weighted regression與Knn linear regression的關係：

LWR在tau值小時的實驗結果和KNN linear regression在neighbors數量大時的實驗結果相似。LWR的做法是針對所有點找權重，越近的點權重越高，越遠的點權重越小，廣義來看，其概念和knn linear regression相似。因此得到的實驗結果也相似。

三、Knn linear regression和Knn nonlinear regression比較：

觀察實驗結果可以發現，knn nonlinear regression在neighbors數小時，效果比knn linear regression。其原因和第一點講的一樣，是因為鄰居數太小以至於線性方程式失真。然而，一但鄰居數增加，knn nonlinear regression的表現及完全比不上knn linear regression。