Nonparametric regression

WEI-CHIH CHANG Q36111281
November 23, 2022

Abstract

調整 KNN-f, KNN-means, lw 三種演算法的各種參數，觀察其對一維資料和二維資料回歸的影響。

1. Related Work

這次報告我希望模仿 sklearn 以及 tensorflow 等模組化的函式庫，實作出以下三種 Nonparametric regression 的演算法，分別是：K-nearest-Neighbors frequency Linear Regression, K-nearest-Neighbors means Linear Regression 以及 Locally Weighted Regression。

程式的部分，我將常用的數學算法(minkowski distance 和 weight matrix)包裝成函式在 base.py 內。兩種不同的 KNN 被封裝成相同 class 類別，以 mod 參數區別兩種演算法模式，mod='f'對應 KNN-f 演算法，mod='m'對應 KNN-means 演算法，最後將其一同放在 KNN.py 中。參數 p 表示 minkowski distance 的計算階數，參數 k 表示所選取的鄰居數量。fit()函式和 predict()函式分別對模型的訓練以及模型對測試資料的預測。

Locally Weighted Regression 的程式規劃和 KNN.py 類似，一樣將常用數學函式 weight matrix 放在 base.py 中，一樣具有 fit()函式和 predict()函式並且功能相同，其中參數 tau 代表高斯鐘形曲線的寬度，最後將該類別封裝在 lw.py 中。

2. K-nearest-Neighbors Linear Regression

KNN 演算法大致分為三個部分，分別是距離計算、排序以及鄰居的選取。因為有排序的需求在，因此選擇 python 特有的 dictionary 變數型態做為主要的資料類別，以加速資料之間的查找以及排序。而 KNN-f 和 KNN-means 兩模式之間的不同在於，前者是取 k 個鄰居中出現頻率最高者，後者則是取 k 個鄰居值的平均。

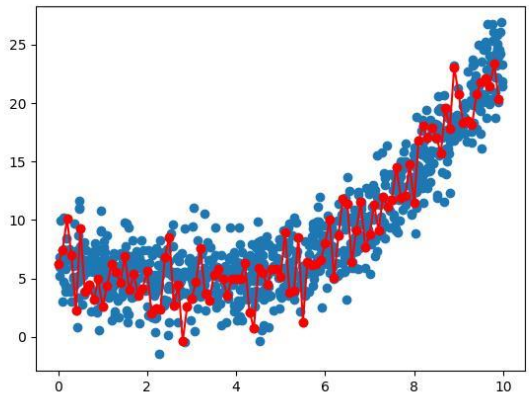
2.1 The influence of different order of Minkowski distance on the regression of 2-D data

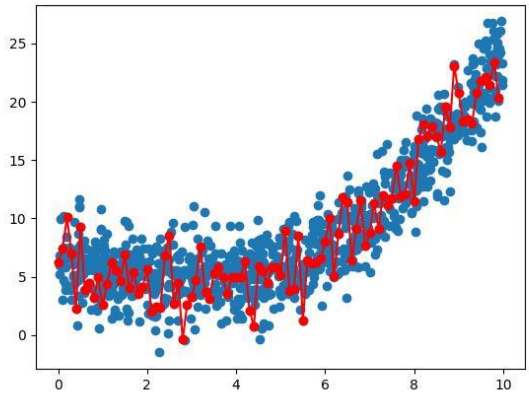
Minkowski distance 可以被解釋為歐氏距離和曼哈頓距離的推廣，當階數 p=2 時為歐氏距離，當階數 p=2 時為曼哈頓距離，當 p 取無窮時的極限情況下，可以得到切比雪夫距離

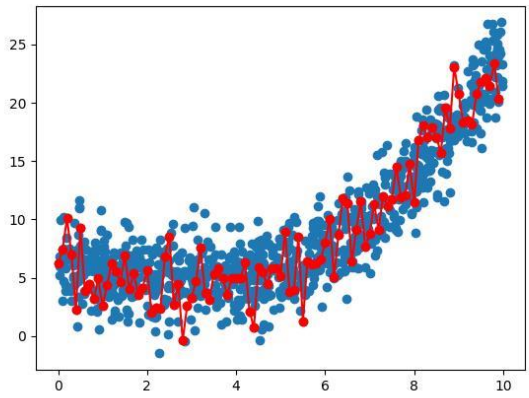
$$\left(\sum_{i=1}^{n}|x_i-y_i|^p\right)^{1/p}.$$

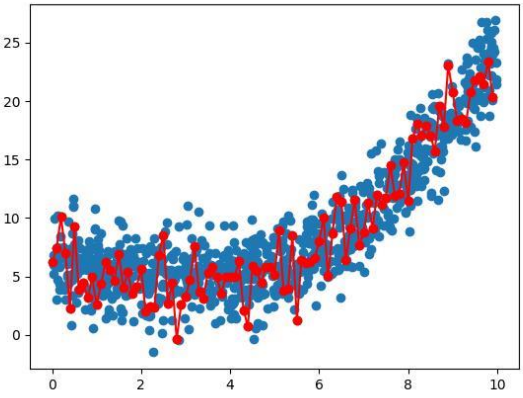
分別取 p=1, 2, 10, 100 比較不同階數的 Minkowski distance 對 K-nearest-Neighbors Linear Regression 的影響。其中 k=5，取樣點數為 100，範圍從 0 到 10。

KNN-f

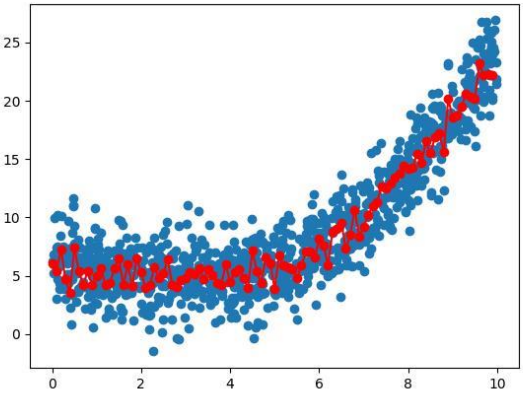


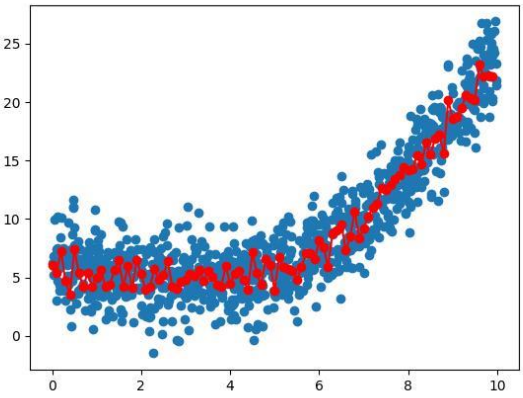


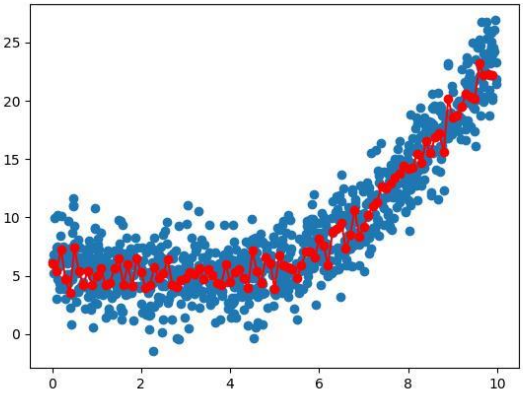


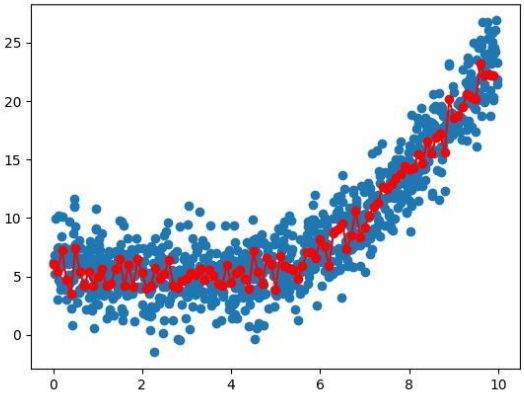


KNN-means







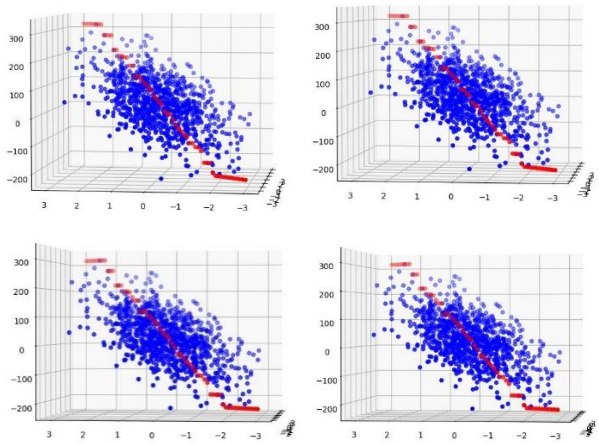


從 KNN-f 和 KNN-means 的八張圖中可以發現 Minkowski distance 的階數對一維資料的 K-nearest-Neighbors Linear Regression 並沒有影響。

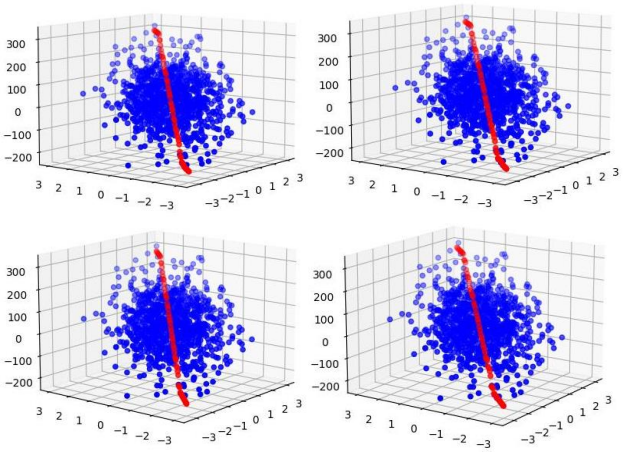
2.2 The influence of different order of Minkowski distance on the regression of 3-D data

分別取 p=1, 2, 10, 100 比較不同階數的 Minkowski distance 對 K-nearest-Neighbors frequency Linear Regression 的影響。其中 k=5，取樣點數為 60，參數範圍維從-3 到 3 的(x1, x2)。

KNN-f



KNN-means

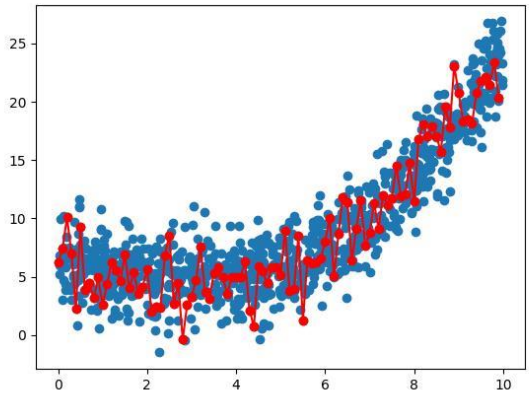


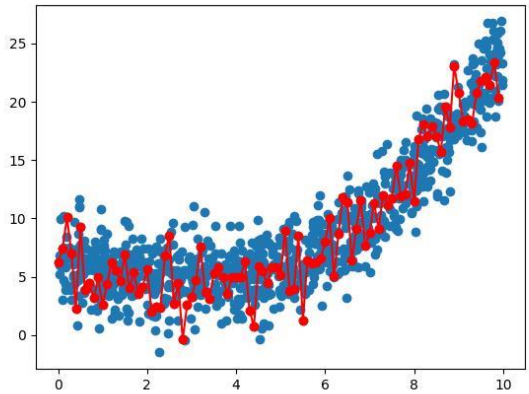
從四張圖可以發現 Minkowski distanc 的階數對二維資料的 K-nearest-Neighbors Linear Regression 並沒有影響。

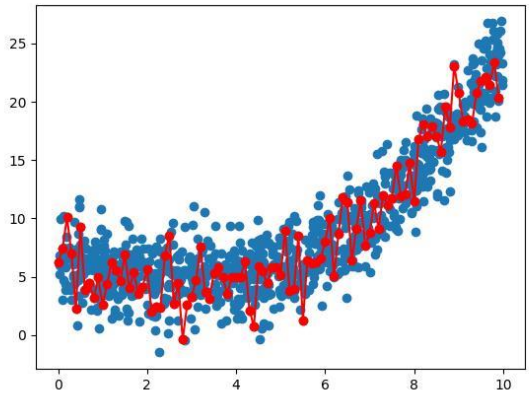
2.3 The influence of different number of nearest-neighbors k on the regression of 2-D data

參數 k 為 KNN 參考的鄰居數，如果 k 的值越小表示參考的點越少，同理 k 的值越大表示參考的點越多。分別取 k=2, 5, 10, 20 比較不同 k 值對 K-nearest-Neighbors Linear Regression 的影響。其中 p=2，取樣點數為 100，範圍從 0 到 10。

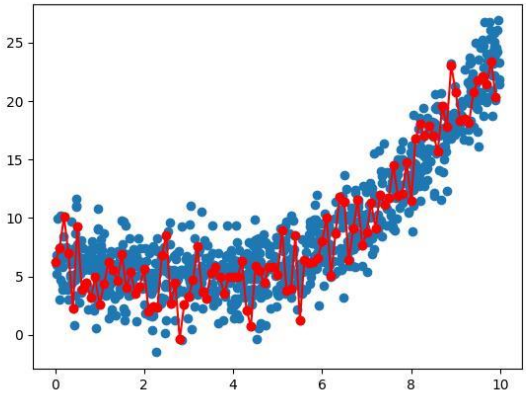
KNN-f

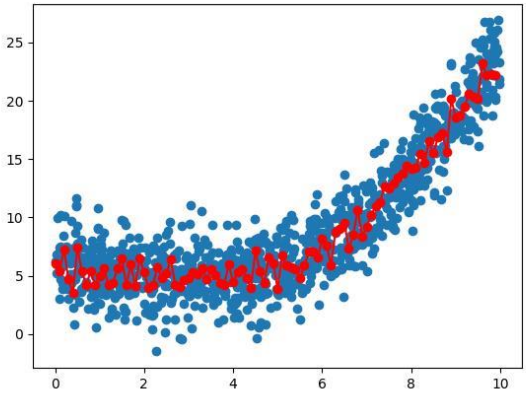


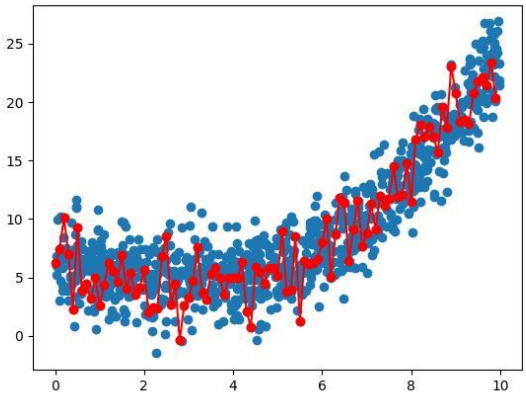


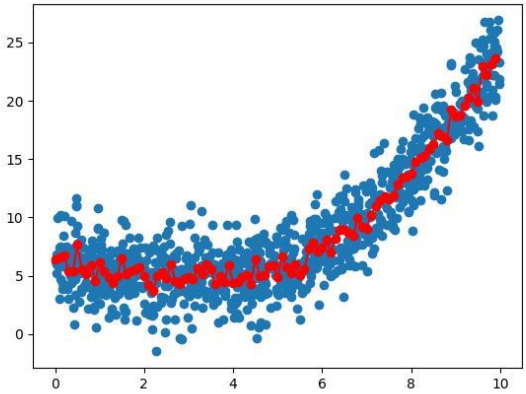


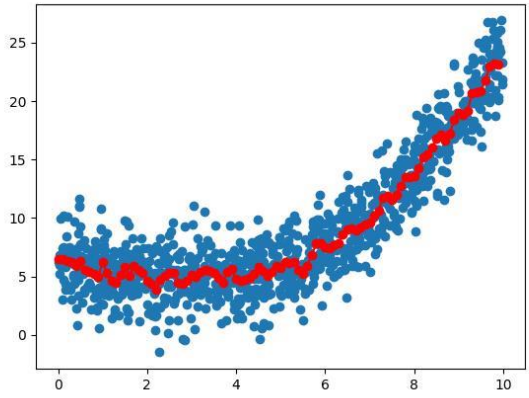
KNN-means

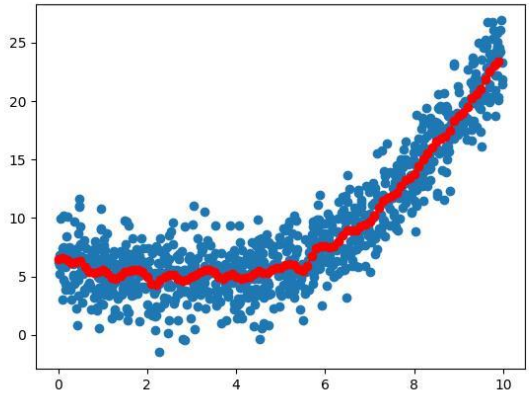




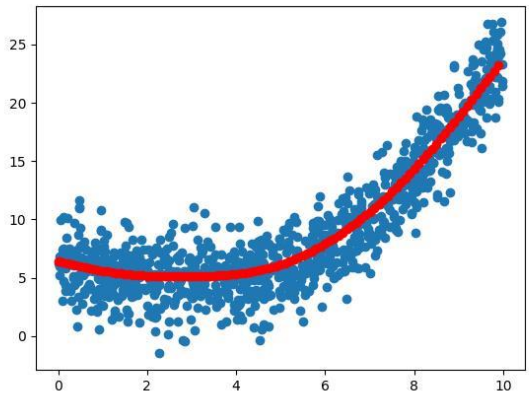








從圖中可以發現鄰居數 k 對 KNN-f 並沒有影響，因為 KNN-f 是考慮點出現的頻率，因此在每一個點幾乎沒有重複的情況下 KNN-f 就退化到 k=1 的 KNN-f。但考慮 KNN-means 可以發現 k=5 時回歸線最不平滑，可以推論是若考慮 5 個鄰居時資料不夠具有代表性，k=2 則是因為都是取該點的最近點因此回歸線較為平滑，k=10 和 k=20 因為取樣點足夠、足夠具有代表性，因此回歸線較為平滑。

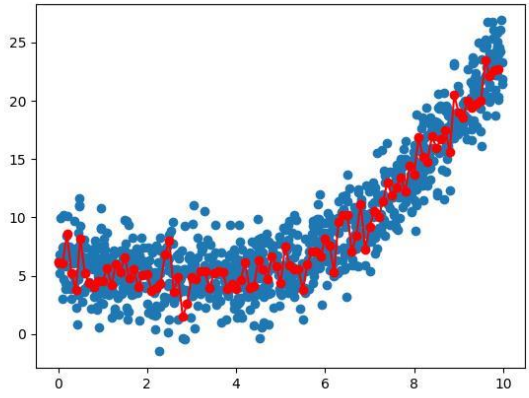


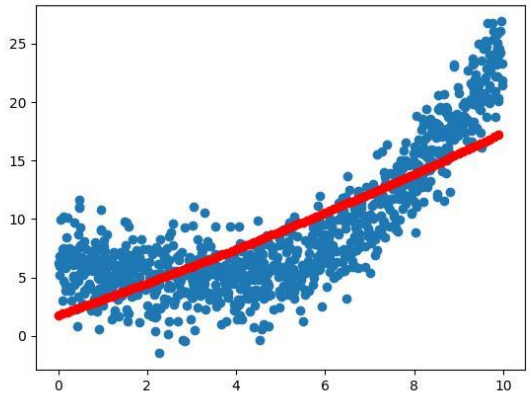
3. Locally Weighted Regression

因為 lw 在計算時權重矩陣時會將特徵視為向量來處理，有大量的矩陣運算，因此選擇 np 的資料型態來處理資料。

3.1 The influence of different bandwidths on the regression of 2-D data

Locally Weighted Regression 的參數 tau 代表演算法中高斯分布的寬度，其中取 tau=0.01, 0.1, 1, 10 來分析高斯分布寬帶對回歸的影響。其中取樣點數為 100，範圍從 0 到 10。

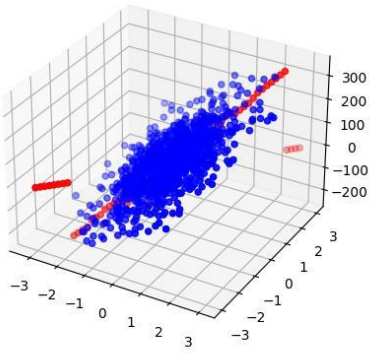


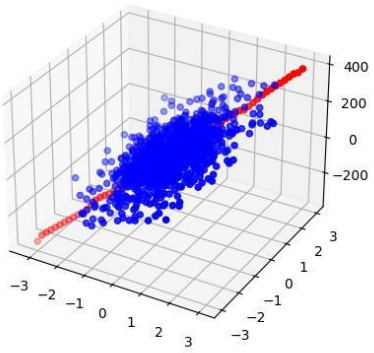


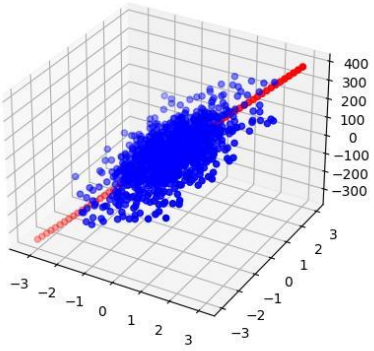
從以上幾張圖可以發現，tau 越小對資料就越敏感，tau 越大則對資料的分布越不敏感，因此甚至當 tau=10 時，回歸線甚至退化成近乎直線。而從四張圖可以觀察出，tau=1 具有最好的回歸效果，而 tau=0.1 則對邊界具有適當的敏感度。

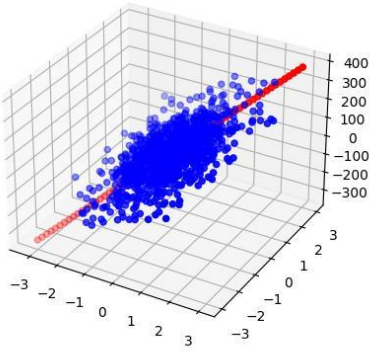
3.2 The influence of different bandwidths on the regression of 3-D data

從以下四張圖可以發現，tau=0.01 對資料仍具有敏感度，但方向並不是好的，而當 tau=0.1, 1, 10 時，對二維資料的回歸幾乎相同。









4. Conclusion

在 KNN-f 時，因為是考慮鄰居集中鄰居出現的頻率，因此在資料重疊機率非常低的情況下，KNN-f 退化成 k=1 的 KNN-f。而 KNN-means 則會因為 k 的不同有不同的回歸效果。其中使用的距離 Minkowski distance 在任意 p 時對一維和二維基料的距離計算幾乎相同。

Locally Weighted Regression 的參數 tau 越小對資料就越敏感，tau 越大則對資料的分布越不敏感，當 tau=10 時，回歸線甚至退化成近乎直線。

使用函式庫有：matplotlib, numpy, math, decimal, collections