人工智慧導論

Homework 4

系所： 資訊所
姓名： 張少鈞
學號： P76114545

k-nearest-neighbors linear regression

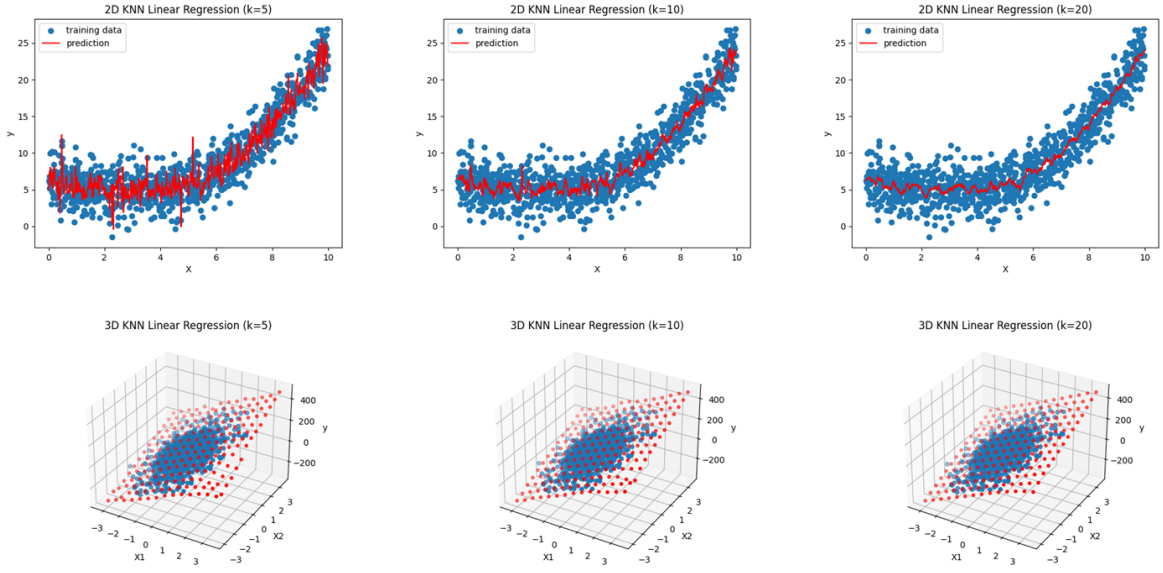
Method

在進行 n 維的線性回歸問題，k-nearest-neighbors linear regression 利用 KNN 的概念從訓練資料(x_t)中利用歐式距離(式 1)計算離查詢點(x_q)最接近的 k 個點。取出 k 個點(x_k)後利用最小平方法進行線性迴歸分析，最後回推查詢點所對應的輸出數值。

$$distance = \sqrt{\sum_{x_i \in x_k} (x_q - x_i)^2} \qquad \text{式 1}$$

Result

k-nearest-neighbors linear regression 的結果如圖表 1。觀察結果可以發現，隨著 k 值的增加，KNN linear regression 預測出來的數值越穩定，較不容易產生震盪，主要原因是 k 值的增加代表參考點增加，預測比較不會受到雜訊的干擾。



(a) (b) (c)

圖表 1 KNN linear regression result。(a) k=5 (b) k=10 (c) k=20

Locally weighted regression

Method

Locally weighted regression(LWR)透過給予不同的權重來改進傳統線性回歸模型欠擬合與過度擬合的問題，甚至能使用 LWR 來擬合非線性問題。

LWR 的權重會依據查詢點而不同，越靠近查詢點的訓練資料權重會越高，反

之則越低。第i筆訓練資料的權重w_i計算方式如(式 2)。式中的$(x_q - x_i)^2$代表查詢點與各點的距離，而$-2k^2$則代表權重隨距離下降的速度，最後透過exp的計算將w_i限定在 0~1 之間。

$$w_i = \exp\left(\frac{(x_q - x_i)^2}{-2k^2}\right) \qquad \text{式 2}$$

在 LWR 的假設中其 Loss 函式定義如式 3，透過計算出各筆訓練資料的權重後，解出能使 Loss 降到最低的θ值，最後用此θ值來計算出查詢點的對應數值(式 4)。

$$J(\theta) = \sum_{x_i \in x_q} w_i (y_i - \theta^T x_i)^2 \qquad \text{式 3}$$

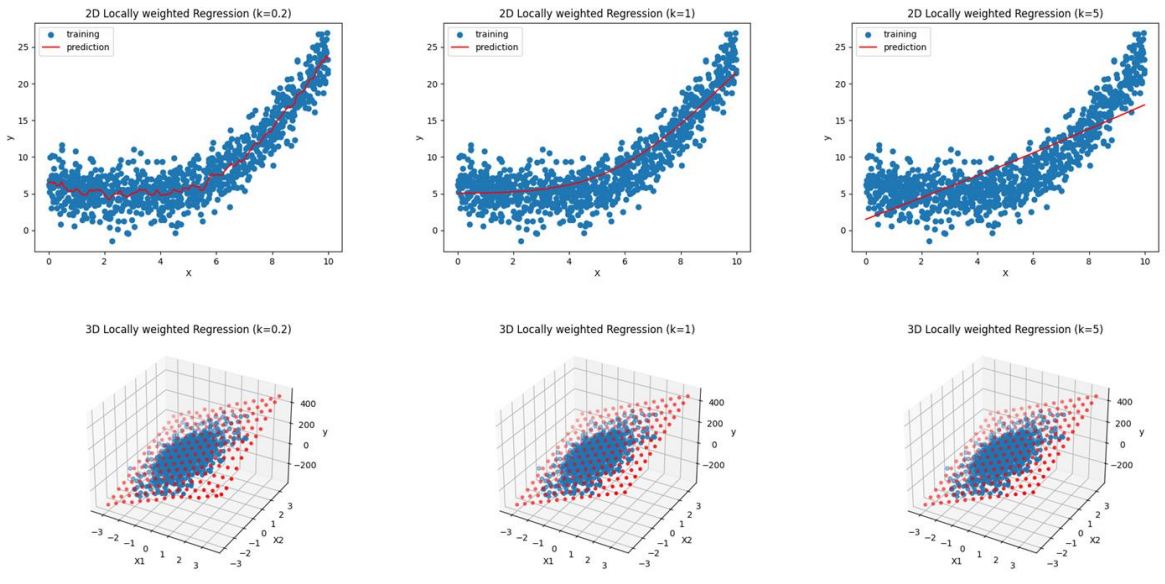
$$y_q = \theta^T x_q \qquad \text{式 4}$$

在求解θ的問題上直接使用矩陣計算來進行，將$J(\theta)$假設為 0 進行計算，最後將方程式改寫為式 5，即可計算出相對應的θ值。

$$\theta = (X^T W X)^{-1} X^T W Y \qquad \text{式 5}$$

Result

透過觀察 LWR 的結果(圖表 2)發現，在 k＝5 時 LWR 的預測結果與線性回歸的結果相近，從 LWR 的權重公式來看，當 k 足夠大時每筆訓練資料的權重都會趨近於 1，也就會變成傳統的線性迴歸分析問題。因此隨著 k 值的縮小，距離查詢點越遠的訓練點的權重會隨之變小，使預測更接近非線性的迴歸分析，因此震盪也隨之增加。



(a) (b) (c)

圖表 2 Locally weighted regression result。(a) k=0.2 (b) k=1 (c) k=5

k-nearest-neighbors regression

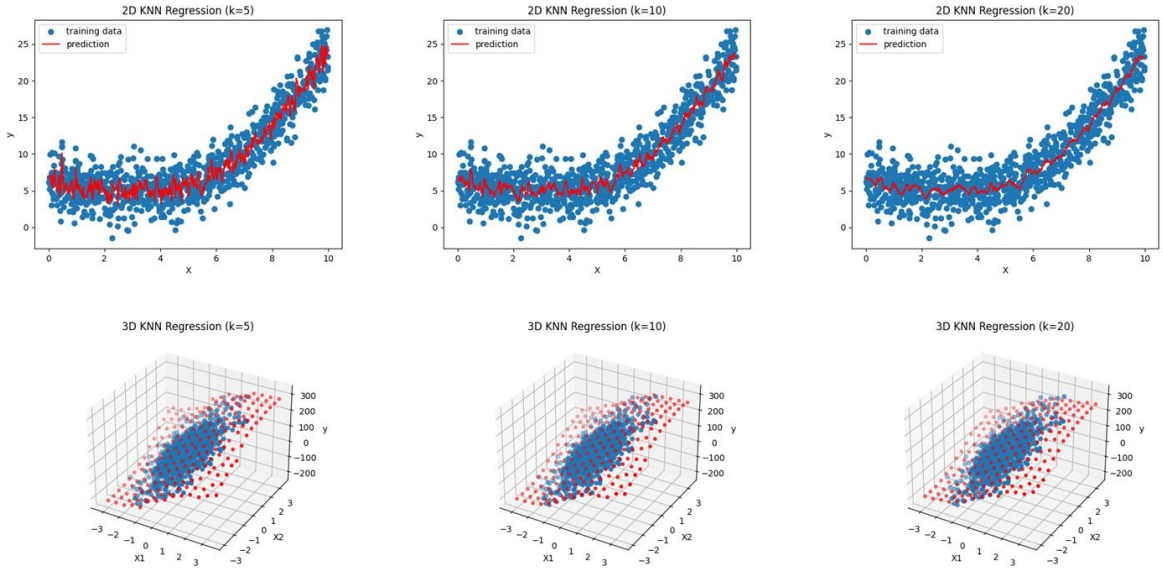
Method

k-nearest-neighbors regression 是 k-nearest-neighbors linear regression 的前身，兩者之間的差別在於，從訓練資料取出 k 個點後，k-nearest-neighbors linear regression 是使用這 k 筆資料進行線性迴歸分析來求查詢點的對應數值；反之，k-nearest-neighbors regression 是將這 k 個點的對應值(y_k)取平均後當成查詢點的對應數值(如式 6)。

$$y_q = \sum_{y_i \in y_k} {y_i}/{k} \qquad \text{式 6}$$

Result

k-nearest-neighbors regression 為 k-nearest-neighbors linear regression 的前身，因此將兩者結果進行比較。透過觀察 k-nearest-neighbors regression 的結果可以發現，在相同的 k 值下 k-nearest-neighbors regression 的預測震盪結果會較小，因為其採用平均的方法來進行預測，而 k-nearest-neighbors linear regression 因為採用最小平方法計算線性回歸，容易造成預測之回歸直線斜率變化大，因此其預測結果也較容易產生震盪。



(a) (b) (c)