國立臺北教育大學 110學年度第1學期

資料探勘 期中考試卷 日期： 2021/11/16

班級： 　　學號： 　　姓名：

1. 請根據附件1檔案中的*x*和*y*數據資料進行計算，並填入下表：
   1. *x*和*y*數據的平均值、標準差、最大值、最小值。(5分)
   2. *x*和*y*數據的相關係數*r*。(5分)
   3. 根據最小平方法並以RMSE(Root Mean Square Error)作為判斷標準，求出線性迴歸*y* = *w*0+*w*1\**x*的係數、RMSE值、與判定係數*R*2，並畫出*x*和*y*數據散佈圖與線性迴歸曲線。(10分)
   4. 如上題，求出多項式迴歸*y* = *w*0+*w*1\**x*+*w*2\**x*2的係數、RMSE值、與判定係數*R*2，並畫出*x*和*y*數據散佈圖與多項式迴歸曲線。(10分)

Ans：(1) (2) 相關係數 *r* =

|  |  |  |
| --- | --- | --- |
|  | *x*數據 | *y*數據 |
| 平均值 |  |  |
| 標準差 |  |  |
| 最大值 |  |  |
| 最小值 |  |  |

(3) (4)

|  |  |  |  |
| --- | --- | --- | --- |
| 迴歸式 | 係數 | RMSE值 | 判定係數*R*2 |
| 線性迴歸  *y* = *w*0+*w*1\**x* | *w*0 = , *w*1 = , |  |  |
| 多項式迴歸  *y* = *w*0+*w*1\**x*+*w*2\**x*2 | *w*0 = , *w*1 = ,  *w*2 = , |  |  |

1. 說明數據資料的標準化(Standardization)和正規化(Normalization)的方法與程序，並比較兩者的差異處。以附件1檔案進行數據資料的相關計算。(15分)

Ans：

1. 使用最近鄰居分類法(k-Nearest Neighbor Classification)，採用加權平均的方法，調整自變數權重(*w*1, *w*2)與半徑平方(*σ*2)，並以誤差均方根(RMSE)作為判斷標準，前70筆資料為訓練集、後30筆資料為驗證集，進行附件2檔案的KNN分類情形，並將相關數值填入下表。
   1. 限制條件：0.1≤ *w*1, *w*2 ≤10、0.1≤ *σ*2 ≤10，求*w*1, *w*2, *σ*2數值、誤差均方根、誤判率，並畫出分類情形散佈圖。(25分)
   2. 限制條件：0.1≤ *w*1, *w*2 ≤10，分別求出*σ*2 = 10、2、1、0.1時的*w*1, *w*2數值、誤差均方根、誤判率，並畫出分類情形散佈圖。(20分)
   3. 畫出半徑平方倒數vs. 誤差均方根、誤判率的趨勢圖。(10分)

Ans：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *w*1 | *w*2 | *σ*2 | 1/*σ*2 | 誤差均方根 | 誤判率 |
|  |  | 10 | 0.1 |  |  |
|  |  | 2 | 0.5 |  |  |
|  |  | 1 | 1 |  |  |
|  |  |  |  |  |  |
|  |  | 0.1 | 10 |  |  |