分類流程:

步驟一 : 將測試資料集(65筆)帶入各方法建立模型

步驟二 : 計算其訓練資料之錯誤率

步驟三 : 將測試資料帶入各模型並計算真實錯誤率(True Error Rate)

測試資料集(65筆共19個文字變數)

分類方法:

1. 隨機森林

隨機森林是一個集成方法，將數個建立好的模型結果整合在一起，以提升預測的準確性，雖然這方法提供比較好的預測，但他在推論和解適度方面上就會有所限制。隨機森林由好幾個決策樹組合而成，而不同決策樹是由不同抽取的預測變數與觀察值所組成，所以每一棵樹的模型也不盡相同，也正因為如此，是由隨機建立的樹所組成的模型，故稱為隨機森林。

流程:

Step 1 : 將訓練資料集使用拔靴法製造出更多的樣本。

Step 2 : 生成更多的決策樹，每棵決策樹皆由隨機的方式抽取變數及觀察值組成。

Step 3 : 生成的每棵樹都不進行修剪。

Step 4 : 重複Step 1 – Step 3，獲得N棵隨機決策樹。

Step 5 : 將N棵樹的預測進行投票，選取最適合的預測。

結果:

Apparent Rate (表面錯誤率: 1.53 %) True Error Rate (真實錯誤率: 0%)

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 51 | 0 |
| 2 | 1 | 13 |

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 0 | 0 |
| 2 | 0 | 12 |

1. 羅吉斯迴歸

羅吉斯迴歸適用於依變數為二元類別的情形，羅吉斯迴歸不需要考慮資料服從常態性假設，

帶入Logistic regression模型進行計算並算出Apparent Rate與True Error Rate。

結果:

Apparent Rate (表面錯誤率: 1.53 %) True Error Rate (真實錯誤率: 0%)

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 51 | 0 |
| 2 | 2 | 12 |

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 0 | 0 |
| 2 | 0 | 12 |

1. SVM支持向量機

分類資料是機器學習中的一項常見任務。 假設某些給定的資料點各自屬於兩個類別之一，而目標是確定新資料點將在哪個類中。對於支援向量機來說，資料點被視為 p 維向量，而我們想知道是否可以用 (p-1) 維超平面來分開這些點。這就是所謂的線性分類器。可能有許多超平面可以把資料分類。最佳超平面的一個合理選擇是以最大間隔把兩個類分開的超平面。因此，我們要選擇能夠讓到每邊最近的資料點的距離最大化的超平面。如果存在這樣的超平面，則稱為最大間隔超平面，而其定義的線性分類器被稱為最大間隔分類器，或者叫做最佳穩定性感知器。

流程:

Step 1 : 將訓練資料帶入SVM模型，並跑出最佳的懲罰項及gamma值

(cost : 在Lagrange formulation中的大C，決定給被分錯資料的懲罰值)

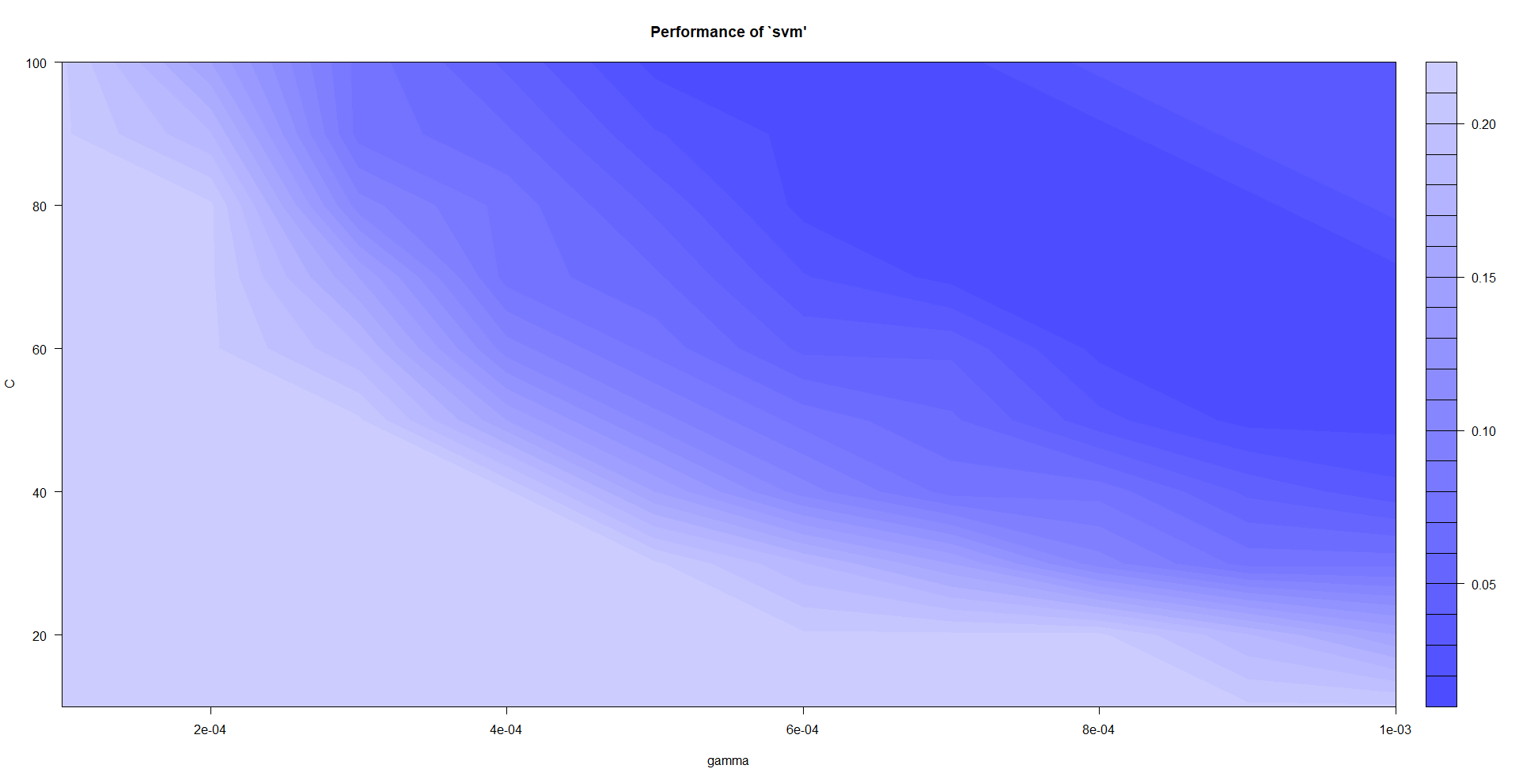
(gamma : 值越大，資料點的影響力範圍越近，對超平面來說，近點的影響力權重

較大，也容易造成overfitting。)

Step 2 : 找出最好的懲罰值及gamma值帶入模型

Step 3 : 將測試資料集帶入模型預測結果

我們從下圖可以觀察出訓練資料集的最適值偏右上方，於是我們選取了cost = 100 作為被分錯資料的懲罰值，並選取了gamma值為0.001作為點與點之間的權重。



結果:

Apparent Rate (表面錯誤率: 0 %) True Error Rate (真實錯誤率: 0 %)

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 51 | 0 |
| 2 | 0 | 14 |

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 0 | 0 |
| 2 | 0 | 12 |

1. LDA 線性判別分析

LDA是一種監督式學習的方法，它的原理是將帶上標籤的數據(點)，通過投影的方法，投影到維度更低的空間中，使得投影後的點，會形成一群一群對應各類別的群體，相同類別的點，將會在投影後的空間更為接近。LDA的基本概念是將高維的樣本變數空間投影到最佳鑑別向量空間，以達到抽取分類信息和壓縮特徵空間維度的效果，因此它是一種有效的特徵選取方法，使用這種方法能夠使投影後的觀察值在新的空間中有最小的類別內的距離和最大類別間的距離，及該模型在該空間中有最佳的可分離性。

流程:

Step 1 將訓練資料集帶入模型建立出最佳的投影空間

Step 2 找出最佳的線性分類器在投影空間中分割各類別

Step 3 將測試訓練集帶入模型中進行分類

結果:

Apparent Rate (表面錯誤率: 1.53 %) True Error Rate (真實錯誤率: 0 %)

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 50 | 1 |
| 2 | 0 | 14 |

|  |  |  |
| --- | --- | --- |
|  | 1 | 2 |
| 1 | 0 | 0 |
| 2 | 0 | 12 |

1. 結論: