312512011 李效賢 Machine Learning HW#4

RBF SVM格子搜尋法輸出結果：

橫軸為C值，縱軸為log1.05 (sigma)值，格子內為取到小數點第四位的分類正確率，而紅字者為同樣C value下，accuracy最高的(C, sigma)組合；紅底白字的格子則為一組表格內分類率最高者(意味這對這組dataset而言的最佳參數)。

Fold No.1

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 5 | 10 | 50 | 100 | 500 | 1000 |
| -100 | 48.00 | 62.67 | 36.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| -95 | 36.00 | 0.00 | 34.67 | 0.00 | 54.67 | 0.00 | 54.67 |
| -90 | 62.67 | 38.67 | 38.67 | 62.67 | 62.67 | 62.67 | 62.67 |
| -85 | 40.00 | 66.67 | 40.00 | 38.67 | 38.67 | 40.00 | 66.67 |
| -80 | 42.67 | 66.67 | 41.33 | 41.33 | 66.67 | 66.67 | 66.67 |
| -75 | 66.67 | 44.00 | 66.67 | 44.00 | 48.00 | 44.00 | 66.67 |
| -70 | 46.67 | 46.67 | 46.67 | 46.67 | 46.67 | 46.67 | 46.67 |
| -65 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 |
| -60 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 |
| -55 | 68.00 | 68.00 | 68.00 | 68.00 | 68.00 | 68.00 | 68.00 |
| -50 | 72.00 | 72.00 | 72.00 | 72.00 | 72.00 | 72.00 | 72.00 |
| -45 | 80.00 | 80.00 | 80.00 | 80.00 | 80.00 | 80.00 | 80.00 |
| -40 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 |
| -35 | 92.00 | 90.67 | 90.67 | 90.67 | 90.67 | 90.67 | 90.67 |
| -30 | 94.67 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 |
| -25 | 93.33 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| -20 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| -15 | 94.67 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 |
| -10 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| -5 | 96.00 | 97.33 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| 0 | 96.00 | 94.67 | 93.33 | 97.33 | 97.33 | 97.33 | 97.33 |
| 5 | 96.00 | 94.67 | 97.33 | 94.67 | 97.33 | 97.33 | 97.33 |
| 10 | 94.67 | 94.67 | 94.67 | 93.33 | 97.33 | 97.33 | 97.33 |
| 15 | 94.67 | 94.67 | 93.33 | 93.33 | 94.67 | 97.33 | 97.33 |
| 20 | 93.33 | 94.67 | 93.33 | 93.33 | 93.33 | 97.33 | 97.33 |
| 25 | 93.33 | 94.67 | 94.67 | 93.33 | 93.33 | 97.33 | 97.33 |
| 30 | 94.67 | 94.67 | 94.67 | 93.33 | 93.33 | 94.67 | 97.33 |
| 35 | 90.67 | 94.67 | 94.67 | 94.67 | 93.33 | 93.33 | 96.00 |
| 40 | 90.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| 45 | 90.67 | 92.00 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| 50 | 89.33 | 92.00 | 93.33 | 94.67 | 94.67 | 94.67 | 94.67 |
| 55 | 89.33 | 94.67 | 92.00 | 94.67 | 94.67 | 94.67 | 94.67 |
| 60 | 90.67 | 94.67 | 93.33 | 94.67 | 94.67 | 94.67 | 94.67 |
| 65 | 77.33 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 |
| 70 | 69.33 | 94.67 | 94.67 | 92.00 | 94.67 | 94.67 | 94.67 |
| 75 | 69.33 | 94.67 | 94.67 | 93.33 | 92.00 | 94.67 | 94.67 |
| 80 | 70.67 | 94.67 | 94.67 | 94.67 | 93.33 | 94.67 | 94.67 |
| 85 | 66.67 | 94.67 | 94.67 | 94.67 | 93.33 | 94.67 | 94.67 |
| 90 | 66.67 | 94.67 | 94.67 | 94.67 | 94.67 | 93.33 | 94.67 |
| 95 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 92.00 | 94.67 |
| 100 | 94.67 | 94.67 | 94.67 | 94.67 | 94.67 | 93.33 | 92.00 |

最佳參數組合：

(C = 5, sigma = 1.05^-5)  
(C = 10, sigma = 1.05^5)、

(C = 100, sigma = 1.05^0)、(C = 100, sigma = 1.05^5)

(C=500, sigma = 1.05^0)、(C=500, sigma = 1.05^5)、(C=500, sigma = 1.05^10)、

(C=500, sigma = 1.05^15)、(C=500, sigma = 1.05^20)、(C=500, sigma = 1.05^25)

(C = 1000, sigma = 1.05^0)、(C = 1000, sigma = 1.05^5) 、(C = 1000, sigma = 1.05^10)

(C = 1000, sigma = 1.05^15)、(C = 1000, sigma = 1.05^20)、(C = 1000, sigma = 1.05^25)

(C = 1000, sigma = 1.05^30)

(分類率97.33%參數組合，共17組；另外分類率0%者表示出現Tied情況)

Fold No.2

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 5 | 10 | 50 | 100 | 500 | 1000 |
| -100 | 34.67 | 45.33 | 45.33 | 34.67 | 34.67 | 36.00 | 34.67 |
| -95 | 36.00 | 50.67 | 34.67 | 50.67 | 34.67 | 36.00 | 34.67 |
| -90 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 |
| -85 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 |
| -80 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 |
| -75 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 |
| -70 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 | 36.00 |
| -65 | 40.00 | 40.00 | 40.00 | 40.00 | 40.00 | 40.00 | 40.00 |
| -60 | 52.00 | 52.00 | 52.00 | 52.00 | 52.00 | 52.00 | 52.00 |
| -55 | 58.67 | 58.67 | 58.67 | 58.67 | 58.67 | 58.67 | 58.67 |
| -50 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 | 60.00 |
| -45 | 69.33 | 69.33 | 69.33 | 69.33 | 69.33 | 69.33 | 69.33 |
| -40 | 81.33 | 81.33 | 81.33 | 81.33 | 81.33 | 81.33 | 81.33 |
| -35 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 | 89.33 |
| -30 | 93.33 | 92.00 | 92.00 | 92.00 | 92.00 | 92.00 | 92.00 |
| -25 | 97.33 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 |
| -20 | 96.00 | 94.67 | 97.33 | 97.33 | 97.33 | 97.33 | 97.33 |
| -15 | 96.00 | 94.67 | 94.67 | 97.33 | 97.33 | 97.33 | 97.33 |
| -10 | 96.00 | 96.00 | 94.67 | 96.00 | 96.00 | 96.00 | 96.00 |
| -5 | 96.00 | 96.00 | 96.00 | 94.67 | 96.00 | 96.00 | 96.00 |
| 0 | 96.00 | 96.00 | 96.00 | 96.00 | 94.67 | 96.00 | 96.00 |
| 5 | 97.33 | 97.33 | 96.00 | 96.00 | 97.33 | 96.00 | 96.00 |
| 10 | 97.33 | 96.00 | 97.33 | 96.00 | 96.00 | 94.67 | 96.00 |
| 15 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 | 97.33 | 96.00 |
| 20 | 94.67 | 96.00 | 97.33 | 96.00 | 96.00 | 96.00 | 97.33 |
| 25 | 92.00 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 | 97.33 |
| 30 | 94.67 | 97.33 | 97.33 | 97.33 | 96.00 | 97.33 | 96.00 |
| 35 | 93.33 | 96.00 | 97.33 | 97.33 | 98.67 | 98.67 | 97.33 |
| 40 | 93.33 | 96.00 | 97.33 | 97.33 | 98.67 | 98.67 | 98.67 |
| 45 | 90.67 | 94.67 | 94.67 | 97.33 | 97.33 | 98.67 | 98.67 |
| 50 | 89.33 | 93.33 | 96.00 | 96.00 | 97.33 | 98.67 | 98.67 |
| 55 | 90.67 | 93.33 | 94.67 | 97.33 | 96.00 | 98.67 | 98.67 |
| 60 | 89.33 | 93.33 | 93.33 | 94.67 | 97.33 | 97.33 | 98.67 |
| 65 | 85.33 | 92.00 | 93.33 | 96.00 | 96.00 | 97.33 | 98.67 |
| 70 | 82.67 | 92.00 | 92.00 | 94.67 | 94.67 | 97.33 | 97.33 |
| 75 | 77.33 | 92.00 | 92.00 | 93.33 | 96.00 | 97.33 | 97.33 |
| 80 | 73.33 | 92.00 | 92.00 | 93.33 | 94.67 | 97.33 | 96.00 |
| 85 | 92.00 | 92.00 | 92.00 | 92.00 | 93.33 | 94.67 | 97.33 |
| 90 | 92.00 | 92.00 | 92.00 | 92.00 | 93.33 | 96.00 | 96.00 |
| 95 | 92.00 | 92.00 | 92.00 | 92.00 | 92.00 | 94.67 | 96.00 |
| 100 | 92.00 | 92.00 | 92.00 | 92.00 | 92.00 | 93.33 | 94.67 |

以下為分類率98.67%的參數組合

(C = 100, sigma = 1.05^35)、(C = 100, sigma = 1.05^40)

(C = 500, sigma = 1.05^35)、(C = 500, sigma = 1.05^40)、(C = 500, sigma = 1.05^45)、(C = 500, sigma = 1.05^50)、(C = 500, sigma = 1.05^55)

(C = 1000, sigma = 1.05^40)、(C = 1000, sigma = 1.05^45)、

(C = 1000, sigma = 1.05^50)、(C = 1000, sigma = 1.05^55)、(C = 1000, sigma = 1.05^60)

(C = 1000, sigma = 1.05^65)

(分類率98.67%參數組合，共13組)

Average of 2 folds

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 5 | 10 | 50 | 100 | 500 | 1000 |
| -100 | 41.33 | 54.00 | 40.67 | 17.33 | 17.33 | 18.00 | 17.33 |
| -95 | 36.00 | 25.33 | 34.67 | 25.33 | 44.67 | 18.00 | 44.67 |
| -90 | 49.33 | 37.33 | 37.33 | 49.33 | 49.33 | 49.33 | 49.33 |
| -85 | 38.00 | 51.33 | 38.00 | 37.33 | 37.33 | 38.00 | 51.33 |
| -80 | 39.33 | 51.33 | 38.67 | 38.67 | 51.33 | 51.33 | 51.33 |
| -75 | 51.33 | 40.00 | 51.33 | 40.00 | 42.00 | 40.00 | 51.33 |
| -70 | 41.33 | 41.33 | 41.33 | 41.33 | 41.33 | 41.33 | 41.33 |
| -65 | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 |
| -60 | 56.00 | 56.00 | 56.00 | 56.00 | 56.00 | 56.00 | 56.00 |
| -55 | 63.33 | 63.33 | 63.33 | 63.33 | 63.33 | 63.33 | 63.33 |
| -50 | 66.00 | 66.00 | 66.00 | 66.00 | 66.00 | 66.00 | 66.00 |
| -45 | 74.67 | 74.67 | 74.67 | 74.67 | 74.67 | 74.67 | 74.67 |
| -40 | 85.33 | 85.33 | 85.33 | 85.33 | 85.33 | 85.33 | 85.33 |
| -35 | 90.67 | 90.00 | 90.00 | 90.00 | 90.00 | 90.00 | 90.00 |
| -30 | 94.00 | 92.67 | 92.67 | 92.67 | 92.67 | 92.67 | 92.67 |
| -25 | 95.33 | 95.33 | 95.33 | 95.33 | 95.33 | 95.33 | 95.33 |
| -20 | 95.33 | 94.67 | 96.00 | 96.00 | 96.00 | 96.00 | 96.00 |
| -15 | 95.33 | 95.33 | 95.33 | 96.67 | 96.67 | 96.67 | 96.67 |
| -10 | 95.33 | 95.33 | 94.67 | 95.33 | 95.33 | 95.33 | 95.33 |
| -5 | 96.00 | 96.67 | 95.33 | 94.67 | 95.33 | 95.33 | 95.33 |
| 0 | 96.00 | 95.33 | 94.67 | 96.67 | 96.00 | 96.67 | 96.67 |
| 5 | 96.67 | 96.00 | 96.67 | 95.33 | 97.33 | 96.67 | 96.67 |
| 10 | 96.00 | 95.33 | 96.00 | 94.67 | 96.67 | 96.00 | 96.67 |
| 15 | 95.33 | 95.33 | 94.67 | 94.67 | 95.33 | 97.33 | 96.67 |
| 20 | 94.00 | 95.33 | 95.33 | 94.67 | 94.67 | 96.67 | 97.33 |
| 25 | 92.67 | 95.33 | 95.33 | 94.67 | 94.67 | 96.67 | 97.33 |
| 30 | 94.67 | 96.00 | 96.00 | 95.33 | 94.67 | 96.00 | 96.67 |
| 35 | 92.00 | 95.33 | 96.00 | 96.00 | 96.00 | 96.00 | 96.67 |
| 40 | 92.00 | 95.33 | 96.00 | 96.00 | 96.67 | 96.67 | 96.67 |
| 45 | 90.67 | 93.33 | 94.67 | 96.00 | 96.00 | 96.67 | 96.67 |
| 50 | 89.33 | 92.67 | 94.67 | 95.33 | 96.00 | 96.67 | 96.67 |
| 55 | 90.00 | 94.00 | 93.33 | 96.00 | 95.33 | 96.67 | 96.67 |
| 60 | 90.00 | 94.00 | 93.33 | 94.67 | 96.00 | 96.00 | 96.67 |
| 65 | 81.33 | 93.33 | 94.00 | 95.33 | 95.33 | 96.00 | 96.67 |
| 70 | 76.00 | 93.33 | 93.33 | 93.33 | 94.67 | 96.00 | 96.00 |
| 75 | 73.33 | 93.33 | 93.33 | 93.33 | 94.00 | 96.00 | 96.00 |
| 80 | 72.00 | 93.33 | 93.33 | 94.00 | 94.00 | 96.00 | 95.33 |
| 85 | 79.33 | 93.33 | 93.33 | 93.33 | 93.33 | 94.67 | 96.00 |
| 90 | 79.33 | 93.33 | 93.33 | 93.33 | 94.00 | 94.67 | 95.33 |
| 95 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 | 95.33 |
| 100 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 | 93.33 |

最佳參數：

(C = 500, sigma = 1.05^50)、(C = 1000, sigma = 1.05^55)共兩組，平均分類率98.67%

討論：

1. 請問在grid search的結果中，C的大小與分類率高低有何關係?

Ans. C在SVM演算法中代表Punish weight(懲罰權重)，C較小時訓練出的模型相對能夠容忍錯誤分類，一般會讓bias較大，但泛化程度會較佳；反過來C較大時，模型容忍的錯誤較少，會讓bias較小，但較有可能overfitting。

而依照我所製作出的圖表觀察結果，在σ**∈** [1.05-40,1.05100]之間時，C值對兩個fold以及全體平均的影響相當小，而以橫向row來看，相同σ值之下，C值對整體分類正確率影響有限。

2. sigma的大小的改變與分類率是否有關係? 若有，請探討sigma的差異與特徵的數值有什麼關聯性?

Ans. 是，在作業要求σ**∈** [1.05-100,1.05100]的集合中，當σ極小(越接近1.05-100)，分類率越低，特別是σ在1.05-35以下時，不論哪個Fold的分類正確率都低於90%，依照RBF kernel定義，應為overfitting現象(整體Kernel function越尖，此時分類器對dataset中的noise越敏感)；另一方面，當σ在1.0580以上時，分類正確率會從95%以上顯著下降，則是underfitting (整體Kernel function較平滑，此時分類器可以說”不夠敏感”，但是泛化性(Generalization)較好)。當σ**∈** [1.0550,1.0570]時，通常會是最佳解發生的區域。

關於sigma差異與特徵數值的關聯性：如果特徵的數值範圍很大，那麼小的 sigma 可能會導致模型對特定數據的變化非常敏感，因為相似度的計算會更加依賴特徵值之間的距離。

在sigma值過小的情況下，在Fold1會出現平手情況(“Tied” in one-against-one strategy)，個人認為就是overfitting所造成的，其次sigma值相當大的時候，大多數的C值搭配下，模型分類率並沒有降到90%以下，對比之下更能觀察出sigma過小時過度擬合(Overfitting)現象。

3. 若分析過程不採用two-fold cross validation，則分類率是否會更高? 請探討之。

Ans. 以數據來說有，但是差異不明顯，甚至可以說在誤差範圍。

我在excel上有做出一個Fold1正確率減去平均正確率的表格，Fold1整體正確率大約比平均的高1.04%左右(也就是大約1組數據在Fold1會分對，在Fold2的模型會分錯)，個人認為這可能是由於數據組數量偏少(testset 75組)造成的誤差，而非cross-validation所造成分類率下降。因為cross-validation本身就是在評估生成模型的泛化性，確認是在排列組合都有不差的表現，還是只對某一組數據表現很好，其他排列組合分類率不佳。

在sigma極小的情況下會出現兩者相差很大，但在sigma較大時則無。

此圖為Fold1 accuracy - average accuracy對log(sigma)的分布圖