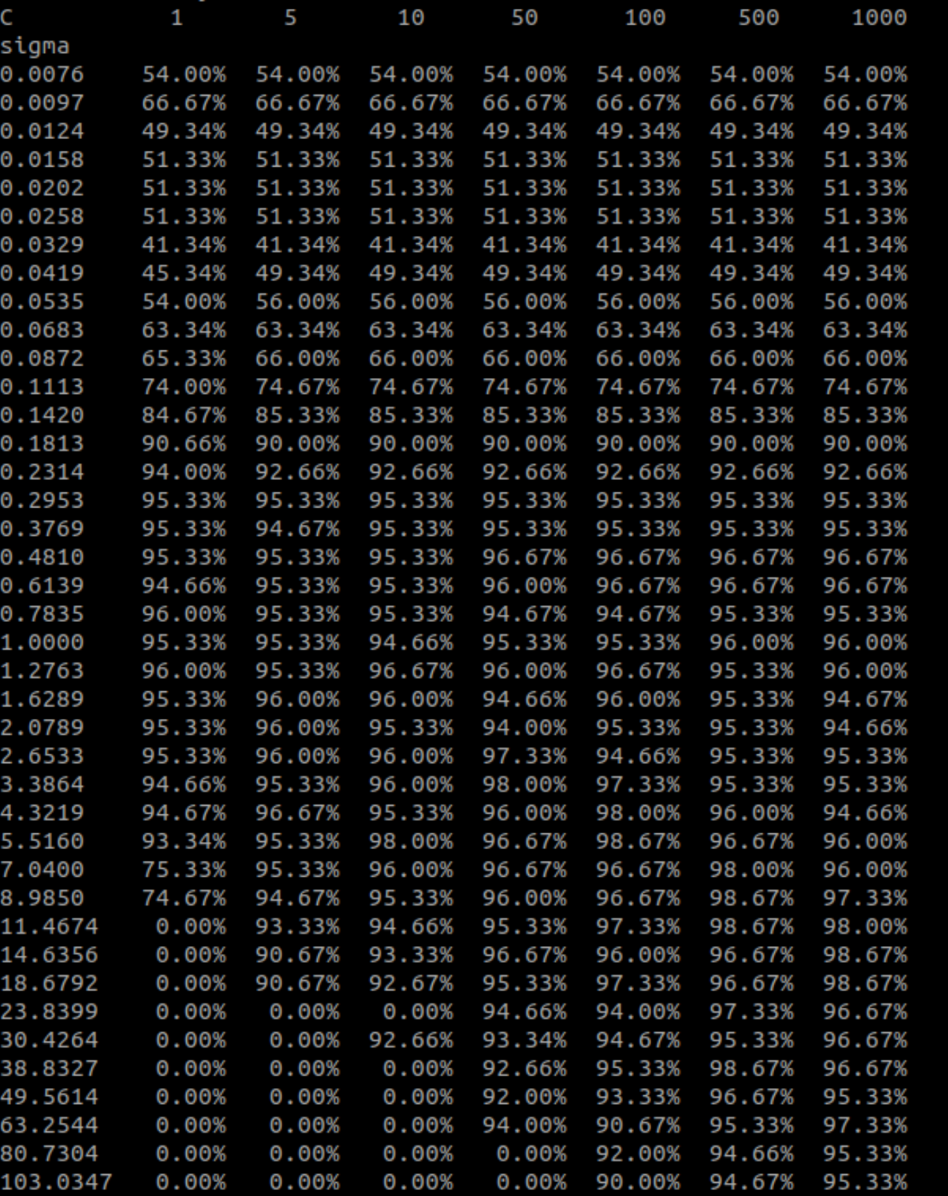
NYCU Machine Learning HW1

312605001

機器人碩士 歐庭維

* 實驗結果

1. Grid search ( SVM + 2 fold )參數組合對應平均分類率



1. 最佳分類率

A black background with white text

Description automatically generated

1. 不採用2-fold validation進行平均之分類率

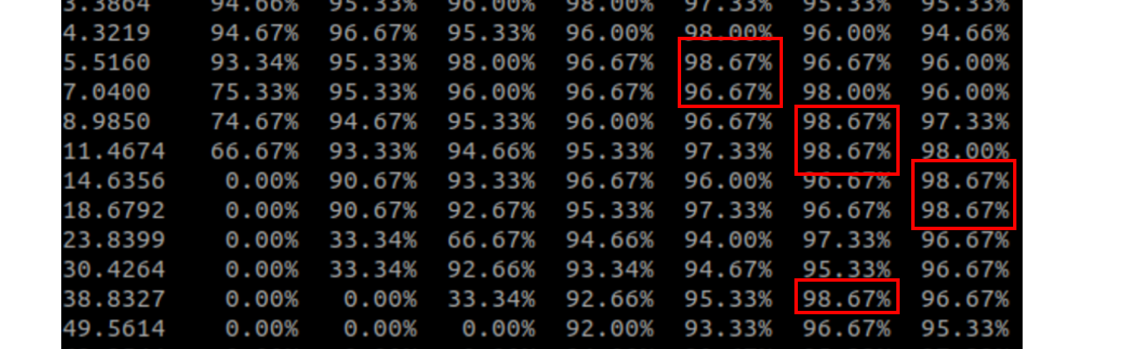
A black background with white text

Description automatically generated

* 請問在grid search 的結果中，C的大小與分類率的高低有何關係？

在SVM模型訓練中，C參數代表著對於分類錯誤的懲罰力度。用簡單的話來說，C就像是一個控制模型「犯錯成本」的項。 如果C值太小，模型對於訓練數據的錯誤不太敏感，這可能導致模型對於新數據的分類表現不佳。反之，如果C值太大，模型會極力避免任何訓練數據的分類錯誤，這可能導致模型對於新數據的泛化能力下降，即過度擬合。 因此，調整適當的C值就像是在平衡模型的「容錯度」，使其在訓練和預測時都能夠有良好的表現。

而在這次作業四中，我使用了grid search去找到C從1 ~ 1000時的表現，如下所示



C : 5 🡪 1000

可以發現分類率最高為98.67，並主要集中於C : 100 , 500, 1000。

而我也想觀察當C變得更大時會不會有overfitting的表現，因此我多加了C=1500 / 2000/ 3000（sigma = 14.6356），結果如下：



可以發現隨c的變大，分數也隨之降低，我推測是overfitting導致模型的泛用性降低。透過整體觀察，可以發現最高分類率的分佈也大致符合在C值大小適中的位置，符合一開始討論的部分。

* Sigma的大小的改變與分類率是否有關係？

在上一次作業討論過，SVM的Sigma主要控制RBF函數的形狀，當sigma越大，會使得rbf kernel 更平緩（泛化性強），當sigma越大，則會導致考慮太多noise導致overfitting。在本次作業中，對於相同的C來說，最佳的sigma分佈大約是在5.51 ~ 38.3左右，大致符合理論所述。

* 若分析過程不採用2-fold validation ，分類率是否會更高？

我發現當只有使用單獨的前半或後半部分時，分類率最高可以達到100%，而我推測這是因為有可能子集的表現剛好被包含於另一部分，導致表現非常的完美，但是當其反過來時，表現就會更差，因此使用s-fold validation就是為了避免這種極端狀態，使結果可以更平均、客觀。

A black background with white text

Description automatically generated

* 其他討論

我發現在執行SVM訓練時，會跑出以下錯誤

A screen shot of a computer code

Description automatically generated

這是因為當alpha沒有在0~C之間時，會導致b（bias）無法算出，這會導致三個分類器只有兩個有用，這會導致投票時會出問題，因此我將正確率過低的結果都以分類錯誤表示，直接設為0，除了過濾不佳的結果外，也去除有問題的數據。

另外由於one against one在三個分類器的預測結果進行投票時，可能會有三個同票的問題，而我也將這種狀況的分類率設為0，視作分類失敗。

