Machine Learning Homework1

# Introduction

此篇文章為機器學期課程的第一份作業，透過 crx.csv 與 data.csv 兩種資料集來實作多種分類模型，並比較多種調整方式與已被廣泛使用的分類套件之間的成效差異。

# Methodology of Use

## Linear classifier with update rule

訓練階段一次讀取單筆資料，若分類正確則讀取下筆資料，若分類不正確，透過標籤答案與預測答案之差異，判斷各特徵的調整方向，個特徵調整的幅度透過 Learning Rate 超參數控制，其值會隨著訓練迭代次數而逐漸降低，整個訓練過程直至分類結果收斂或第最大迭代次數。

## Linear classifier with Least-squared manner

根據最小平方法的原則，訓練過程一步到位，一次讀取所有資料點，透過公式解搭配矩陣運算，求出最佳的模型參數。

## Voted perceptron and Average voted perceptron

基於 A. Linear classifier的架構，但是在訓練過程中將每次更改模型時的歷史參數都記錄起來，搭配對應的訓練正確率作為投票權重，以 Voted 和 Average voted 兩種加權方式進行預測任務。

## Large margin classifier

同樣基於 A. Linear Classifier 的架構，也與 C. 同樣在訓練過程中將所有的模型歷史參數都記錄起來，但不同的是選擇最佳模型時，要先計算所有歷史模型參數的 ||W||^2，並選擇最小值的對應參數當作訓練最終的模型參數.

## Soft large margin classifier

當訓練過程中，遇到模型無法正確分類的資料點，是無法計算及對應的 margin 的，為了使得其產生反向影響力，增加分錯時的成本，加入 C 超參數來調整 penalty 程度。

# Experiment

## Experiment settings

將兩種資料 crx.csv 和 data.csv 皆以 70:30 比例切分訓練資料集與驗證資料集，將各種模型以訓練資料及訓練，並以測試資料集測試模型分類成效，RANDOM\_STATE 超參數預設設定為 11，Learning Rate 預設設定為 0.1，最大疊代次數預設設定為 50 次。

## Performance on classification task

| Dataset | Method | Accuracy |
| --- | --- | --- |
| crx | Linear classifier  (update rule) | 0.702 |
| Linear classifier  (least-squared manner) | 0.880 |
| Voted perceptron | 0.875 |
| Average Voted perceptron | 0.875 |
| Large margin classifier | 0.755 |
| Soft large margin classifier | 0.755 |
| Existing SVM | 0.875 |
| data | Linear classifier  (update rule) | 0.953 |
| Linear classifier  (least-squared manner) | 0.947 |
| Voted perceptron | 0.953 |
| Average Voted perceptron | 0.947 |
| Large margin classifier | 0.942 |
| Soft large margin classifier | 0.965 |
| Existing SVM | 0.965 |

# Conclusion

Linear classifier (update rule) 分類的成效普通 ，最大原因可能是訓練過程中每次只關注單個樣本，可能導致顧此失彼，為了正確分類新的樣本，而讓已經分類正確的樣本又再次變成錯誤的分類，相當依賴訓練過程中樣本輸入的順序，而考量所有資料，一步到位訓練的Linear classifier (least-squared manner) 分類方法，則有不錯的成效，但是 Linear classifier (update rule) 透過加權的考量，不管是 Voted perceptron 或是 Average Voted perceptron 都大大提升了分類成效來到 0.86/0.95（對應 crx/data 資料集），甚至達到了與 Existing SVM 相當的成效。

以上所有模型的相關實作程式碼都有上傳至 Github：

https://github.com/littlemilkwu/ds\_ncku\_machine\_learning