# Data Mining HW5

LIBSVM

Name: 吳睿哲　　Student ID: r06921095

### 5.1 Iris dataset : Testing label is provided.

a. Comparison of performance with and without scaling. [5%]

ans:

| | Scaling(-1,1) | un-scaling |
|---|---|---|
| Training accuracy | 0.96 | 0.986667 |
| Testing accuracy | 0.973333 | 1.0 |

b. Comparison of different kernel functions. [5%]

ans:

| | linear | polynomial | RBF |
|---|---|---|---|
| Training accuracy | 1.0 | 0.986667 | 0.986667 |
| Testing accuracy | 0.986667 | 0.986667 | 0.973333 |

c. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%]

ans:

我是用 grid.py 尋找適合的 kernel,再用排列組合的方式尋找 c 跟 g
svm_type: C-SVC
kernel_type: linear
un-scaling

training accuracy: 0.986667
testing accuracy: 1.0

d. More discussions is welcome. [Bonus 1%]

ans:

通常 kernal 會使用 linear 或者是 RBF,linear 的速度較快,但是 RBF 的解準確率較高。

### 5.2 News dataset : Testing label is provided.

a. Comparison of performance with and without scaling. [5%]

ans:

| | scaling(0,1) | un-scaling |
|---|---|---|
| Training accuracy | 0.9772 | 0.9772 |
| Testing accuracy | 0.2867 | 0.8462 |

b. Comparison of 5-1-a and 5-2-a. [5%]

ans:

5-1 在做完 scaling 之後,performance 沒有太大的改變,但是 5-2 做完 scaling.performance 會變很差。

c. Comparison of different kernel functions. [5%]

ans:

| | linear | polynomial | RBF |
|---|---|---|---|
| Training accuracy | 0.842 | 0.2778 | 0.2778 |
| Testing accuracy | 0.9767 | 0.2776 | 0.2776 |

d. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%, Surpass baseline 5%]

ans:

我是用 grid.py 尋找適合的 kernel,再用排列組合的方式尋找 c 跟 g
我用的 svm type 是 C-SVC
kernel type 是： RBF
gamma: 1
c: 64

training accuracy: 0.9772
testing accuracy: 0.8462

e. We know that the curse of dimensionality causes overfitting. How does it influence Naive Bayesian, Decision Tree and SVM separately? [5%]

ans:

Naive bayes: 對 naive bayes 的影響較小。
Decision tree: curse of dimensionality 會造成 decision tree 準確率瞬間下降很多。
svm: 會造成 svm 龐大的計算量。

f. More discussions is welcome. [Bonus 1%]

ans:

因為資料本身就已經是 tfidf 的呈現方式,所以不需要再特別去做 scaling。

### 5.3 Abalone dataset : Testing label is provided.

a. Your data preprocessing and scaling range. Please state clearly. [10%]

ans:

這題的前處理主要在於轉換成 libsvm 能讀的格式,先將 label 放在每行的第一個位置,之後再依序方入 feature。

再來是 scaling 的部份,我選擇的 range 是[0,1]

b. Comparison of different kernel functions. [5%]

ans:

| | linear | polynomial | RBF |
|---|---|---|---|
| Training accuracy | 0.6340 | 0.5542 | 0.5932 |
| Testing accuracy | 0.6453 | 0.5705 | 0.5983 |

c. Parameter set and performance of your best model. (Report training accuracy and testing accuracy) [5%, Surpass baseline 5%]

ans:

我是用 grid.py 尋找適合的 kernel,再用排列組合的方式尋找 c 跟 g

我用的 svm type 是 C-SVC

kernel type 是： RBF

gamma: 0.5

c: 512

training accuracy: 0.6873

testing accuracy: 0.6702

d. More discussion is welcome. [Bonus 1%]

ans:

如果原本的資料本身範圍沒有差很多,有沒有做 scaling 對 performane 不會造成很大的影響。

### 5.4 Income dataset

a. Your data preprocessing / data cleaning. Please state clearly. [10%]

ans:

這題的前處理,我首先將 categorial 以及 continuous 的資料分開處理,categorical 的資料以 one hot encoding 處理,missing value 的部份則直接忽略,最後再統一以[-1,1]的範圍做 scaling。

b. How do you choose parameters set and kernel function ? [5%]

ans:
我是用 grid.py 尋找適合的 kernel,再用排列組合的方式尋找 c 跟 g
我用的 svm type 是 C-SVC
kernel type 是： RBF
c 跟 g 的 range 是跟 libsvm 講義中選的一樣。

c. Report cross validation accuracy, and testing accuracy. [5%]

ans:
10 ford validation accuracy: 0.8562
testing accuracy(80% training 20% testing):0.8518

d. Parameter set of your best model. [Surpass baseline 5%, Top 20% in class: 5%]

ans:
我是用 grid.py 尋找適合的 kernel,再用排列組合的方式尋找 c 跟 g
我用的 svm type 是 C-SVC
kernel type 是： RBF
gamma 是 0.005
c 是 24
scaling range 是[-1,1]

e. More discussion or observation are welcome. [Bonus 1%]

ans:
這提使用 linear model 的效果叫跟 RBF 差不多,因為我使用的是 sparse 的資料形式,而且還可以達到加速的效果。