Data Mining HW5

LIBSVM

Name: Student ID:

5.1 Iris dataset : Testing label is provided.

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

SVM(default) :
Without scaling : 97.333%
Scaling : 98.667%

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

(只變動 –t kernel function , 其他參數固定)
0(linear): 100%
1(polynomial): 98.6667%
2(rbf): 97.333%
3(sigmoid): 96%

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

Parameter set : SVM –t 0
Training score : 98.6667%
Testing score : 100%

d. More discussions is welcome. [Bonus 1%]
Iris dataset 算是 feature 維度相當少的資料集，在簡單的任務上單純的線性分類(ex: linear)表現可能會優於複雜的非線性分類器(ex: rbf)

5.2 News dataset : Testing label is provided.

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

SVM(default) :
Without scaling : 27.7622%
Scaling : 27.7622%
(表現似乎一樣糟..)

b. Comparison of 5-1-a and 5-2-a. [5%]
5-1-a 中：做完 scaling 有助於將資料的數值大小維持在適當的範圍內，有助於增長訓練的效果
5-2-a 中：在 news dataset 中，資料是很 sparse 的狀態，相當數量的維度都為 0，此在 libsvm 的資料表示中能夠不予以記錄，但在做 scale 之後，會使得很多原本為 0 的項變得不為 0，在 libsvm 型式的資料量會變得相當龐大，推測可能於此有關。

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

0(linear): 84.1958%
1(polynomial): 27.7622%
2(rbf): 27.7622%
3(sigmoid): 27.7622%

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

Parameter set : SVM –t 0 -c 1.25
Training score : 97.6733%
Testing score : 84.3357%

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

將維數災害對於分類器的影響分成兩個面向: 線性分類以及非線性分類

線性分類器，由於不會畫出過於擬合資料的非線性邊界所以比較不會受到維數災害影響而 overfitting ex: naive Bayes ,SVM(kernel : linear 等線性 kernel)
非線性分類器，能夠建立非常精確的非線性的決策邊界，易在資料數量不足夠多又有過高的維度時發生維數災害造成 overfitting ex: decision tree, neural network, SVM(kernel : RBF 等非線性 kernel)

f. More discussions is welcome. [Bonus 1%]
在遇到很 sparse 的資料的時候，一開始直覺為高維度的問題，就會想利用 rbf，但查了一些資料之後顯示，在處理 sparse 資料的時候，linear 可以有很不錯的表現

5.3 Abalone dataset : Testing label is provided.

a. Your data preprocessing and scaling range. Please state clearly. [10%]
(沒有做 scale)
   1. 首先先將 rings 的 label 321 改成 210 來符合 libsvm 的格式
   2. 由於 feature 中有 sex，利用 pd.get_dummies 將一維的 sex 映射到三個維度: sex_m , sex_f , sex_i 來表示
   3. 將上述出來的結果輸出成 libisvm 可接受的格式
   (train 和 test 的資料都要一起做上述處理)

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

   0(linear): 64.5254%
   1(polynomial): 57.047%
   2(rbf): 59.8274%
   3(sigmoid): 57.4305%

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

   Parameter set : SVM –t 0 -c 2
   Training score : 63.9438%
   Testing score : 65.3883%

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

5.4 Income dataset

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

   1.先 drop 掉一些認為不需要的 attribute
   2.接著對資料做 pd.dummies，映射到高維度
   3.在對資料做 normalization

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

   Kernel function 的選擇部分，因為資料是屬於高維度，考慮可以使用 rbf 或 linear 去做嘗試。

參數的部分，c 的部分經過嘗試一旦大於 100 便難以收斂出結果
所以考慮參數調整的範圍可以用 default c = 1 去做嘗試

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

Validation accuracy(-v 5) : 86.2601%
Training accuracy : 87.0432%

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

Parameter set : SVM(default)(kernel : RBF)

e. More discussion or observation are welcome. [Bonus 1%]
在過程中，一開始一直執著於參數的調整，希望能夠找到最佳組合
但嘗試的過程中表現一直不如預期，最後才回歸到最基本的資料前處理
Drop 掉多餘的資訊，將資料利用 dummy 映射到較高維的空間，做標準化或正規化，這些都應該要在進行 model 訓練之前進行完善的處理，表現才能夠提升。