機器學習實務與應用

Homework #6
Due 2019 April 8 9:00AM

本次作業含以下四個部分，而每個部分都有提供相對應資料檔(hw6_partx_data.csv)供處理。每個部分的作業除了繳交 jupyter notebook 檔案做為資料處理結果作為評分依據（請記得在 jupyter notebook 加入報告說明內容）之外，另一部分評分將會依據你所另外提供之 py 檔，針對我們另行保留的測試資料檔之執行結果來評分。底下有說明保留測試檔的名稱及要求處理後輸出的檔案格式，附件有提供保留測試檔的範例，但非實際的保留檔。

==========

Part 1: Linear Regression
資料檔案： hw6_part1_data.csv
目標：預測 MeanTemp
模型：使用 Regression 的 model
評分標準：均方誤差(mean-square error)越小越好
使用 jupyter notebook 繳交報告，其中需有模型訓練的方式，檔名為 ML_HW6_part1_<學號>.ipynb
可參考的處理方式
1. 使用預處理 Normalization, Standardization 等
2. Regularization
3. 找出比較好的因子

Part 2: Classification
資料檔案：hw6_part2_data.csv
目標：預測 Survived == 1 or Survived == 0，是否成功生存
模型：Logistic Regression
評分標準：AUC 越大越好
使用 jupyter notebook 繳交報告，其中需有模型訓練的方式，檔名為 ML_HW6_part2_<學號>.ipynb
可參考的處理方式
1. 比較 attribute 對預測結果的影響強弱
2. 使用 confusion matrix，進行預測結果的展示
3. 畫出 ROC curve

Part 3: Decision Tree
資料檔案： hw6_part3_data.csv
目標：預測 target == 1 or target == 0，是否有心臟病
模型：Decision Tree
評分標準：AUC 越大越好
使用 jupyter notebook 繳交報告，其中需有模型訓練的方式，檔名為

ML_HW6_part3_<學號>.ipynb
可參考的處理方式
1. 使用不同的深度
2. 找出比較好的因子

Part 4: SVM
資料檔案： hw6_part4_data.csv
目標：class == e or class == p，香菇可食用或有毒
模型：SVM
評分標準：AUC 越大越好
使用 jupyter notebook 繳交報告，其中需有模型訓練的方式，檔名為
ML_HW6_part4_<學號>.ipynb
可參考的處理方式
1. 採用不同的 kernel function
2. 畫出 Decision Boundary

請繳交壓縮檔，檔名為 ML_HW6_<學號>.zip
其中需包含的檔案及資料夾為：
1. 資料夾 Part1
   ML_HW6_part1_<學號>.ipynb
   ML_HW6_part1_<學號>_test.py
2. 資料夾 Part2
   ML_HW6_part2_<學號>.ipynb
   ML_HW6_part2_<學號>_test.py
3. 資料夾 Part3
   ML_HW6_part3_<學號>.ipynb
   ML_HW6_part3_<學號>_test.py
4. 資料夾 Part4
   ML_HW6_part4_<學號>.ipynb
   ML_HW6_part4_<學號>_test.py

除此之外，你可以存放你訓練完權重的檔案（在 sklearn 下，可以使用 joblib 或是 pickle，如果使用 numpy，則可以保存為 npy），以便到時候執行你的模型。

批改作業時不會執行. ipynb 的檔案，也不會重新訓練模型，這個檔案內容須包含模型產生的方法（對資料的預處理、模型參數的調整等）， …_test.py 則為實際用來測試模型的檔案。

模型保存的參考方式
from sklearn.externals import joblib

```
### 保存模型
joblib.dump(model, 'model.pkl')
### 讀取模型
model = joblib.load('model.pkl')
```

批改方式

批改作業時將執行

1. ML_HW6_part1_<學號>_test.py
2. ML_HW6_part2_<學號>_test.py
3. ML_HW6_part3_<學號>_test.py
4. ML_HW6_part4_<學號>_test.py

測試資料會放在前一層的目錄，請使用相對路徑讀取檔案，測試檔案名稱分別為 hw6_part1_test.csv, hw6_part2_test.csv, hw6_part3_test.csv，預測完後請分別產生一個 hw6_part<?>_<學號>_pred.csv 的檔案在前一層的目錄，以便進行答案的評分。

所以目錄會像這樣

ML_HW6 _M06304XXXX

part1

…

ML_HW6_part1_M06304XXXX_test.py

part2

…

ML_HW6_part2_M06304XXXX_test.py

part3

…

ML_HW6_part3_M06304XXXX_test.py

part4

…

ML_HW6_part4_M06304XXXX_test.py

hw6_part1_test.csv

hw6_part2_test.csv

hw6_part3_test.csv

hw6_part4_test.csv

hw6_part1_pred_ M06304XXXX.csv

hw6_part2_pred_ M06304XXXX.csv

hw6_part3_pred_ M06304XXXX.csv

hw6_part4_pred_ M06304XXXX.csv

黑色的部分為你建立的資料夾及撰寫的程式

紅色的部分是我們會加入的測試檔案，藍色的部分是經由 …_test.py 產生的檔案，此檔案將作為評分的依據，需要分別對每筆…_test.csv 的資料進行預測，產生結果，每筆資料的預測結果以換行隔開。

附檔中有提供一個測試用的…_test.csv，可以自行確定能否正常讀入，同時也附上預期輸出的…pred_....csv 的檔案，可以確認格式是否正確。

注意事項

請勿使用網路上額外找到的資料進行訓練