Machine Learning HW2

曾以諾  
Institute of Data Science  
National Cheng Kung UniversityTainan, ROC  
re6111032@gs.ncku.edu.tw

*Abstract*—*機器學習HW2，更詳細的說明請見連結中hw2.ipynb*

Github link: [MachineLearning-HW2](https://github.com/butterfly2012010/MachineLearning-HW2)

Keywords—Naive Bayes classifier, random forest classifier, XGBoost, Catboost, LightBGM

# Classification Task

1. Naive Bayes Classifier
   1. 說明：fit()時計算每個class的prior probability、每一個類別中的每一個feature的prior probability，predict()時計算posterior probability
   2. 遇到的困難：(method 1)刻完之後，訓練需要48分鐘，預測效果也不理想，

Accuracy = 0.9350571769926609

Confusion =

[[10957 0]

[ 761 0]]

(method 2)也沒有很理想，在交叉驗證的時候，出現了keyerror，有點小問題需要修正。

1. Random Forest Classifier
   1. 說明：用entropy決定要用什麼特徵來區分data，以及用data的purity、是否達到樹的最大深度、以及節點的最小樣本數來決定樹節點要不要繼續長下去。
   2. 遇到的困難：預測時結果不理想，而訓練時有時候樹會是”0”，樹長不出來。
2. Random Forest Classifier (scikit-learn)
   1. 說明：使用了scikit-learn的RandomForestClassifier，也做了交叉驗證
3. XGBoost, Catboost, LightBGM
   1. 說明：使用了現有的package來做XGBoost、Catboost和LightBGM，也做了交叉驗證，其實發現結果和我自己做的一樣爛(LightGBM的混淆矩陣和手刻的天真貝氏和隨機森林)

# References

1. Vamc-stash, Naïve-Bayes,

<https://github.com/vamc-stash/Naive-Bayes/blob/master/src/naive_bayes.py>,

1. SebastianMantey, Random-Forest-from-Scratch, <https://github.com/SebastianMantey/Random-Forest-from-Scratch>.

[3] XGBoost, <https://xgboost.readthedocs.io/en/stable/>

[4] Catboost, <https://catboost.ai/>

[5] LightGBM, <https://lightgbm.readthedocs.io/en/v3.3.2/>