電子商務技術作業 8

第一部份：Python

載入「income.csv」(預測薪水是否能超過 50K)，請將資料切成訓練集與測試集(random_state=15, test_size=0.3)並做必要的前處理。

1. 建立 KNN 模型，印出模型對訓練集、測試集的 Accuracy。(5%)
2. (承上題) 當限制 feature 數量為 n 時 (n = [1, 14])，請為每一個 n 利用 Wrapper feature selection 技巧尋找最適合此模型的 feature set。 搜尋策略請使用 forward selection，並以 5 Cross Validation 後的 Accuracy 為選擇的依據。(30%)
3. (承上題) 請以折線圖呈現模型在不同 feature 數量下的最佳表現。(X 軸：feature 數量，Y 軸 Accuracy)。(5%)
4. 請印出最適合此模型的 feature set。(包含 feature 數量及名稱) (5%)
5. 請使用挑選出的最佳 features 重新訓練模型 (10%)，並比較挑選前與挑選後的模型表現。(10%)

第二部份：Weka

載入資料集 income.csv (預測薪水是否能超過 50K)。

1. 針對 RandomForest 尋找最適合此模型的 feature set，並以 OOB 或 5 CV、Accuracy 為選擇的依據。請將過程與輸出結果截圖到作業中。(20%)
2. 請寫出挑選的 feature set。(並以 weka 輸出佐證) (5%)
3. 請使用挑選前與挑選後的 features 分別建立 RandomForest (使用 OOB 或 5 CV) 並做比較(至少寫出 2 點)，並將過程與輸出結果截圖到作業中。(10%)

- 繳交期限：5/26(三) 中午 12:00，檔名為 ECT_HW8_學號
- 請繳交答案卷(.pdf)和程式檔(.ipynb)，pdf 檔內容包含 Python 與 Weka，請註明題號將該小題的程式及執行結果截圖貼上，如題目有詢問額外的問題也請一併寫上。

  上傳至 ee-class 作業區，遲交一天扣該次作業 5%