機器學習實務與應用

Homework #6
Due 2020 Nov 9 11:00PM

(一) 針對員工離職率(left)進行預測

資料檔案：*hr-analytics.csv*

作業要求：

1. 讀入資料，並判斷出那些數據格式不是數字，或是有缺失值。
2. 將非數字類型的資料進行必要的編碼。
3. 若有缺失值請填補。
4. 將資料切割成訓練集 70%，預測集 30%。
5. 利用 SVM 及 Decision tree 模型進行預測。
6. 請比較 Decision Tree 與 SVM 及 Logistic Regression 的預測準確率。請探討那個模型比較適合，其可能原因為何?

(二) 針對所提供的鐵達尼號 873 名乘客的資料建立死亡預測之模型。

資料檔案：*titanic-train.csv*、*titanic-test.csv*

作業要求：

1. 請利用 SVM、RandomForest 及 XGBoost 三種機器學習方法建立對於 *titanic-train.csv* 內 873 位乘客的死亡預測模型。並比較這三種方法的效能。
2. 每筆資料的欄位說明如下，總共有 14 個欄位。其中所要預測代表存活的欄位為 *survival*。
   - *pclass*: Passenger class (1 = 1st; 2 = 2nd; 3 = 3rd)
   - *survival*: A Boolean indicating whether the passenger survived or not (0 = No; 1 = Yes); this is our target
   - *name*: A field rich in information as it contains title and family names
   - *sex*: male/female
   - *age*: Age, asignificant portion of values aremissing
   - *sibsp*: Number of siblings/spouses aboard
   - *parch*: Number of parents/children aboard
   - *ticket*: Ticket number.
   - *fare*: Passenger fare (British Pound).
   - *cabin*: Does the location of the cabin influence chances of survival?
   - *embarked*: Port of embarkation (C = *Cherbourg*; Q = *Queenstown*; S = *Southampton*)
   - *boat*: Lifeboat, many missing values
   - *body*: Body Identification Number
   - *home.dest*: Home/destination

3. 用以預測存活與否的欄位，除了代表存活與否的 *survival* 欄位當然不可採用，其餘可以自行決定要如何採用。
4. 對非數字類型的資料如何編碼、缺失值如何填補也是自行決定。
5. 特徵是否進行轉換，或是增加額外特徵，也請自行決定。
6. 每個模型針對 873 位乘客的存活預測結果與實際結果繪製 confusion map。且計算每個模型預測結果之 F1 score。
7. 綜合在 *titanic-train.csv* 的表現，請最終選出一個機器學習的方法，針對附件檔案 *titanic-test.csv* 中 436 位乘客資料，進行存活預測。請於所繳交之 notebook 檔案的最末了一個 cell 撰寫針對測試檔 *titanic-test.csv* 的執行程式，該程式請產出一個結果檔 *titanic-pred.csv*。該 csv 檔請包含 *name* 以及 *survival* 兩個欄位。其中 *survival* 為針對相對應之 *name* 的預測結果。助教會依據預測結果評分。

繳交說明：請繳交 jupyter notebook 之檔案。若有討論部分也利用 jupyter notebook 說明。