機器學習導論

Homework #2
Due 2023 Oct 23 11:00PM

(一) 參與 kaggle 網站上提供的 House Prediction 的競賽 (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/)，該競賽是在預測房價(**SalePrice**)。

資料檔案：*HW2_house-prices.csv*

作業要求：

1. 請依據 kaggle 競賽的機制，於作業繳交期限前完成比賽。
2. 請繳交 JupterNotebook 之檔案，以及比賽結果的分數及排名。
3. Kaggle 上面有提供許多範例程式，鼓勵同學多多閱讀，不限制同學是否採用他們提供的做法。
4. 別外請利用相關方法只挑選五個特徵，針對訓練資料 *HW2_house-prices.csv* 進行模型建構，並提供最後模型的預測結果之指標。並與不限制特徵使用數量時的預測結果比較。

(二)使用 Linear Regression 預測在不同的時間，租借共享單車的人數預測 (count)

資料檔案：*HW2_bike-sharing_train.csv, HW2_bike-sharing_test.csv,*

作業要求：

1. 讀入訓練資料 *HW2_bike-sharing_train.csv*。
2. 對日期欄位進行處理。
3. 利用 regression 推測使用人數。產生準確率的指標。
4. 請利用訓練後的模型預測測試資料 *HW2_ bike-sharing _test.csv* 的離職情況，並將結果存成 *HW2_ bike-sharing _test_sol.csv*，儲存格式如下範例。該結果的準確率將佔此一題分數的 ***35%***

	A	B
1	count	
2	32	
3	234	
4	32	
5	32	
6	214	

(三) 針對員工離職率(left)進行離職與否的預測

資料檔案：*HW2_hr-analytics_train.csv, HW2_hr-analytics_test.csv*

作業要求：

1. 讀進訓練資料 *HW2_hr-analytics_train.csv*，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. 建立 Logistic Regression 模型並進行訓練。請呈現訓練後模型預測的混淆矩陣。
5. 請針對各個特徵與離職率的關係進行探討。看是否透過特徵轉換提高預測之準確率。
6. 請利用訓練後的模型預測測試資料 *HW2_hr-analytics_test.csv* 的離職情況，並將結果存成 *HW2_hr-analytics_test_sol.csv*，儲存格式如下範例。該結果的準確率將佔此一題分數的 ***35%***。

	A	B
1	left	
2	1	
3	0	
4	1	
5	1	

(四) 依助教通知方式填寫並繳交參與 AI、機器學習競賽資訊與證明。