機器學習導論

Homework #3
Due 2019 Oct 14 11:00PM

(一)

題目說明：使用 Logistic Regression 預測是否有得肝臟疾病(**Selector**)

資料檔案：*liver.csv*

資料說明：相關欄位說明請參閱下表。其中 Selector 欄位為要預測的目標，1 表示有肝臟疾病，2 表示沒有

欄位	說明	Alcohol(+)
Mean corpuscular volume (MCV)	MCV>100：酒精使用、B12 或葉酸缺乏 MCV<80：地中海貧血、缺鐵性貧血	↑, MCV > 100 (Macrocytic)
Alkaline phosphotase (ALK-P)	消化道疾病、膽道疾病、肝臟疾病、腎病、 骨頭病變、懷孕	↑
Alanine aminotransferase (GPT)	肝臟疾病	↑
Aspartate aminotransferase (GOT)	肝臟疾病、骨骼肌、心肌受損	↑, GOT > 2xGPT
Gamma-glutamyl transpeptidase (rGT)	酒精使用、藥物、膽道疾病	↑
Drinks	飲酒量（個人飲酒習慣）	
Selector	**是否有罹患肝臟疾病**	

作業要求：

1. 讀入資料，將資料切割成訓練集 70%，預測集 30%。分別使用 Logistic regression 及 Bayes 兩種方法來預估。比較那一種模型較佳。
2. 根據臨床經驗中 GPT/GOT 和 GOT/GPT 兩種比值，可作為判斷肝病的參考依據。請找出這兩個比值與預測目標(Selector)的相關係數。並比較單純 GTP 與預測目標，以及 GOT 與預測目標的相關係數大小。
3. 新增前述兩項比值做為新加入特徵，取代原來 GTP 與 GOP 特徵。再重新做一次預測，比較特徵轉換前後的預測準確度

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。

(二) 請上網查詢一個機器學習方法的範例，整理範例其 (a) 資料處理的方法重點，包含特徵處理及模型使用，以及(b)應用的特性。並準備每人三分鐘的投影片（可多人合報）。範例來源之一為 kaggle 的網站。請上網填寫所要報告範例之網址與題目，及報告的時段。先填寫者有較高優先權，但資料得填寫完整。