

機器學習導論

Homework #2

Due 2019 Oct 2 11:00PM

(一) 題目說明：使用 regression 預測房價(SalePrice)。

資料檔案：[house-prices.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編碼。
 3. 填補缺失值。
 4. 將資料切割成訓練集 70%，預測集 30%。分別使用 Linear、Ridge、及 Lasso 三種 regression 模型預測 Rating，並使用 MSE (Mean-Squared-Error) 作為預測準確度的指標。比較那一種模型較佳。
 5. 依據最佳結果的模型，對預測集資料繪製出預測房價 vs 實際房價之散佈(scatter plot)圖
 6. 比較將特徵值進行標準化前處理後之預測準確度
 7. 利用相關係數選取特徵使用：
 - 利用 pandas 套件中 dataframe 之函數 corr()找出各特徵之間的相關係數，並利用 seaborn 套件之 heatmap()函數繪製。
- ```
import seaborn as sns
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True)
```
- 僅使用與房價最相關的前五高係數之特徵進行預測
  - 僅使用與房價最相關的前五低係數之特徵進行預測
  - 比較使用前五高、前五低及所有特徵三種狀況所得到預測準確度的差異
8. 利用 matplotlib 套件繪製特徵 GrLivArea 與房價 SalePrice 之散佈 (scatter plot)圖，判斷是否有極端之 outliers，請將之移除後再比較預測準備度。

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。