



機器學習基礎與演算法

Kaggle 實戰簡介

課程投影片

「版權聲明頁」

本投影片已經獲得作者授權台灣人工智慧學校得以使用於教學用途,如需取得重製權以及公開傳輸權需要透過台灣人工智慧學校取得著作人同意;如果需要修改本投影片著作,則需要取得改作權;另外,如果有需要以光碟或紙本等實體的方式傳播,則需要取得人工智慧學校散佈權。

課程內容

Kaggle 實戰

[實作課程] Kaggle 平台介紹

[實作課程] Kaggle 機器學習實戰

Kaggle 實戰簡介

[實作課程] Kaggle 平台介紹





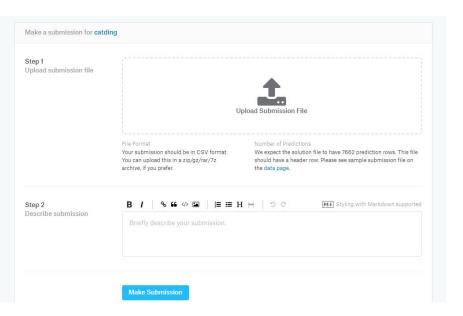
[實作課程] Kaggle 機器學習實戰

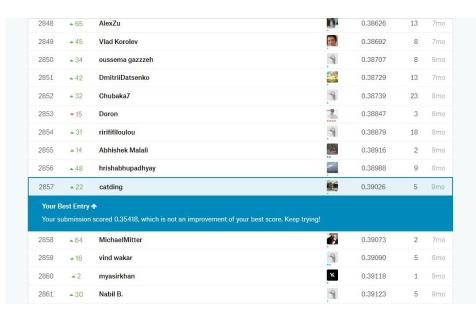
Kaggle 平台簡介

- Kaggle 為一個全球性的資料科學競賽網站, 任何人都可以 上傳數據資料來舉辦比賽, 很多公司會發布一些接近真實業 務的問題, 並提供獎金來吸引愛好數據科學的人來一起解決 問題
- 有些公司甚至會用 Kaggle 上的排名來評估應徵者



上傳預測結果就會看到排名







機器學習 Kaggle 實戰:波士頓房價預測

題目連結: House Prices: Advanced Regression Techniques

- 總共有 76 個 features, features 的說明 請點此
- 每天最多上傳 5 次結果
- 請各位試著完成本題練習,請注意期中考可能包含類似(機器學習類)題型
- 歡迎各位寫作時討論,或參考本課程說明以及參閱 競賽Kernel,建議以能讀懂並能自己修改 Code 為目標



(optional) 完成 scikit-learn-practice 比賽

題目連結: <u>scikit-learn-practice</u>

- 這是單純讓大家熟悉 Scikit-learn 的比賽。總共有一千筆訓練資料、40個 features,簡單的二元分類問題。可在房產預測競賽後作為練習,以提升自己實力
- 資料並沒有提供 features 的意義,純粹是讓大家練習 features scaling、建模、調參數等步驟
- 每天最多上傳 10 次結果
- 請在 private / public leaderboard 上取得 0.8 以上的準確度。達標代表你對 Scikit-learn 的操作有一定水準囉!



資料下載

可至競賽中 Data 頁面下載,或直接使用課程練習題目錄資料

	seconds ago
☐ data-science-london-scikit-learn	2 hours ago
house_prices_advanced_regression_techniques	2 minutes ago

□
test.csv test.csv
☐ train.csv
☐ trainLabels.csv



在開始前...

- 以下約略是整個資料分析競賽的流程
 - 資料清洗與轉換
 - 探索式資料分析 (EDA)
 - 特徴工程 (feautre engineering)
 - 建立模型
 - 調整參數
 - 上傳結果
- 以下替各位做簡單的整理





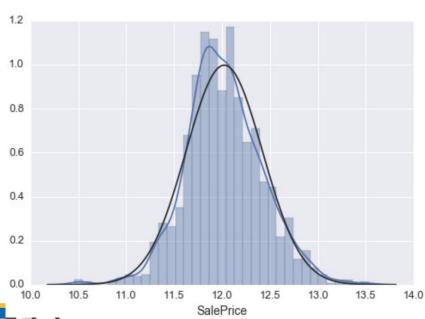
資料清洗與轉換

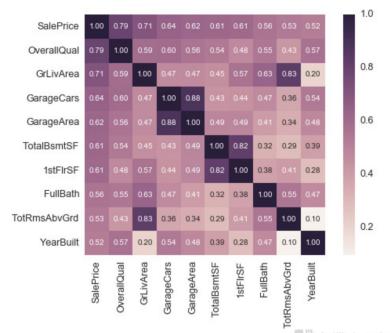
- 注意是否有遺失值 (missing value), 該用什麼方式填補
- 注意是否有 outliers。實務上依照不同 task, outliers 的定義不同, 須自行判斷
- 一些右偏分佈的 feature, 可透過取 log 將其轉為常態分佈
- 類別的變數可透過 one-hot encoding 轉為數字
- 類別太多可試著將相近的類別歸為一類
- 如果會用到一些算法像是 PCA, regression, 要記得將資料 進行標準化 (normalization)



探索式資料分析 (EDA)

● 繪製變數的分佈圖 (hist)、盒鬚圖 (box)、相關係數圖等等







特徵工程 (feature engineering)

 特徵建立是一門藝術, 依靠創意 +經驗 +專家領域知識, 沒有標準答案。能找到最關鍵的 feature, 即使用普通的模型, 也能榜上有名

● 常見的做法:feature 之間的交互關係, 例如相乘、相除、取

log、平方、三次方等等



建立模型

- 每個模型都可以嘗試看看,但建議先從簡單的模型開始!
 e.g. linear regression,並把結果當成 baseline 參考
- 後續的模型如果結果有比 baseline 還差, 就要注意是否參數有問題, 甚至 code 是否寫錯





調整參數

- 強大的模型伴隨著許多參數要調控,但如果一直使用固定的 資料來進行調參,就有可能發生 Overfiithg 的情形
- 善善用 cross-validation + grid search 來尋找最好的參數,再
 使用這個參數,重新訓練你的模型,細節請參考<u>正確調參數</u>

的方式



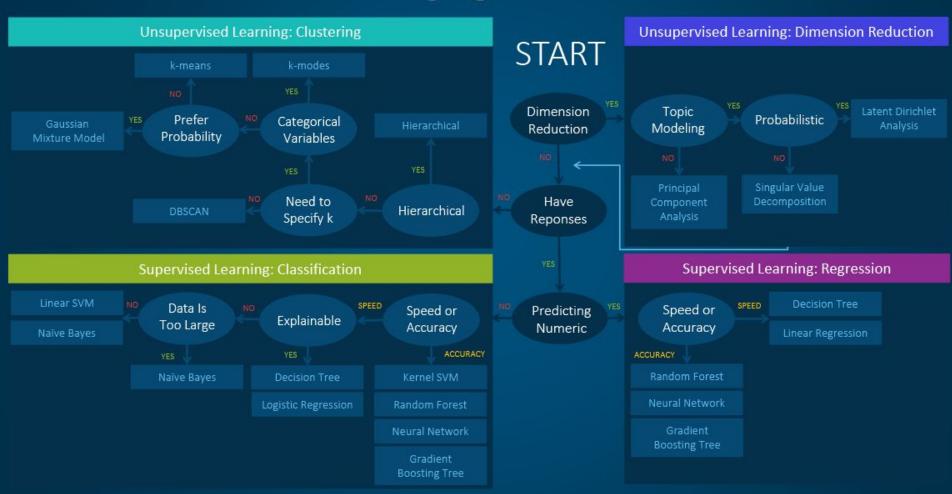


常見問題

- 別人的模型比較好?那是因為別人用了『Ensemble』! 把 RandomForest + XGBoost + GradientBoosting 的結果全 部合併起來, 通常會再提昇一些些 performance
- 若是模型在自己切的 testing data 表現很好, 但是上傳後 publicboard 的分數卻很低, 那很有可能是 Overfitting
- Data 數量較少時,請務必使用 cross-validation 評估結果
- 分類問題如果遇到 data imbalance, 可嘗試使用
 oversampling 或 undersampling 的方式改善, 可參考<u>連結</u>



Machine Learning Algorithms Cheat Sheet



競賽成績繳交注意事項

繳交截止日期 2019/4/02 23:59

Team Name修改為自己的學號

