Day 18 特徵工程

特徵類型

出題教練 陳明佑

知識地圖 特徵工程 特徵類型

CUPOY

機器學習概論 Introduction of Machine Learning

監督式學習
Supervised Learning

前處理 Processing → 探索式數據分析 Exploratory Data Analysis → **特徵工程 Feature Engineering** → 模型選擇 Model selection → 參數調整 Fine-tuning → 集成 Ensemble

非監督式學習
Unsupervised Learning

- 分群 Clustering
- 降維 Dimension Reduction

特徵工程 Feature Engineering

概論

數值型特徵
類別型特徵
時間型特徵

→ 填補缺值 →
- 去離群值 → 去偏態 → 特徵縮放
- 類別型特徵處理
- 時間型特徵處理

→ 特徵組合 / 特徵篩選 → 特徵評估

本日知識點目標

- 複習並思考：資料中常見的有哪些特徵類？
- 上面這些類型：要轉換成對目標的猜測，有哪些要特別注意的地方?

常見特徵類型：數值型特徵 / 類別型特徵

CUPOY

常見特徵有兩大類 (舉Day017的範例)

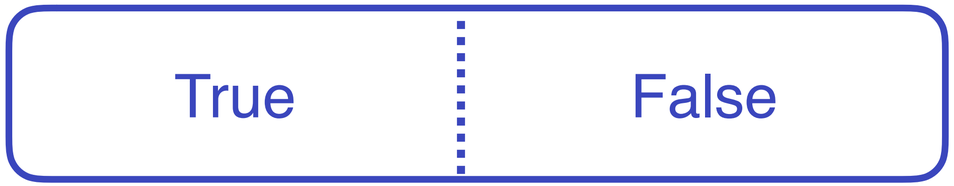
數值型	類別型	類別型	數值型
坪數	行政區	性別	年齡
20坪	新北市	男	50
每坪50點	台北市+500點 其餘+0點	男-2點 女+1點	60以上0點 30~59 +1點 29以下+2點

- 數值型特徵：有不同轉換方式，函數 / 條件式都可以
- 類別型特徵：通常一種類別對應一種分數

其他特徵類型 (1 / 2)

CUPOY

二元特徵



- 只有 True / False 兩種數值的特徵
- 可以當作類別型，也可當作數值型特徵 (True:1 / False: 0)

排序型特徵

- 例如名次 / 百分等級，有大小關係，但並非連續數字
- 通常當作數值型特徵處理，因為當作類別型會失去排序資訊

其他特徵類型 (2 / 2)

CUPOY

時間型特徵

2018/12/15 09:00:00

- 雖然時間型特徵可當作數值型特徵或類別型特徵，但都不適合
  - 取總秒數雖可變為數值，但會失去週期性 (ex 月 / 星期)
  - 使用本身可以當作類別，但會失去排序資訊，類別數量也過大
- 因此時間型特徵我們會個別於 Day 25 的課程中講解

補充說明

因為程式講解需要，會以 cross_val_score 顯示改善效果：分數越高表示效果越好，但不會在現階段講解這部分原理，有興趣提前了解的同學請研讀延伸閱讀內容：k- fold cross validation

重要知識點複習

CUPOY

- 資料中最常見的特徵類型是**數值型特徵**與**類別型特徵**，雖然還有二元特徵、排序型特徵、時間型特徵等多種特徵類型，但仍以前兩者為主
- **數值型特徵**：最容易轉成特徵，但需要注意很多**細節**
- **類別型特徵**：通常一種類別對應一種分數，問題在**如何對應**
- **時間型特徵**：特殊之處在於有**週期性**
- 上述三種特徵，會在之後的課程講述對應的特徵工程

解題時間
It’s Your Turn

請跳出PDF至官網Sample Code＆作業
開始解題