巨匠線上真人

Python 資料科學應用開發

第九堂：資料準備與清理

（Data Preparation and Cleaning）

同學，歡迎你參加本課程

- 請關閉你的FB 、Line等溝通工具，以免影響你上課。
- 考量頻寬、雜音，請預設關閉攝影機、麥克風，若有需要再打開。
- 隨時準備好，老師會呼叫你的名字進行互動，鼓勵用麥克風提問。
- 如果有緊急事情，你必需離開線上教室，請用聊天室私訊給老師，以免老師癡癡呼喚你的名字。
- 軟體安裝請在上課前安裝完成，未完成的同學，請盡快進行安裝。

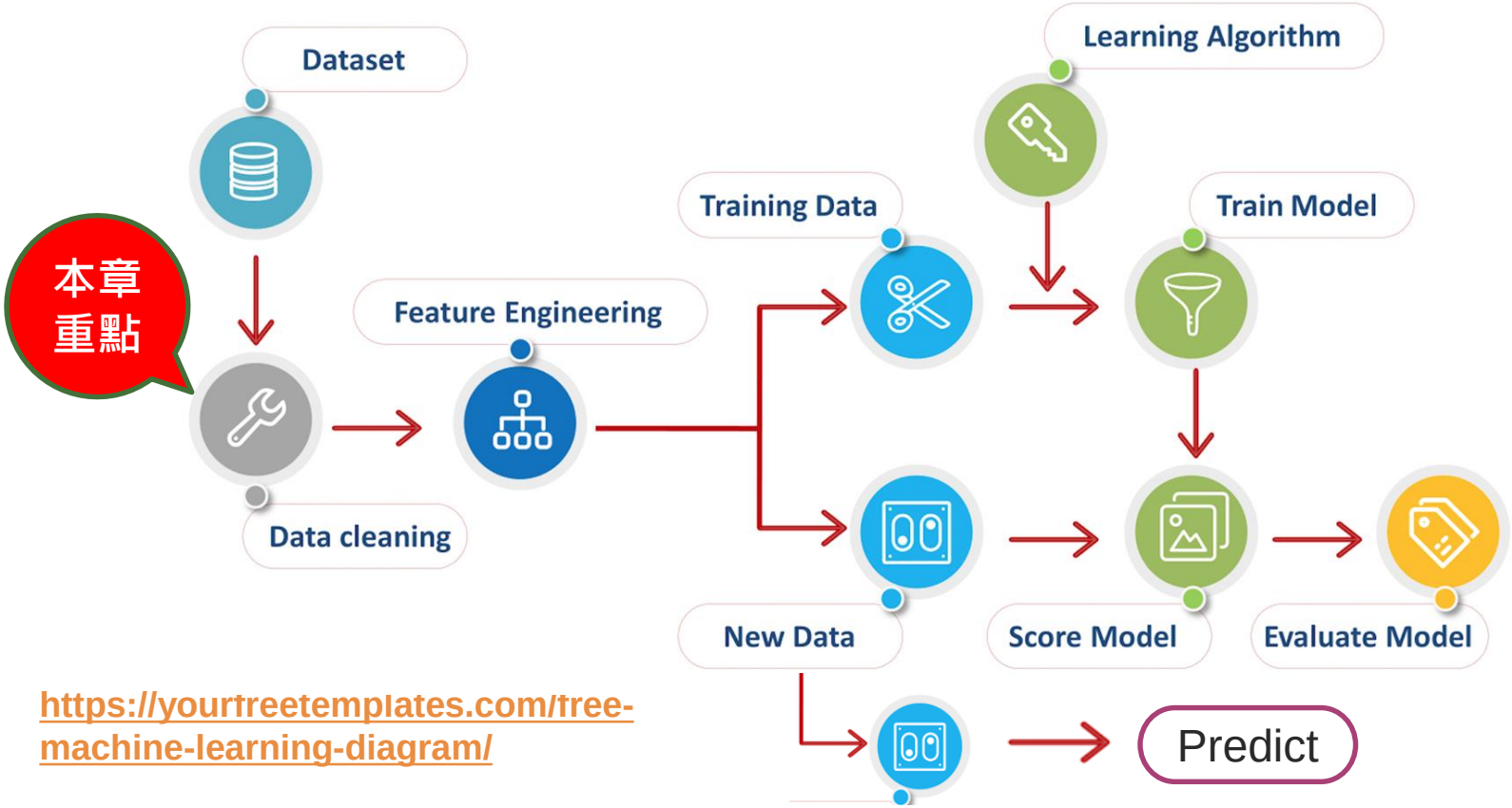
課程檔案下載


巨匠電腦線上真人
開課查詢
免費體驗專區
課程總覽
專業師資
學員專區
講師專區
最新消息
您好!
登出
點數卡產品兌換
APCS檢測專區
公告專區
我的課表
IT真人課程劃位
電腦分校課程劃位
外語真人課程劃
美語分校課程劃
取消劃位
課程檔案下載
上課權益查詢
教學平台測試
學習諮詢
常見問題
個資維護
忘記密碼
登出
課程檔案下載
程式語言好難學?
那是因為
你還沒學過Python!
線上老師LIVE直播教學．搶先看
巨匠電腦真人課程

ZOOM 學員操作說明


www.pcschool.com.tw
您正在觀看蘇世宗的螢幕
查看選項
原始大小
請求遠端控制
共同註記
退出全螢幕
5 查看選項/共同註記/筆（連連看）
滑鼠
文字
筆
箭頭
擦除
格式
撤銷
重做
清除
與會者 (15)
查找參會者
張齡月 (我)
婷婷
解除靜音
舉手
2 共享螢幕（指導演練；點評作品）
老師須先停止共享螢幕
才能請學生共享螢幕
1 聊天
解除靜音
啟動視訊
邀請
與會者
共享螢幕
聊天
錄影
4 解除靜音
3 與會者/舉手

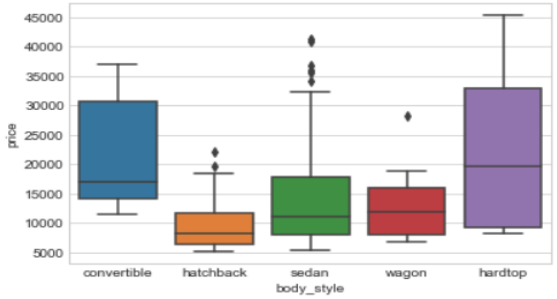
機器學習流程



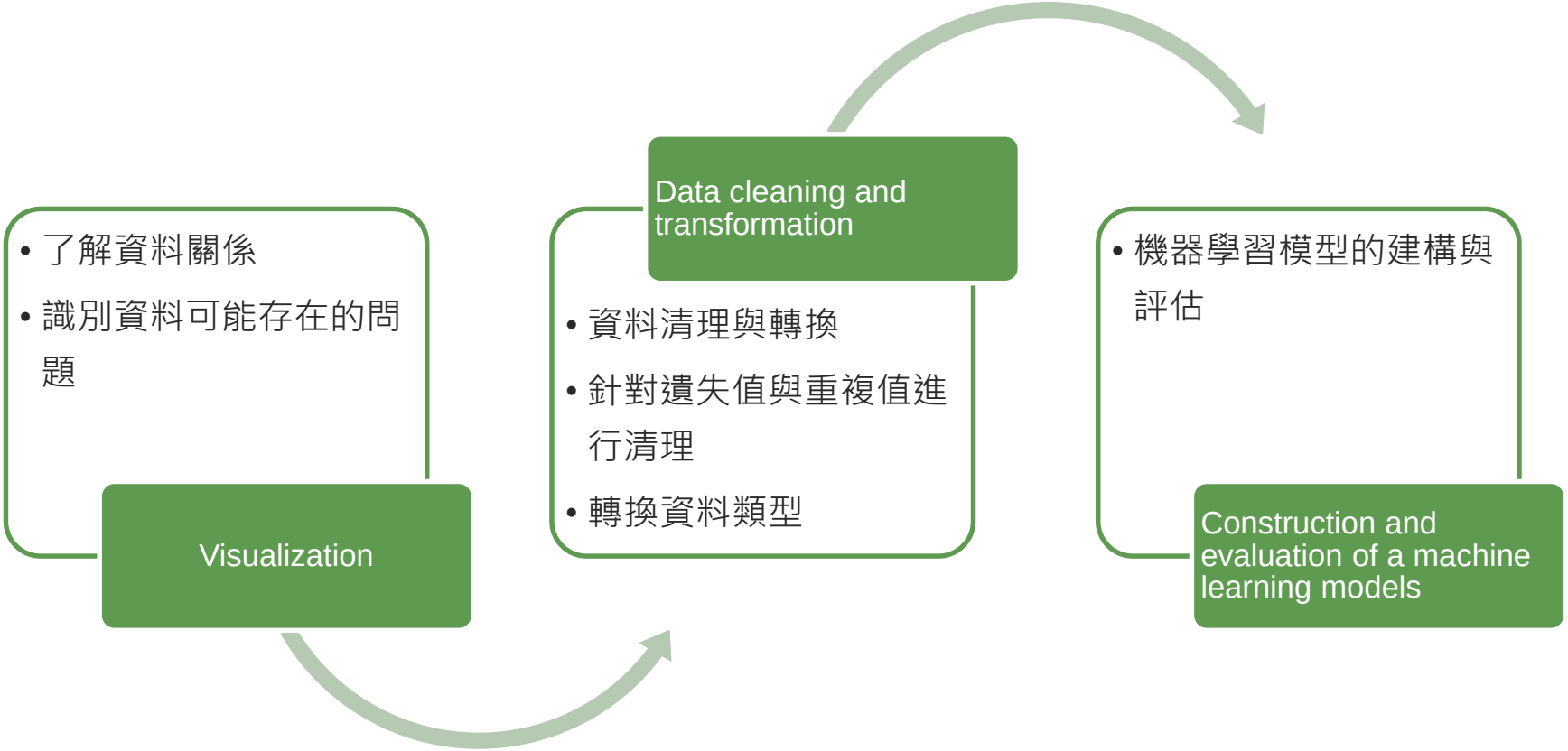
https://yourtreetemplates.com/tree-machine-learning-diagram/

資料準備與清理

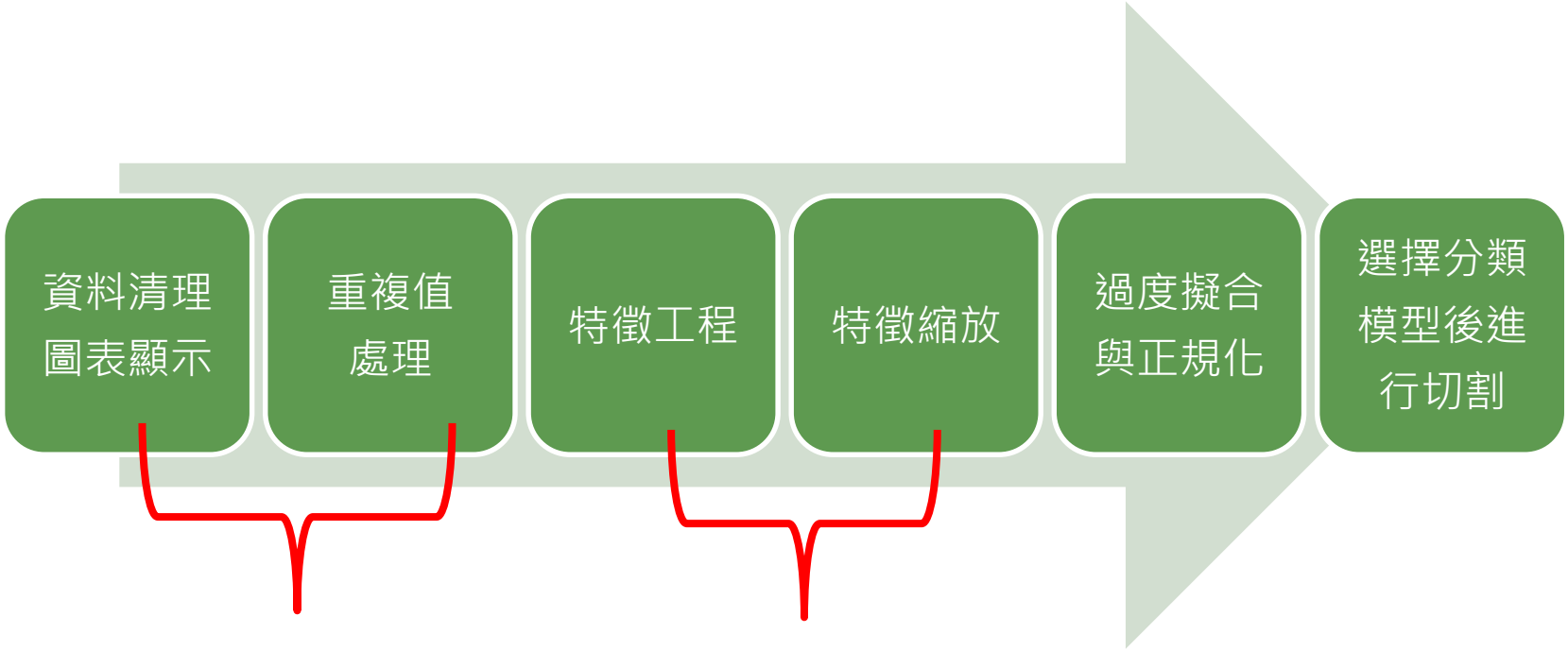
- The process of data preparation is highly interactive and iterative.
- 資料預處理（Pre-processing）
  - 資料探索與分析（Exploratory Data Analysis, EDA）：分析資料集，以了解資料的主要特性，通常採視覺化判斷
  - Data Clean：資料格式統一、代碼編製、遺漏值（Missing Value）、重複記錄處理



資料預處理-1-資料準備 (Data Preparation)



資料預處理-2


資料清理
圖表顯示
重複值
處理
特徵工程
特徵縮放
過度擬合
與正規化
選擇分類
模型後進
行切割
以資料內容為主的處理
以資料標籤為主的處理

資料清理

針對coluimns名稱進行字串的取代

將有問題的資料轉換為遺失值

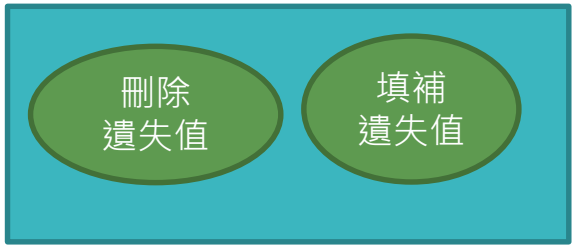
指定欄位資料轉換為數值型態

資料清理

針對coluimns名稱進行字串的取代

將有問題的資料轉換為遺失值

指定欄位資料轉換為數值型態

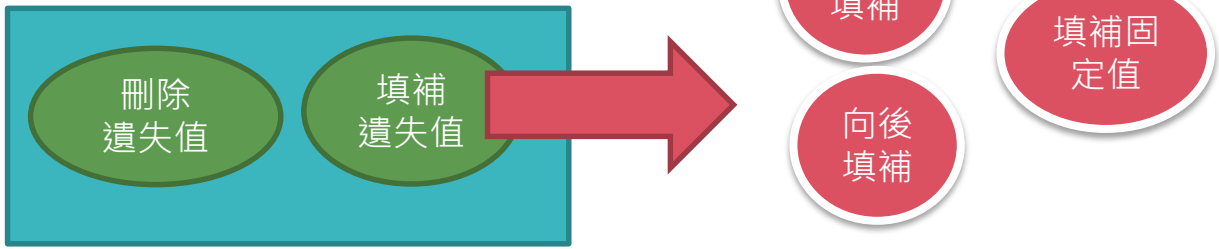


資料清理

針對coluimns名稱進行字串的取代

將有問題的資料轉換為遺失值

指定欄位資料轉換為數值型態



資料清理

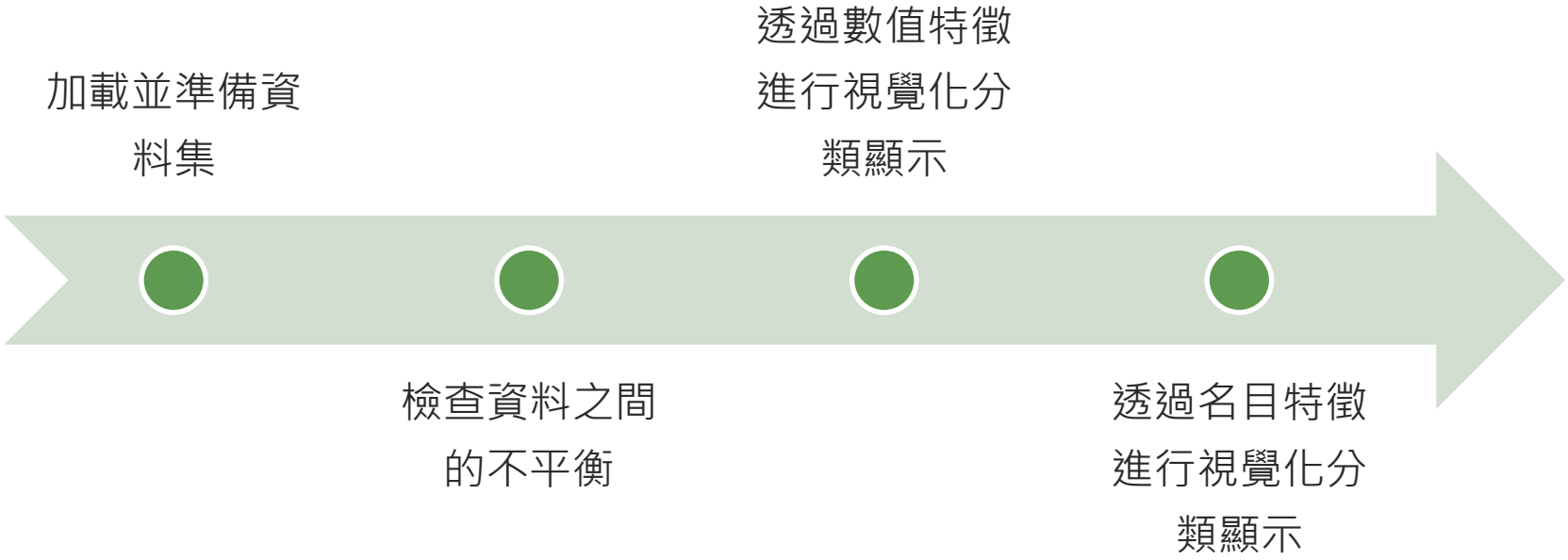
針對coluimns名稱進行字串的取代

將有問題的資料轉換為遺失值

指定欄位資料轉換為數值型態

- Python的map進行映像比對
- 透過Pandas的astype進行轉換
- Sklearn的標籤轉換處理

圖表顯示



案例探討

- 探索 automobile pricing 資料集
- Lab (DAT275x)
  - Module3-275 / DataPreparation.ipynb
- 流程見下頁

資料準備與清理


Import Data
Merge Data Sets
Rebuild Missing Data
Standardise Data
Normalise Data
De-Duplicate
Verify & Enrich
Export Data

處理流程

- 載入資料集
- 重新命名欄位名稱（Recode names）
- 遺漏值（Missing value）處理
- 轉換欄位資料型態（Transform column data type）
- 特徵工程與變數轉換（Feature engineering and transforming variables）
- 數值變數轉換（Transforming numeric variables）
- 移除重複列（Remove duplicate rows）

遺漏值處理

- 觀察遺漏值（Missing Value）個數
  - df.isnull( ).sum( )
  - df.isnull( ).sum( ).plot( )
- 處理遺漏值（Missing Value）
  - df.dropna(axis=0)
- 填補遺漏值
  - 以欄位平均值填補遺漏值
    - from sklearn.preprocessing import Imputer
    - imr = Imputer(missing_values='NaN', strategy='mean', axis=0)
    - imr = imr.fit(df.values)
    - imputed_data = imr.transform(df.values)
    - imputed_data
  - Pandas
    - Pandas Missing Value

Feature engineering and transforming variables

- Feature engineering 特徵工程
  - 大多數情況下，機器學習不是使用原始數據完成。
  - 特徵被轉換或組合以形成更具預測性的新特徵。
- 良好的特徵工程可以使機器學習模型運作更好。

Feature engineering and transforming variables

- Transforming numeric variables
  - 改善它們的分佈特性。
  - 此過程不僅可以應用於功能，還可以應用於回歸問題的標籤。
  - 一些常見的變換包括對數和冪，包括正方形和平方根。

Feature engineering and transforming variables

- Compute new features
  - 來自兩個或更多現有功能。
  - 這些新功能通常稱為交互術語。
  - 當兩個特徵的值的產生比兩個特徵本身更具預測性時，就會發生交互。
  - 考慮購買奢侈品男鞋的機率：
    - 這種可能性取決於作為男性的用戶與富裕的購買者之間的相互作用。

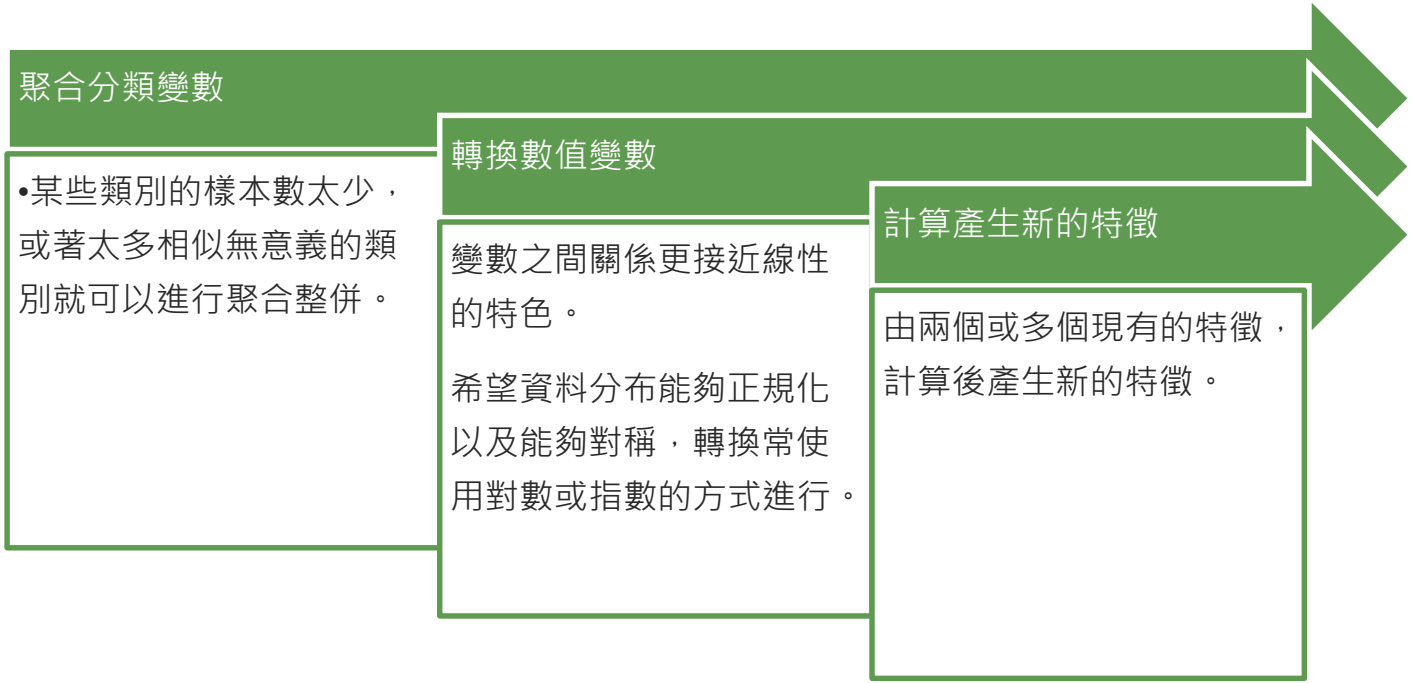
Feature engineering and transforming variables

- Aggregating categories variables
  - 聚合分類變數。
  - 具有太多唯一類別的分類特徵或標籤將限制機器學習模型的預測能力。聚合類別可以在某種程度上改善這種情況。
  - 只聚合在類似的類別才有意義。

轉換欄位資料型態

- 類別特徵（Categorical Feature）
  - 有序（ordinal）：特徵值隱含順序或大小、高低之分
  - 名目（nominal）：不隱含順序或大小、高低之分，例如：T-shirt color
- 處理方式：
  - Python Map
  - Pandas
  - Scikit Learn
- 請參考Transform column data type.ipynb練習

特徵工程


聚合分類變數
•某些類別的樣本數太少，或著太多相似無意義的類別就可以進行聚合整併。
轉換數值變數
變數之間關係更接近線性的特色。
希望資料分布能夠正規化以及能夠對稱，轉換常使用對數或指數的方式進行。
計算產生新的特徵
由兩個或多個現有的特徵，計算後產生新的特徵。

獨熱編碼（One Hot Encoding）

- 針對名目（nominal）特徵將每一類別轉為個別的啞變數（Dummy variable）
- 處理方式： Transform column data type.ipynb

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

獨熱編碼（One Hot Encoding）

- 於新版本執行 preprocessing.OneHotEncoder( ) 語法將出現以下警告訊息：

  FutureWarning: The handling of integer data will change in version 0.22. Currently, the categories are determined based on the range [0, max(values)], while in the future they will be determined based on the unique values.
  If you want the future behaviour and silence this warning, you can specify "categories='auto'".In case you used a LabelEncoder before this OneHotEncoder to convert the categories to integers, then you can now use the OneHotEncoder directly.warnings.warn(msg, FutureWarning)

- 處理方式：preprocessing.OneHotEncoder(categories='auto')

特徵縮放（Feature Scaling）

- 規模一致化
  - 常態化（Normalization）
  - 標準化（Standardization）

常態化（Normalization）

- from sklearn.preprocessing import MinMaxScaler
- mms = MinMaxScaler()
- X_train_norm = mms.fit_transform(X_train)
- X_test_norm = mms.transform(X_test)

$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{\min}}{x_{\max} - x_{\min}}$$

標準化（Standardization）

- from sklearn.preprocessing import StandardScaler
- stdsc = StandardScaler()
- X_train_std = stdsc.fit_transform(X_train)
- X_test_std = stdsc.transform(X_test)

$$x_{std}^{(i)} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

其他欄位轉換

- 格式統一
  - 日期（YYYY-MM-DD）
  - 代碼（F/M or 0/1）
- 級距化
  - np.linspace(2.0, 3.0, num=5)

    array([2. ， 2.25， 2.5 ， 2.75， 3.])
  - np.arange(3)

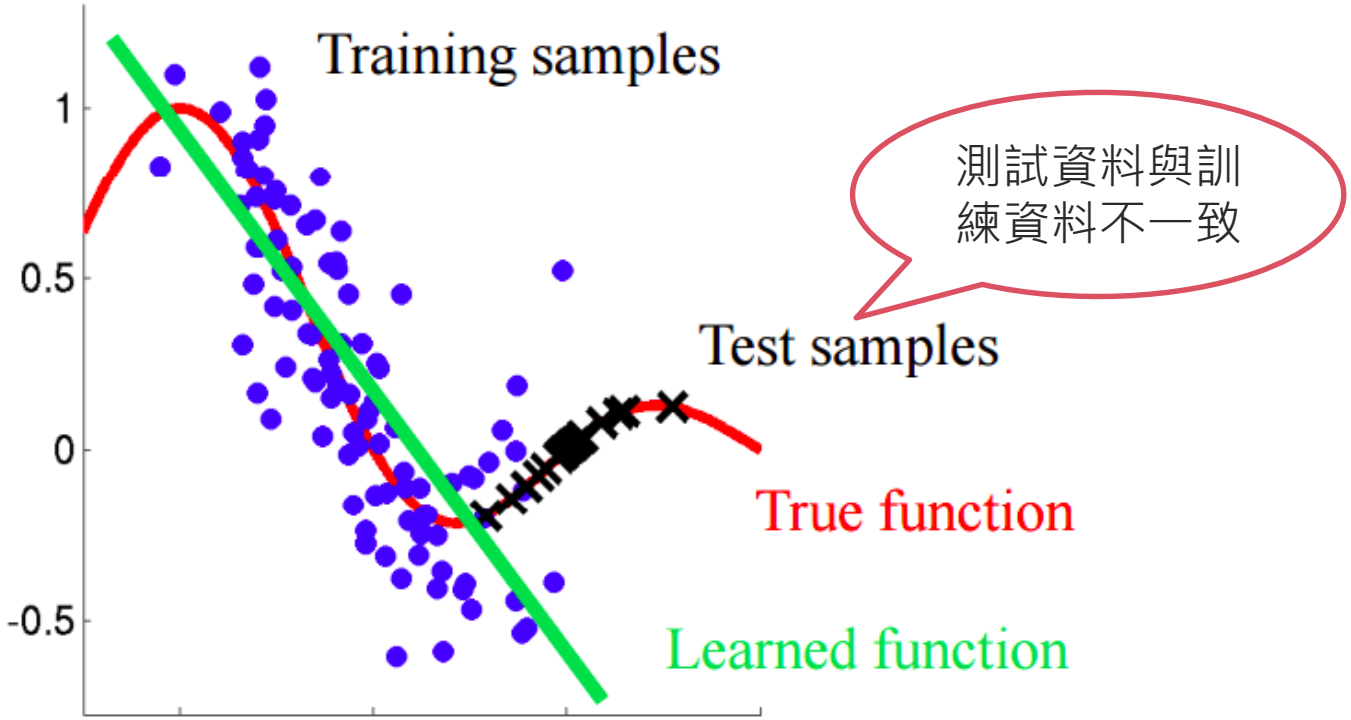
    array([0, 1, 2])
  - np.bincount(np.array([0, 1, 1, 3, 2, 1, 7]))

    array([1, 3, 1, 1, 0, 0, 0, 1])

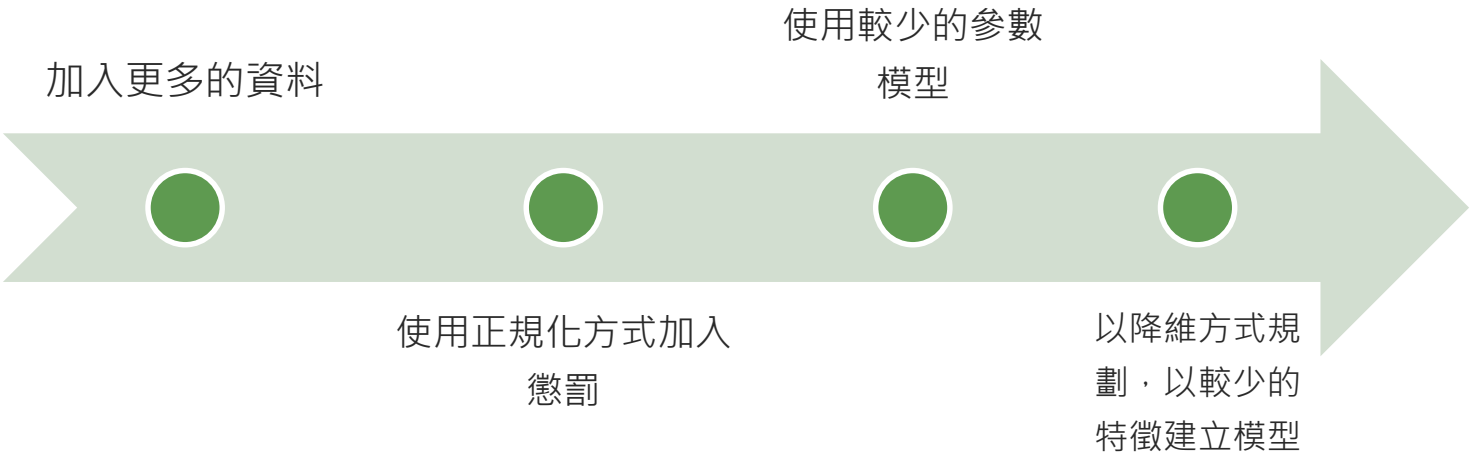
選擇有意義的特徵

- 過度擬合（Overfitting）：訓練時準確率很高，但處理新數據時，效果不佳，這種模型雖然具有低偏差性（Low Bias），但具有高變異性（High Variance）
- 為避免模型過度擬合，可採以下措施：
  - 收集更多資料
  - 經由『正規化』手段，對複雜模型引進懲罰項（Penalty）
  - 使用較少的參數，建立較簡單的模型，例如：決策樹使用較少的層數
  - 使用降維，以較少的特徵建立模型

過度擬合（Overfitting）

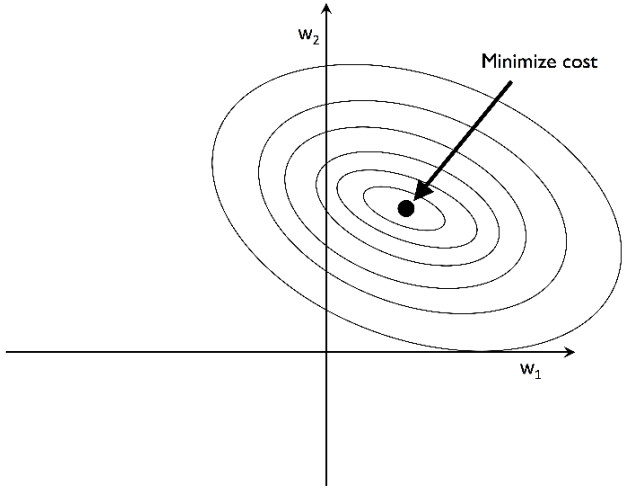


避免過度擬合



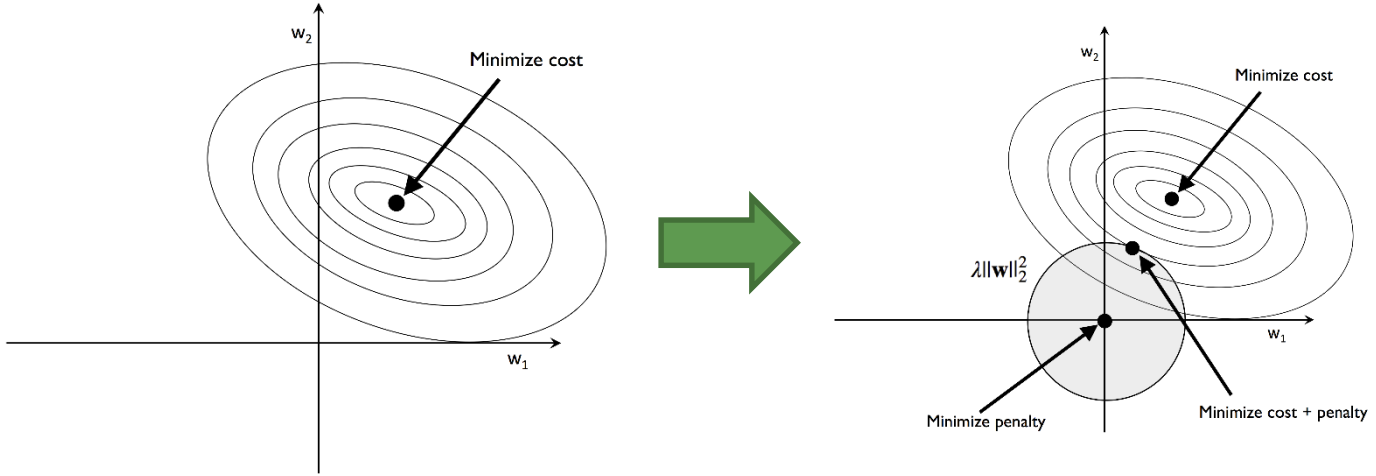
正規化（Regularization）

- 對成本函數加一懲罰項（Penalty）
- 分為兩種：
  - **L2** $L2: \|\boldsymbol{w}\|_2^2 = \sum_{j=1}^{m} w_j^2$
  - **L1** $L1: \|\boldsymbol{w}\|_1 = \sum_{j=1}^{m} |w_j|$



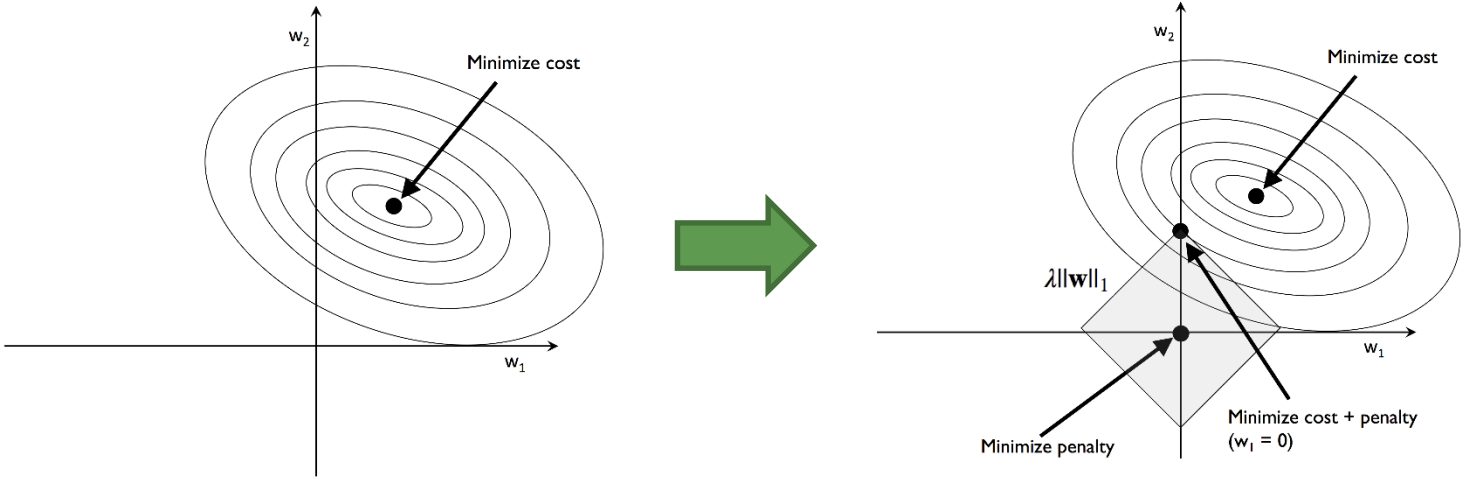
L2 Regularization

- 正規化參數 λ 加大，可增加『正規化強度』



L1 Regularization

- L1 會導致稀疏解（sparse solution），求得最佳解的權重大部分為0。
- 用於高維的資料含有很多不適當的特徵。
- 適合降維的 Feature Selection。



評估模型性能 (Evaluate model performance)

- 在訓練模型後請評估性能。
  - MSE (Mean squared error)：可稱為均方誤差或平均平方誤差。
  - RMSE (Root mean squared error) ：可稱為均方根誤差。
  - MAE (Mean absolute error) ：可稱為平均絕對誤差。
  - Median absolute error：中位絕對誤差。

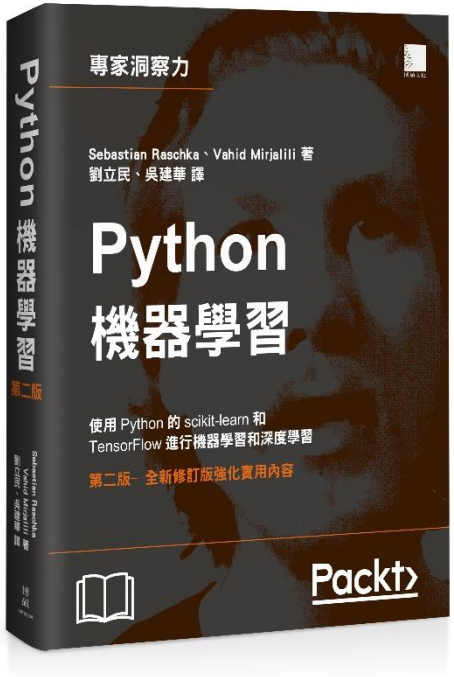
評估模型效能

- Lab (275x)：IntroductionToRegression
  - Module4-275 / IntroductionToRegression.ipynb
- 流程
  - Execute a first linear regression
    - Split the dataset
    - Scale numeric features
  - Train the regression model
  - Evaluate model performance

評估模型效能

- Lab (275x)：ApplyingLinearRegression
  - Module4-275 / ApplyingLinearRegression.ipynb
- 流程
  - Load the dataset
  - Prepare the model matrix
    - Create dummy variables from categorical features
    - Add the numeric features
  - Split the dataset
    - Rescale numeric features
  - Construct the linear regression model
  - Evaluate the model

參考用書



- 書名：Python機器學習（第二版）

  http://www.drmaster.com.tw/bookinfo.asp?BookID=MP11804
- 作者：Sebastian Raschka, Vahid Mirjalili ISBN
- 譯者：劉立民、吳建華
- 出版社：博碩

問卷

http://www.pcschoolonline.com.tw

www.pcschoolonline.com.tw