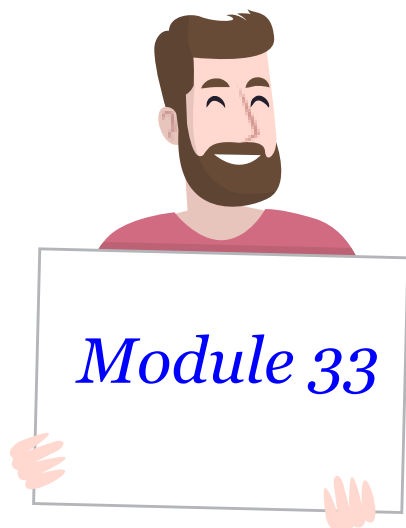




實務上會遇到的問題



designed by  freepik

Estimated time:
45 min.



資訊工業策進會 Institute for Information Industry

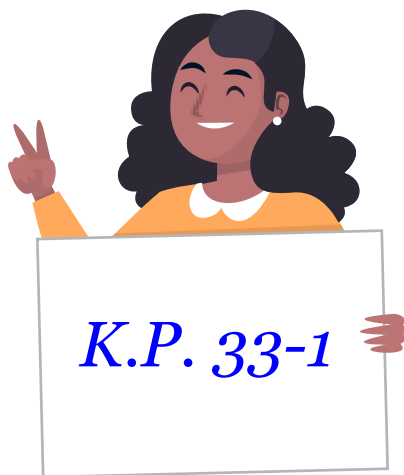
學習目標

- 33-1:類別不平衡
- 33-2:資料標準化
- 33-3:資料科學競賽平台介紹



33-1: 類別不平衡

- 類別不平衡介紹
- 類別不平衡做法
- 類別不平衡準確度問題



designed by freepik

類別不平衡介紹

- 類別不平衡指的是當我們要做分類問題的時候，某些類別的資料特別少
 - 這個在實務上很常發生
 - 通常會建議分類問題，每個類別資料量數量級要一樣

類別不平衡介紹

- 針對類別不平衡，最根本的解決辦法就是蒐集更多少數類別的資料
 - 但這通常需要更多時間、更多成本、更多勞力、更多感應器等



更多時間



更多成本



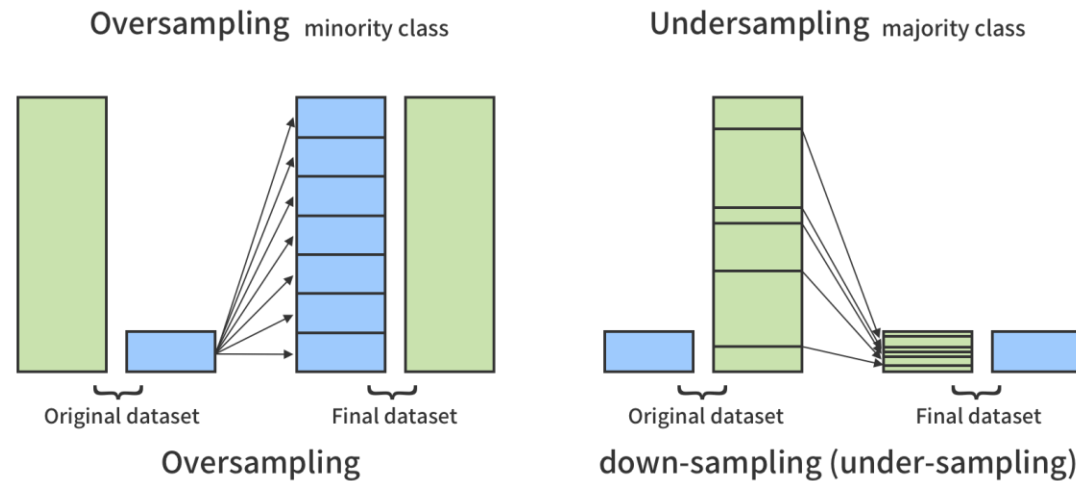
更多努力



更多感應器

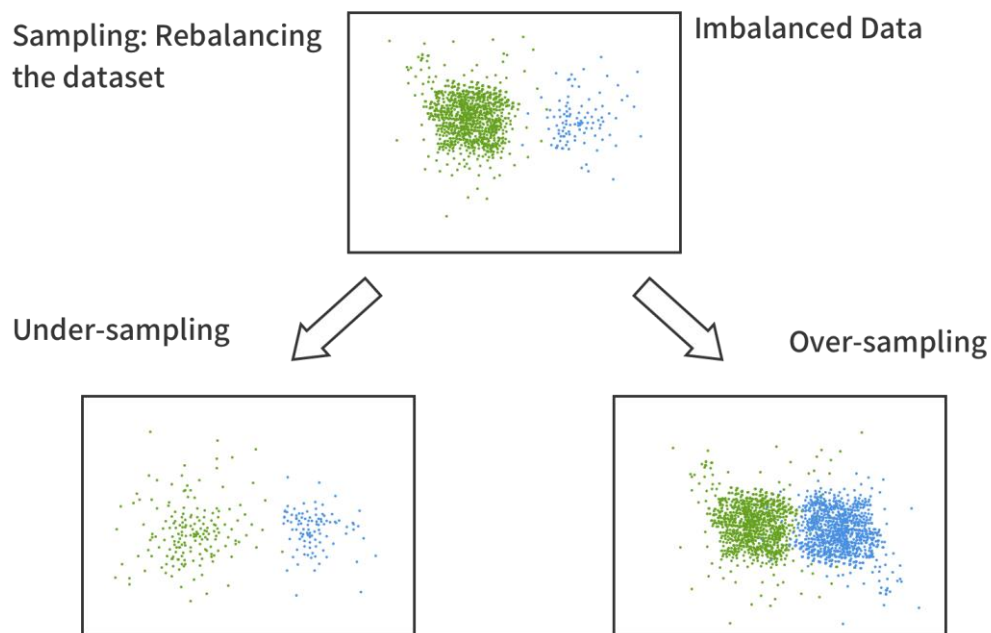
類別不平衡介紹

- 針對類別不平衡，如果沒有辦法在蒐集資料，可以使用resample方法
 - Resample分成over-sampling及down-sampling
 - over-sampling是把少數類別資料複製多次使得其跟多數類別數量級差不多
 - down-sampling表示把多數類別資料砍少讓它們跟少數類別數量級差不多



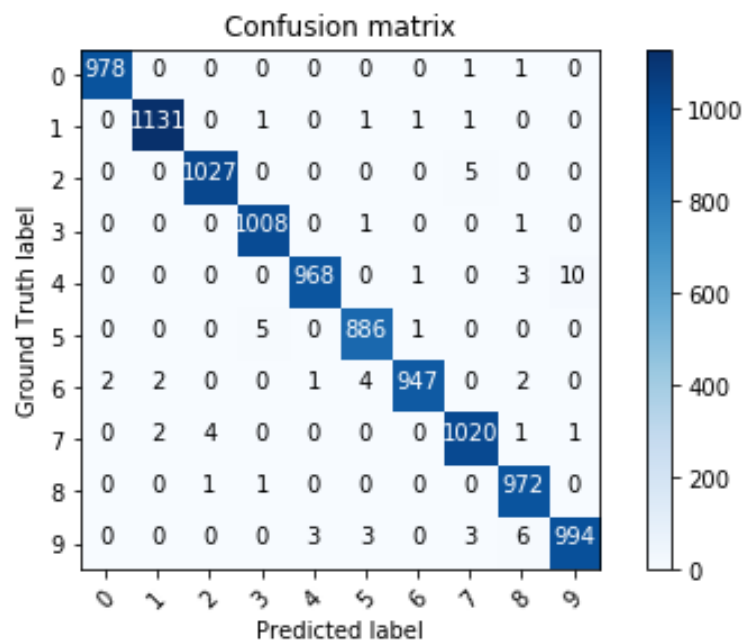
類別不平衡介紹

- 以下是over-sampling以及down-sampling的示意圖
 - 藍色是多數類別的點，紅色是少數類別的點



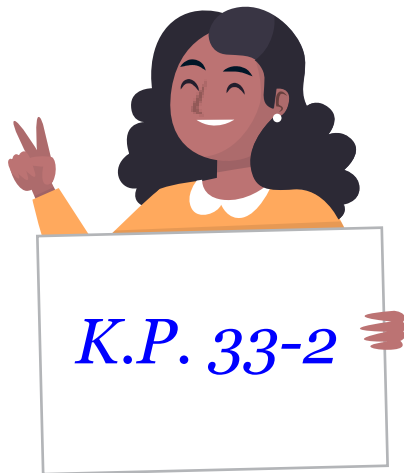
類別不平衡準確度問題

- 處理類別不平衡的問題時，建議不要只單衡量分類的準確度
 - 假設有100筆資料，99筆是多數類別，1筆是少數類別，如果機器偷懶的話會一直猜是多數類別，造成容易誤會準確度99%
 - 建議多參考幾個模型衡量指標，如混淆矩陣



33-2: 資料標準化

- 資料標準化
- **Min-Max Normalization**
- **Z-score Normalization**
- 資料標準化注意事項



designed by freepik

資料標準化

- 資料標準化的目的就是要將資料特徵壓縮到某個特定的數值區間附近
- 常見的做法有
 - Min-Max Normalization
 - Z-score Normalization

Min-Max Normalization

- **Min-Max Normalization**是一種將資料標準化的方法之一
 - 它的概念是將每筆資料減掉最小值並除以最大值及最小值的差 $\frac{x_i - x_{min}}{x_{max} - x_{min}}$
 - 影像資料資料標準化就是最常見使用**Min-Max Normalization**的方法

254	13	190	20
137	148	10	55
200	48	90	100
60	66	20	10

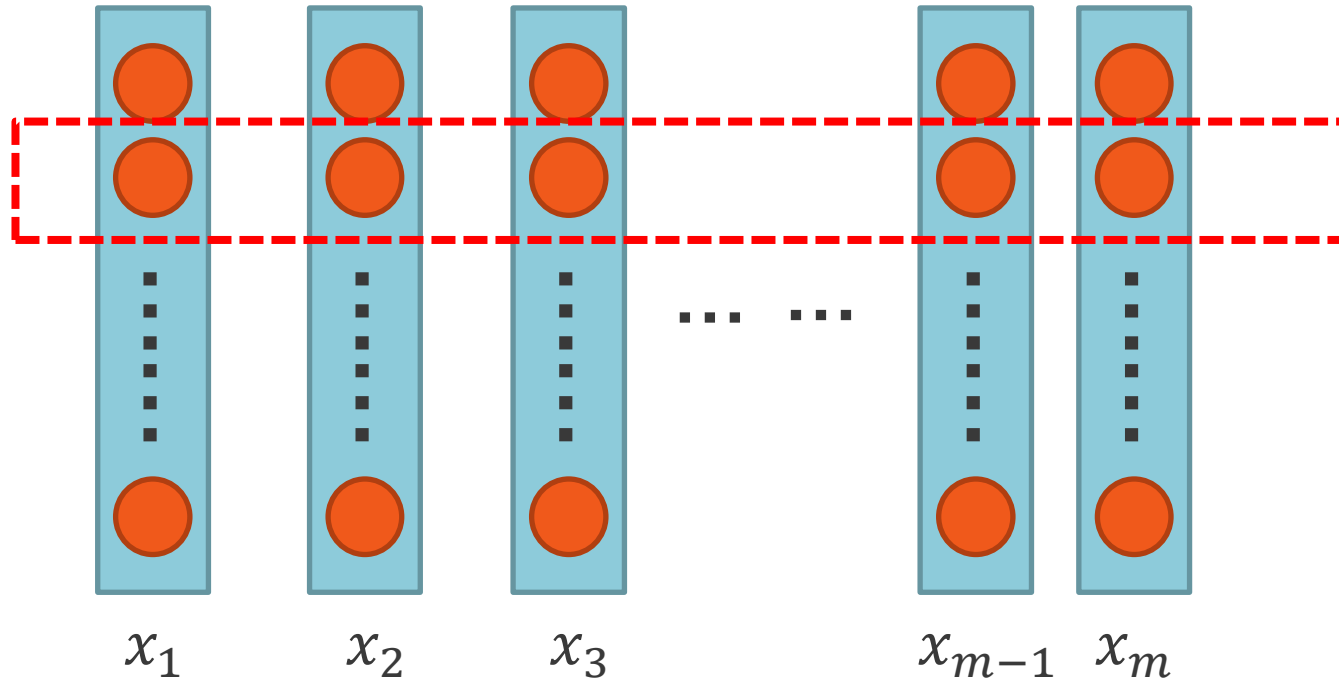
除以255



0.996078	0.05098	0.745098	0.078431
0.537255	0.580392	0.039216	0.215686
0.784314	0.188235	0.352941	0.392157
0.235294	0.258824	0.078431	0.039216

Z-score Normalization

- Z-score Normalization是一種將資料標準化的方法之一
 - 它的概念是將每筆資料減掉整筆資料的平均除以標準差 $\frac{x_i - \mu}{\sigma}$

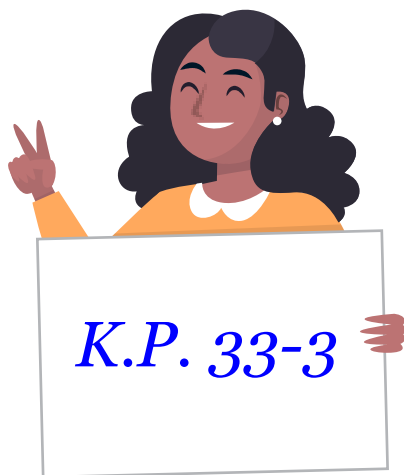


資料標準化注意事項

- 在進行資料標準化的時候，任何統計變量必須使用訓練資料來計算
 - 例如z-score normalization裡的平均以及標準差
- 當測試時，測試資料做資料標準化時，必須使用訓練資料所算出來的統計變量(平均或標準差等.....)

33-3: 資料科學競賽平台介紹

- Kaggle平台介紹
- 天池平台介紹



designed by freepik

Kaggle平台介紹

- **Kaggle是一個資料科學競賽平台**
 - 平時會有很多與資料科學相關的競賽在上面
 - 上面有非常多不同領域的資料集以及各公司想要解決的問題
- **Google於2017年買下Kaggle這個平台**

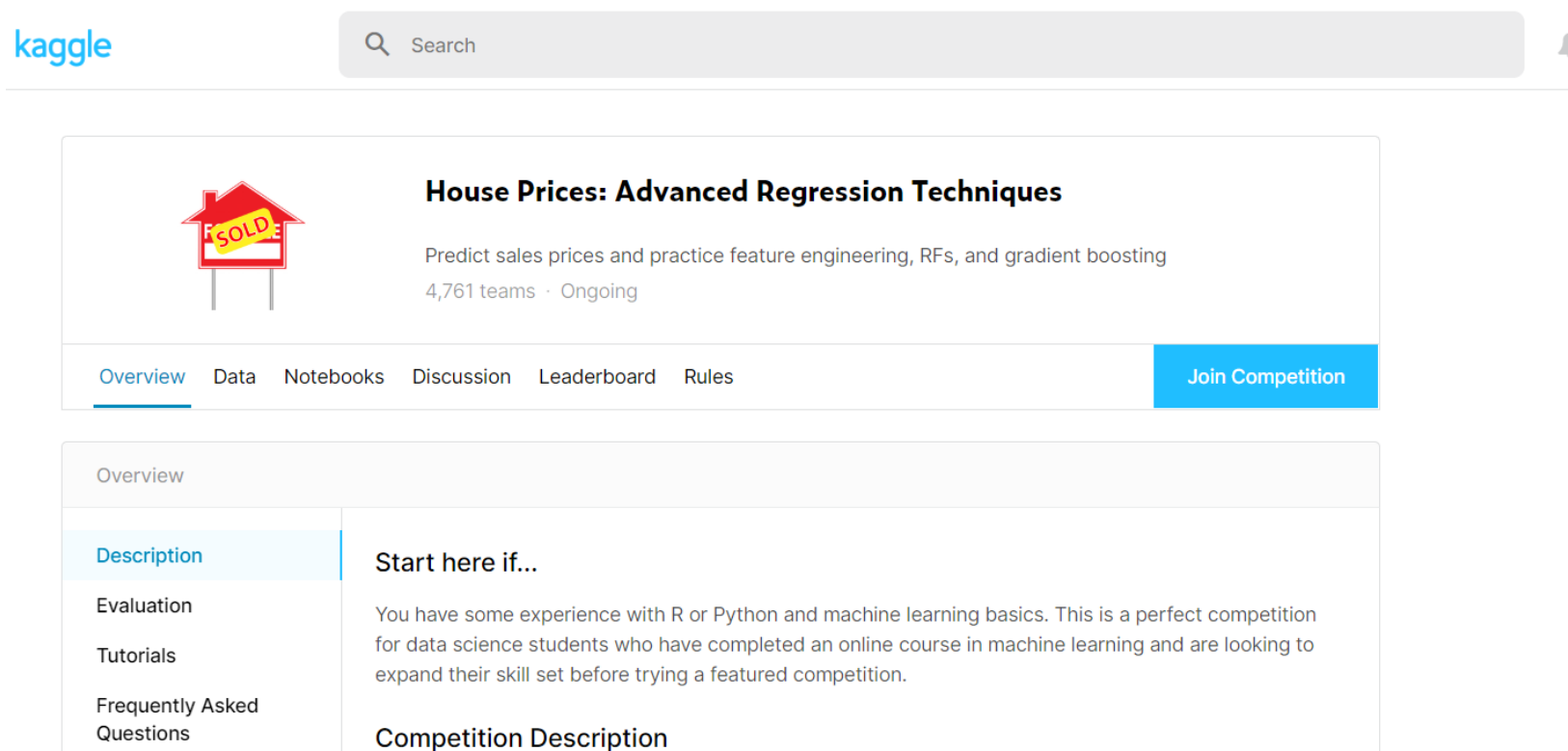
kaggle

Kaggle平台介紹

- Kaggle能看到許多過往比賽的資料集以及觀看許多參賽者如何解決問題
 - 非常適合想要精進自己能力的同學善加利用
 - 同時也能觀察不同領域的企業有哪些問題想要解決

Kaggle平台介紹


- Kaggle針對每個資料集都會有基本描述、資料集下載區域、討論區、即時排名、比賽規則等資訊



The screenshot displays the Kaggle homepage with a search bar and a notification bell. The featured competition is 'House Prices: Advanced Regression Techniques', which includes a 'SOLD' sign icon, a description of predicting sales prices, and statistics showing 4,761 teams and an ongoing status. Navigation tabs for Overview, Data, Notebooks, Discussion, Leaderboard, and Rules are present, along with a 'Join Competition' button. The 'Overview' section is expanded, showing a table of contents with links to Description, Evaluation, Tutorials, and Frequently Asked Questions, and a 'Start here if...' section with a brief introduction to the competition.

kaggle

Search

 **House Prices: Advanced Regression Techniques**

Predict sales prices and practice feature engineering, RFs, and gradient boosting
4,761 teams · Ongoing

[Overview](#) [Data](#) [Notebooks](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#)

Overview

Description	Start here if...
Evaluation	You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.
Tutorials	
Frequently Asked Questions	Competition Description

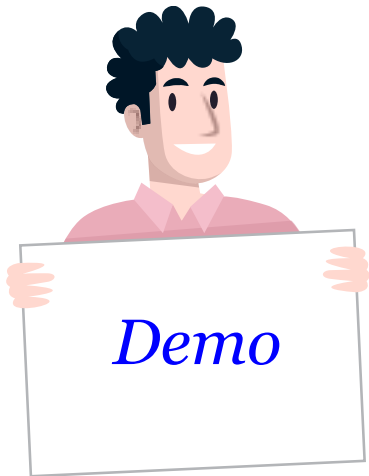
天池平台介紹

- 大陸版本的Kaggle，上面一樣有非常多的資料集以及比賽可以供大家參考
- 天池主要是由阿里巴巴所主持的



Demo 33-3

- 混淆矩陣
- **Kaggle**平台註冊
- **Kaggle**加入比賽



designed by freepik

線上Corelab

- **題目1：TensorFlow使用混淆矩陣**
 - 給予預測結果與實際結果，請用混淆矩陣去評估模型
- **題目2：使用混淆矩陣去衡量DNN神經網路在MNIST上的分類**
 - 給予DNN網路預測結果與資料標籤實際結果，請用混淆矩陣去評估模型
- **題目3：使用混淆矩陣去衡量CNN神經網路在MNIST上的分類**
 - 給予CNN網路預測結果與資料標籤實際結果，請用混淆矩陣去評估模型

本章重點精華回顧

- 類別不平衡
- 資料標準化
- 資料科學競賽平台



Lab: 資料不平衡與Kaggle

- Lab01: 混淆矩陣
- Lab02: Kaggle 平台註冊
- Lab03: Kaggle 加入比賽

Estimated time:

20 minutes

