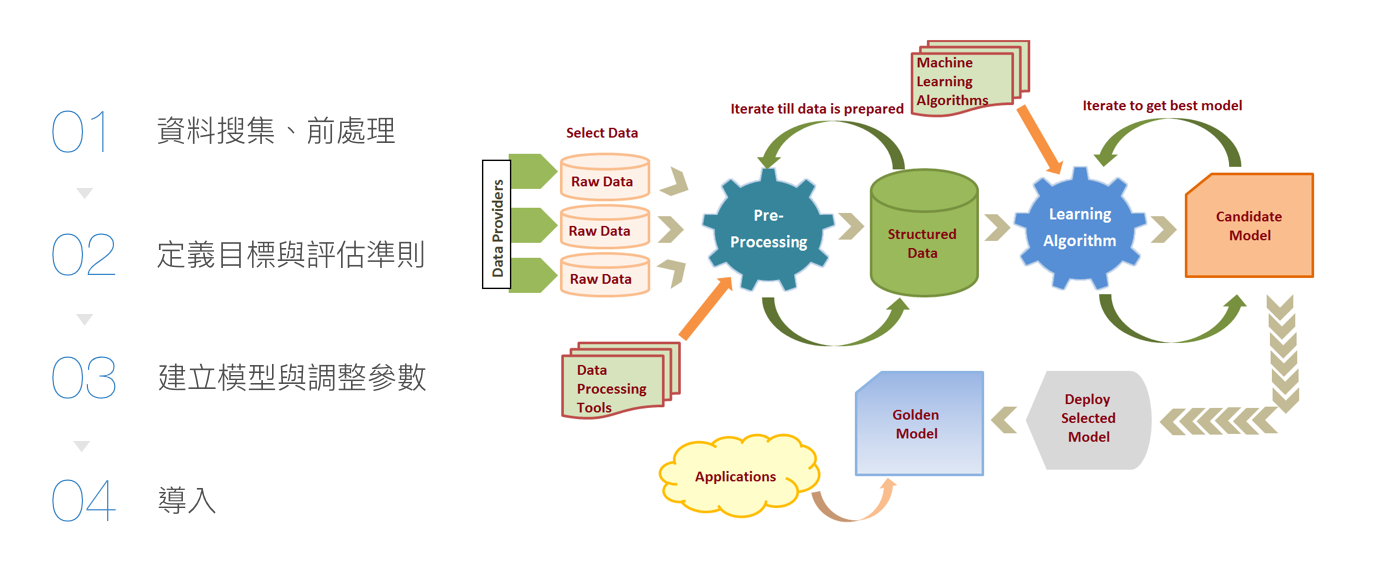
**機器學習監督式學習筆記—林莛翔**

* 流程介紹



圖片來源:機器學習百日

一、資料搜集、前處理

1.資料搜集:

資料:未經過處理的原始記錄

結構化資料:固定欄位、固定格式等數據，如CSV檔

非結構化資料:不規則或無定義之文本，像是圖片、影片、文字

2.資料前處理

離群值處理:數據出現不合規律的極大值或極小值

標準化:提高模型精準度

二、定義目標與評估準則

找出預測目標

分類問題(預測值為類別): 決策樹、隨機森林

回歸問題(預測值為實數): 線性迴歸、多項式迴歸

將資料分訓練集、驗證集和測試集

三、建立模型與調整參數

如回歸模型、樹模型、神經網路模型

根據訓練模型、經驗進行調整參數

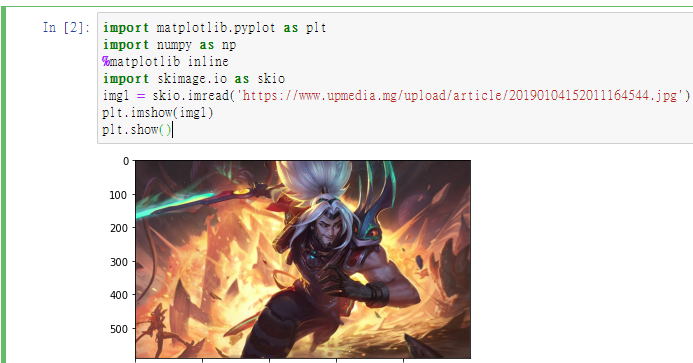
四、導入

進行目標預測，輸出預測結果

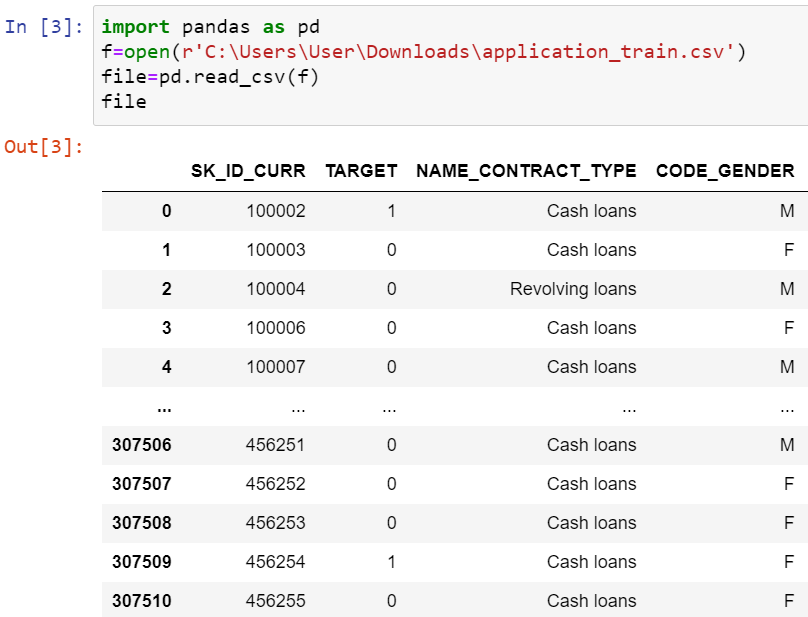
* 完整介紹和程式實作(只列出部分程式碼和**實作超連結(藍色的字**)
* 資料前處理

開啟檔案

開啟圖片: PIL, skimage, open-cv



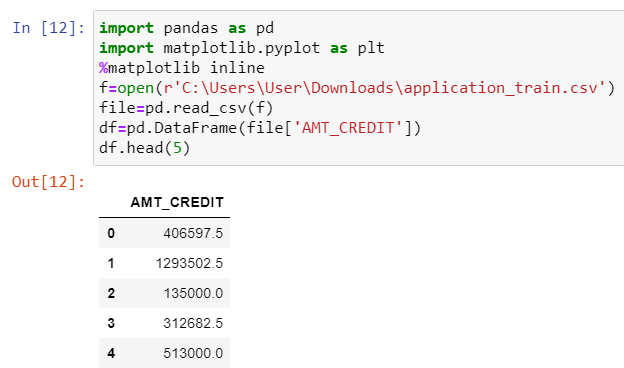
開啟文件:csv,txt檔等



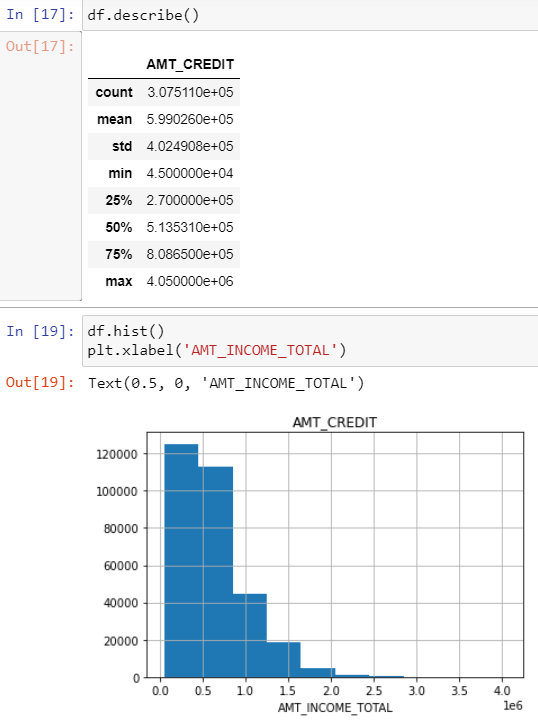
建立dataframe

意義:方便顯示各欄位資料，進行比較和各項數據計算。

取出特定欄位



計算欄位之各項數據和長條圖



也可合併其他欄位和進行其他功能。[欄位合併實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_013_HW%20.ipynb)

常見欄位資料類型:

實數型:實數、排序

類別型object(包含字串)

異常值 (Outliers) [離群值實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_010_HW%20.ipynb) [空缺值實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_012_HW%20.ipynb)

出現原因:手誤、系統性錯誤、未知值隨意填補數字

檢查方法:計算各項數據(標準差、平均數等)、畫圖(直方圖等)

離群值:少數幾筆資料和其他數字相差極大，使預測結果不準

解決方法:捨棄、調整離群值

觀察分數可得知捨棄離群值模型準確度較佳

空缺值:無數字欄位(NA)

解決方法:用平均數、眾數和中位數等填補，還有許多補法要

視資料調整

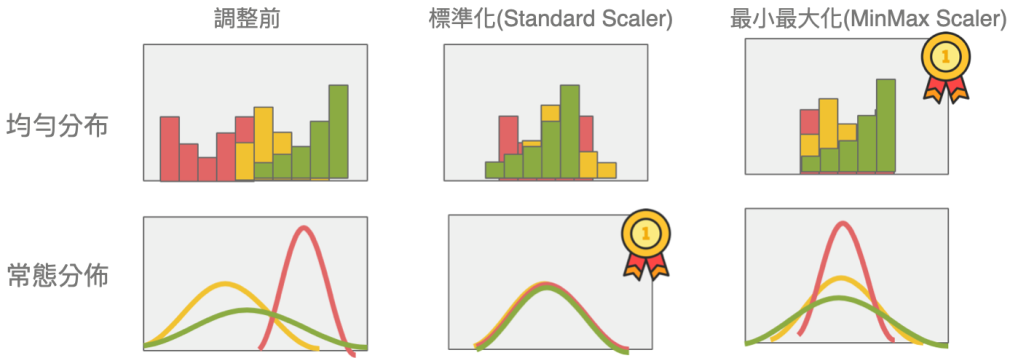
標準化&最大最小化

意義:平衡數值間的影響力

使用場合:非樹狀模型，如線性回歸、羅吉斯回歸

標準化:適用於常態分佈(轉換不易受到極端值影響)

最大最小化:適用於均勻分布(轉換容易受到極端值影響)



圖片來源:機器學習百日

* 探索式數據分析

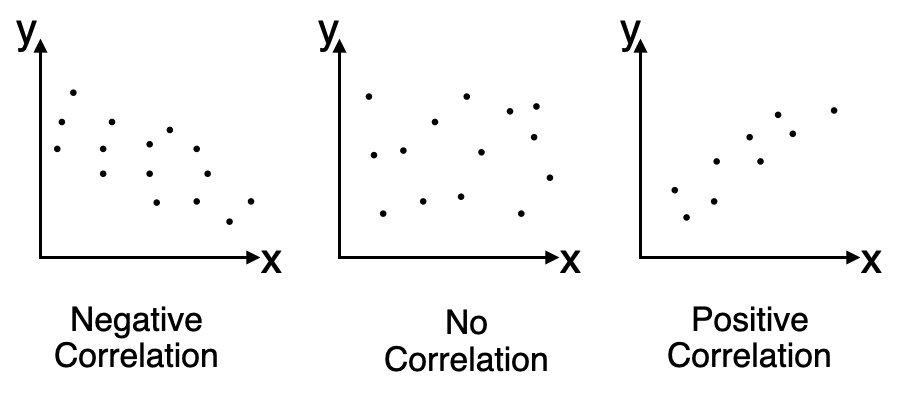
相關係數r[相關係數實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_014_HW%20.ipynb)

意義:了解兩個變數之間的線性關係，範圍介於-1~1之間

相關性強度(假設r為正數):

弱相關r<0.4 中度相關0.4<r<0.6 強相關 r>0.6

正相關和負相關圖



圖片來源:機器學習百日

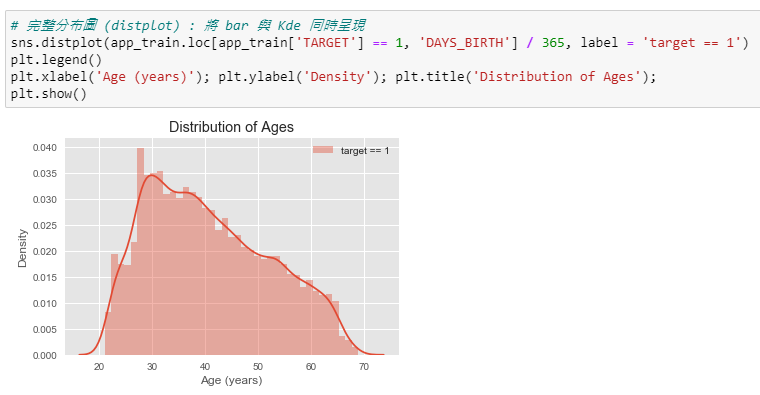
KDE核密度函數

意義:把長條圖換成折線圖，更細分數值間距，使趨勢更容易觀察

優點:無須在意分布(常態分布或均勻分布等)

缺點:計算量龐大

下圖為長條圖和KDE圖同時呈現



離散化 [等頻等寬劃分實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_017_HW%20.ipynb)

意義:讓事情變簡單、減少離群值對分析的影響

方法:等寬劃分：按照相同寬度將資料分群

等頻劃分：將資料分群，每等份資料裡面的個數是一樣的

繪圖排版&常用圖形

Subplot [Subplot實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_019_HW%20.ipynb)

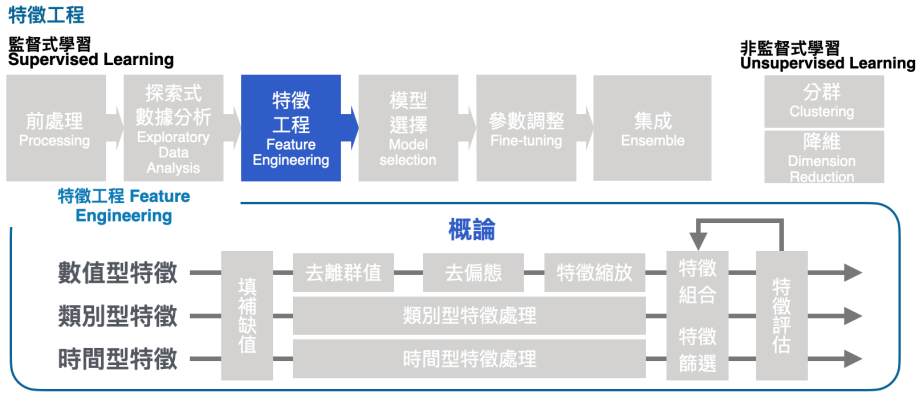
意義:用圖片呈現相似的欄位資訊，方便進行比較

Heatmap [Heatmap實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_020_HW%20.ipynb)

意義:以顏色呈現訊息的強弱和變數的相關性

* 特徵工程

特徵工程簡介

意義:欄位對應到後續評估分數的轉換

圖片來源:機器學習百日

欄位分類

數值型特徵

去偏態 [去偏態實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_023_HW%20.ipynb)

意義:當離群值資料比例太高，或者平均值沒有代表性

時，可使用去偏態調整數據，更接近於常態分佈

方法:

對數去偏:使用對數去除偏態，適用於價格、計數的欄位

方根去偏:將數據減最小值後開根號，適用於成績欄位

分布去偏:功能介於對數去偏和方根去偏之間，隨函數調整

參數可更靈活調整數值。

類別型特徵

編碼 [Label and Onehotencoding實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Label%20and%20OnehotEncoder.ipynb) [均值編碼實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_025_HW%20.ipynb)

意義:字串轉為數值方法

Label encoding:資料有順序的時候使用，如年齡依大小分組

One Hot encoding:資料無順序的時候使用，如國家、姓名

均值編碼:當類別特徵與目標明顯相關時可使用均值編碼

，但容易過擬和

計數編碼:計算類別在資料中的出現次數，當目標與類別

筆數成正/負相關時使用

雜湊編碼:當相異類別數量相當大時，為節省時間使用

雜湊編碼

時間型特徵

編碼 [特徵分解實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_027_HW%20.ipynb)

意義:數值分類

特徵分解:照年、月、日、時、分、秒、星期幾等分類

特徵組合

數值和數值合成 [數值和數值合成實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_028_HW%20.ipynb)

意義:透過各欄位數值進行合成，得到更強力的特徵。如透過

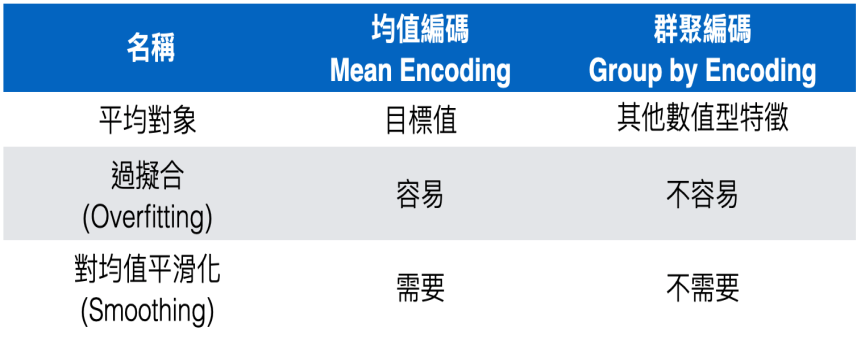
經度和緯度藉由計算點與點之間的距離，得到新特徵。

類別和數值合成 [類別和數值合成實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_029_HW%20.ipynb)

意義:透過各欄位數值和類別進行合成，得到更強力的特徵。

如群聚編碼用各類數值(平均值、中位數等)取代類別欄位

，和數值欄位進行合成，得到新特徵。



圖片來源:機器學習百日

特徵選擇 [相關係數過濾法和Lasso嵌入法實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_030_HW%20.ipynb)

過濾法 意義:選定統計數值與設定門檻，刪除低於門檻的特徵。

如相關係數過濾法。

嵌入法

意義:使用機器學習模型，根據擬合後的係數，刪除係數低於

門檻的特徵。如Lasso嵌入法，GDBT(梯度提升樹)嵌入法

特徵評估 [特徵評估實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_031_HW%20.ipynb)

特徵重要性

意義:是增加重要特徵或刪除不重要特徵的重要判定準則，有

分支次數、特徵覆蓋度、損失函數降低量三種。



圖片來源:機器學習百日

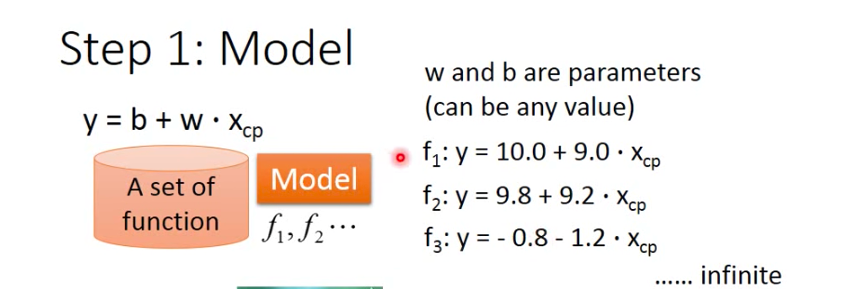
* 模型選擇

概論

機器學習三步驟(以線性回歸舉例)

1.定義好模型

輸入x到模型，透過調整參數，讓所得到的ŷ最接近真實的y值



圖片來源：李宏毅ML Lecture 1: Regression - Case Study

2.評估模型的好壞

觀察損失函數(如MSE)何時出現最小值，可以透過梯度下降找出最佳的模型參數

過擬和

發生情形:在訓練集的損失函數很低，但在實際應用的損失函數超高，導致模型失準。

解決方法: 增加資料量、降低模型複雜度(如五次式變三次式)和使用正規化等。

3.找出讓訓練目標最佳的模型參數

切分資料集 [切分資料集實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/%E5%88%87%E5%88%86%E8%B3%87%E6%96%99%E9%9B%86.ipynb)

意義:避免過擬和，將資料切分為三大類(驗證、測試和訓練集)

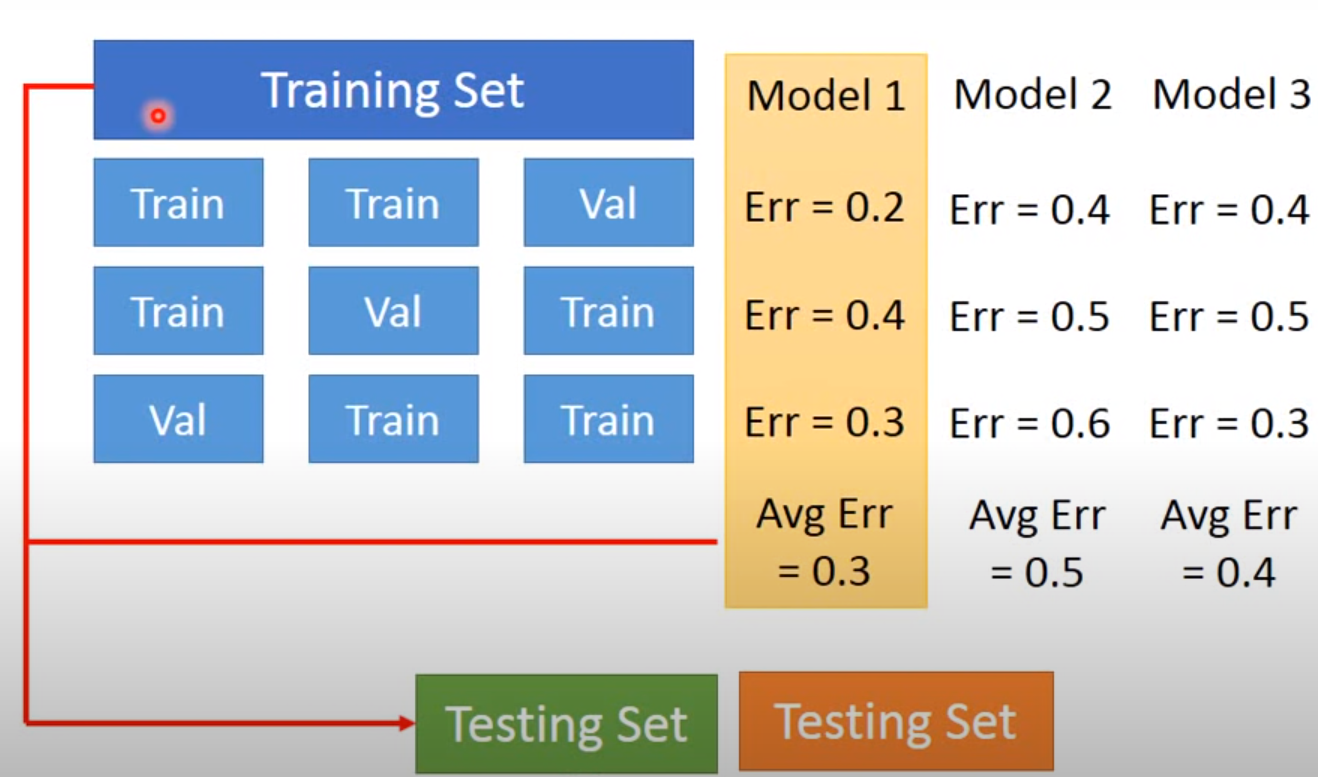
方法: 利用train-test split和K-fold Cross-validation

train-test split:快速將資料進行切分成訓練集和測試集。

K-fold Cross-validation:為了讓結果更穩定，可使用此方法將資料

切分成很多份的訓練和驗證集，再透過 計算與正確答案的誤差，比較每份模型

之誤差平均，選出最小誤差當作模型。



圖片來源：李宏毅ML Lecture 2:Where does the error come from?

預測類型

回歸VS分類

回歸問題:代表預測的目標值為實數 (-∞ 至 ∞)，回歸問題是可

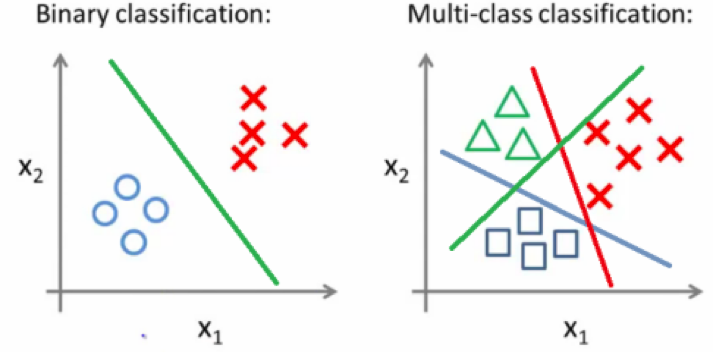
以轉化為分類問題的，如預測成績(實數)，可以改問

分數是否及格(類別)

分類問題:代表預測的目標值為類別 (0 或 1)

二元分類:預測值只有兩個，如分數是否及格(是、否)

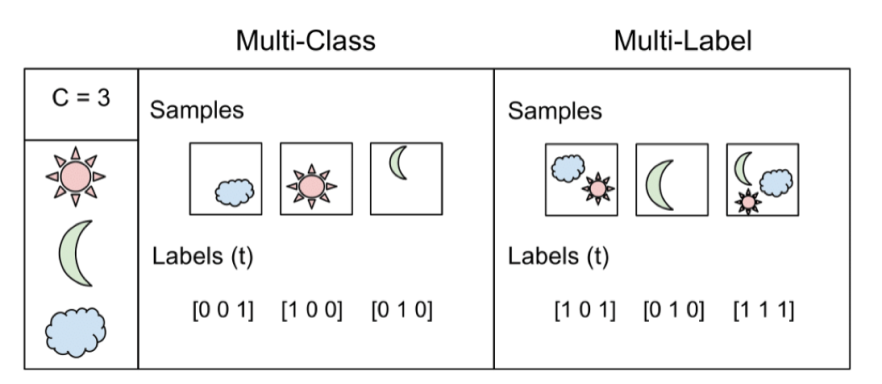
多元分類:預測值兩種以上，如手寫數字辨識(0~9)



圖片來源:機器學習百日

多分類:每個樣本只能歸在一個類別

多標籤:一個樣本可同時歸在多個類別



圖片來源:機器學習百日

評估指標

準確率

意義: 正確分類樣本數/總樣本數，判斷正確率

回歸問題:觀察預測值和正確值之差距，如MAE、MSE

和R-square等

分類問題:觀察預測值和正確值之正確程度，如AUC

F1 – Score等

模型介紹

基礎模型

線性回歸Linear Regression[線性、羅吉斯回歸實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_038_HW%20.ipynb)

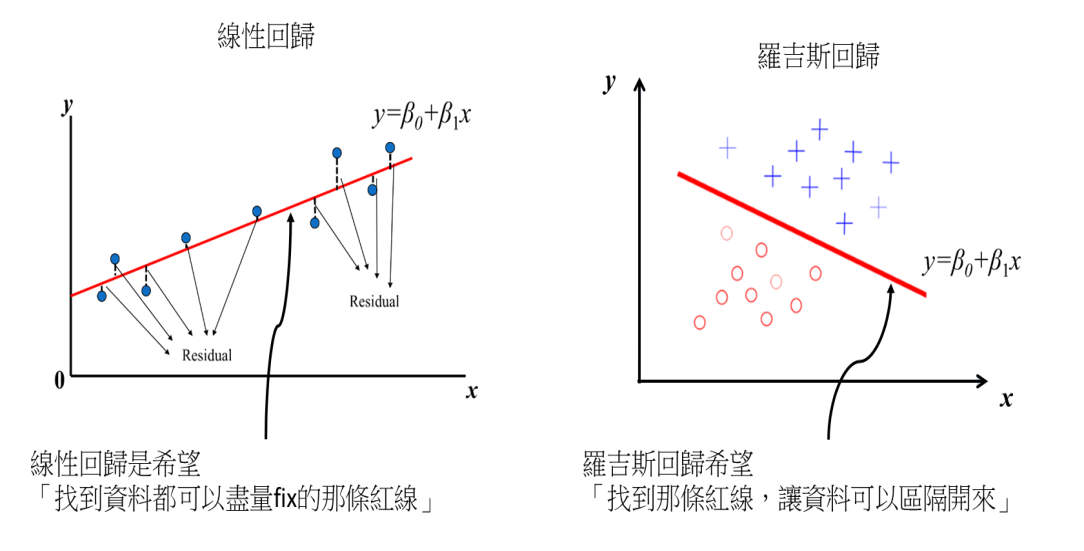
使用於回歸問題，常見、簡單的回歸模型，找出點與點之間的線

性規律進行預測。常利用MSE(平均均方誤差)作為損失函數。

羅吉斯回歸 Logistics Regression

使用於分類問題，線性回歸的分類版，讓資料分成兩部份

target和non-target。



圖片來源: [網址](https://chih-sheng-huang821.medium.com/%E6%A9%9F%E5%99%A8-%E7%B5%B1%E8%A8%88%E5%AD%B8%E7%BF%92-%E7%BE%85%E5%90%89%E6%96%AF%E5%9B%9E%E6%AD%B8-logistic-regression-aff7a830fb5d)

LASSO & Ridge回歸[LASSO&Ridge回歸實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_040_HW%20.ipynb)

為了避免過擬和，會將正則化函數和損失函數做結合，就形成

目標函數=損失函數+正則化，而正則化函數分為兩種L1

和L2，分別和線性回歸結合降低權重(weights)、減少模型

複雜度，組成了LASSO(L1+線性)和Ridge回歸(L2+線性)

L1函數 640?wx_fmt=pngL2函數640?wx_fmt=png

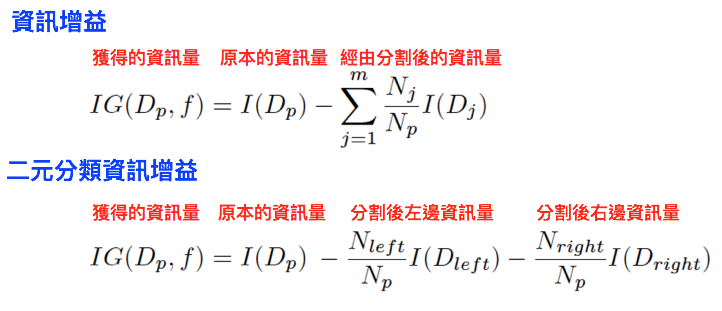
樹狀模型

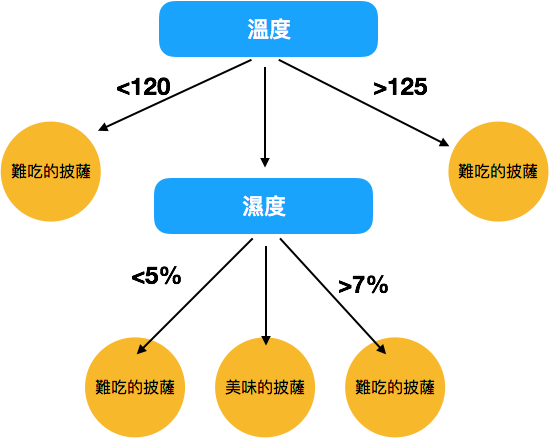
決策樹 [決策樹實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_042_HW%20.ipynb)

決策樹是一個解釋性能力很強的模型，可以讓他人知道模型分類原則，而不是回歸線和數值而已。決策樹透過特徵切分資料讓相 似度高的資料在同一區，得到最大的資訊增益進而找出規律預測目標值，下圖為決策樹模型範例。

資訊增益:獲得的資訊量

資訊量計算方法: 熵(Entropy) 以及 Gini不純度(Gini Impurity)



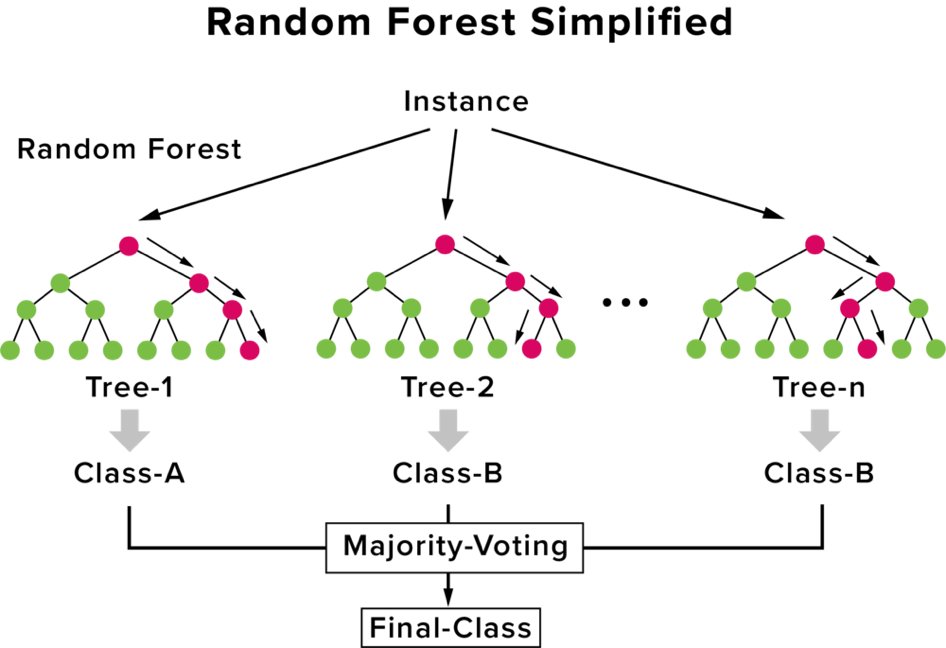


圖片來源: [網址](https://medium.com/jameslearningnote/%E8%B3%87%E6%96%99%E5%88%86%E6%9E%90-%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E7%AC%AC3-5%E8%AC%9B-%E6%B1%BA%E7%AD%96%E6%A8%B9-decision-tree-%E4%BB%A5%E5%8F%8A%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-%E4%BB%8B%E7%B4%B9-7079b0ddfbda)

隨機森林 [隨機森林實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_044_HW%20.ipynb)

若不對決策樹進行參數調整，決策樹很容易過擬和，隨機森林

為決策樹的延伸，隨機使用一部份資料和特徵生成樹，再將多棵 樹組合，透過加權投票的方式得到模型。

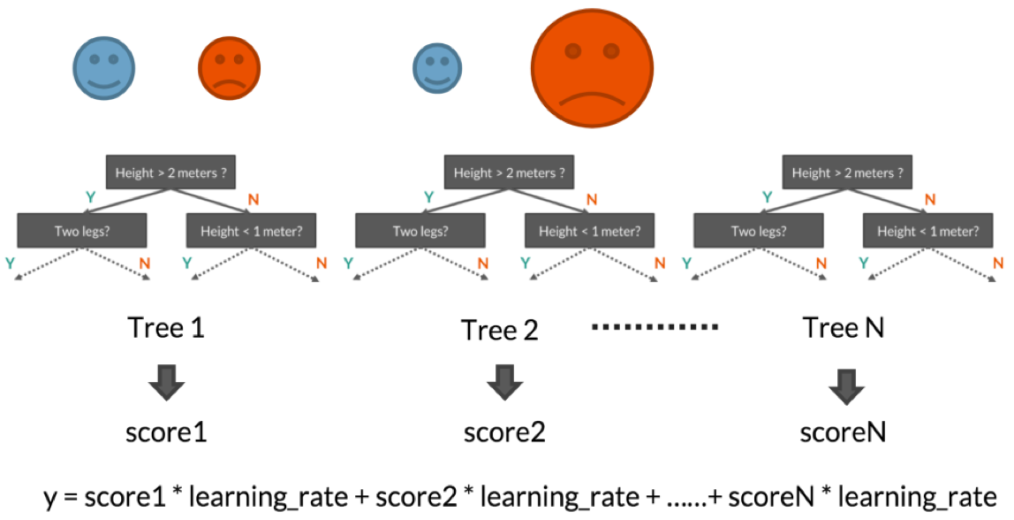


圖片來源:機器學習百日

梯度提升機 [梯度提升機實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_046_HW%20.ipynb)

梯度提升機是決策樹的延伸，透過觀察前一棵樹預測的錯誤

，在下一棵樹修正錯誤，再觀察這棵樹有無預測錯誤，重複以上動 作，將每顆樹組合形成模型，但可能會過擬和。



圖片來源:機器學習百日

* 參數調整 [參數調整實作](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/Day_047_HW%20.ipynb)

意義

許多模型都需要調整參數，如隨機森林、決策樹、梯度提升機等

調整方法:

窮舉法

直接指定參數範圍訓練，再選擇效果較好的參數

隨機搜尋法

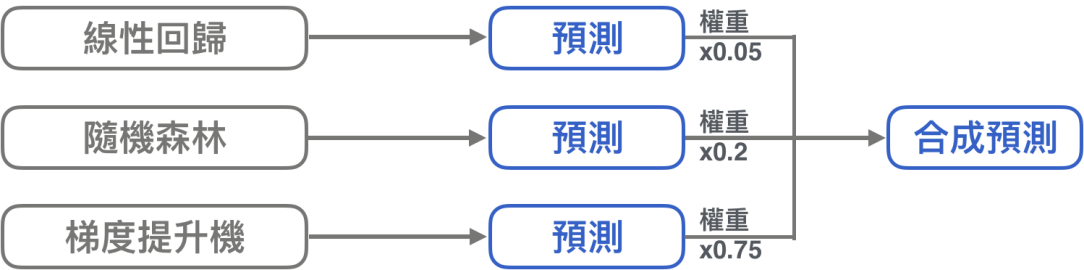
用均勻分布隨機抽取參數訓練，再選擇效果較好的參數

* 集成

混和泛化

基於模型差異大及各個模型都效果佳的情況下，將不同模型的預測

值加權合成，提升準確率。

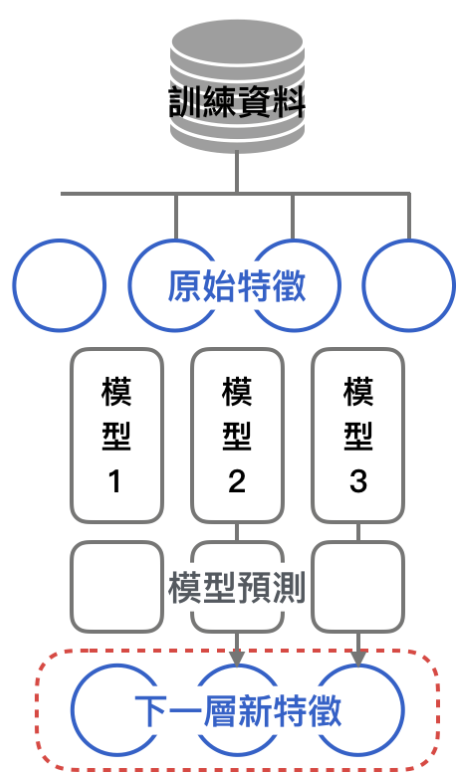


圖片來源:機器學習百日

堆疊泛化

堆疊泛化為混和泛化的改良，將預測結果混合，並使用預測結果當

新特徵，繼續進行模型預測。



圖片來源:機器學習百日

* 期中考\_\_Enron Fraud Dataset 安隆公司詐欺案資料集

題目敘述:安隆公司在2002年宣告破產，才揭露其財報在多年以來均是造假 的醜聞，現在有一份財報顯示公司內部來往的情報、薪資和股票等資訊，請透過機器學習模型找出誰是詐欺犯罪者?

實作重點:處理存在各種缺陷的真實資料(填補缺失值)

適當的評估函數了解預測結果(交叉驗證)

適當的特徵工程提升模型的準確率(最大最小化、增加新特徵)

[題目網站](https://www.kaggle.com/c/2021-ml100marathon-midterm) [期中考程式碼](https://github.com/TimTimLin/-1st-DL-CVMarathon/blob/main/kaggle%20%E6%9C%9F%E4%B8%AD%E8%80%83.ipynb)

成績 Score: 0.82142 排名:4

