資料探勘

第三組

組員：陳育霖，洪理川，向子聰，黃聖傑，曾俊諺

KNNの工作原理

加入新類別的資料x時，判斷最接近x的K個點的資料類別，定義x的類別.

1

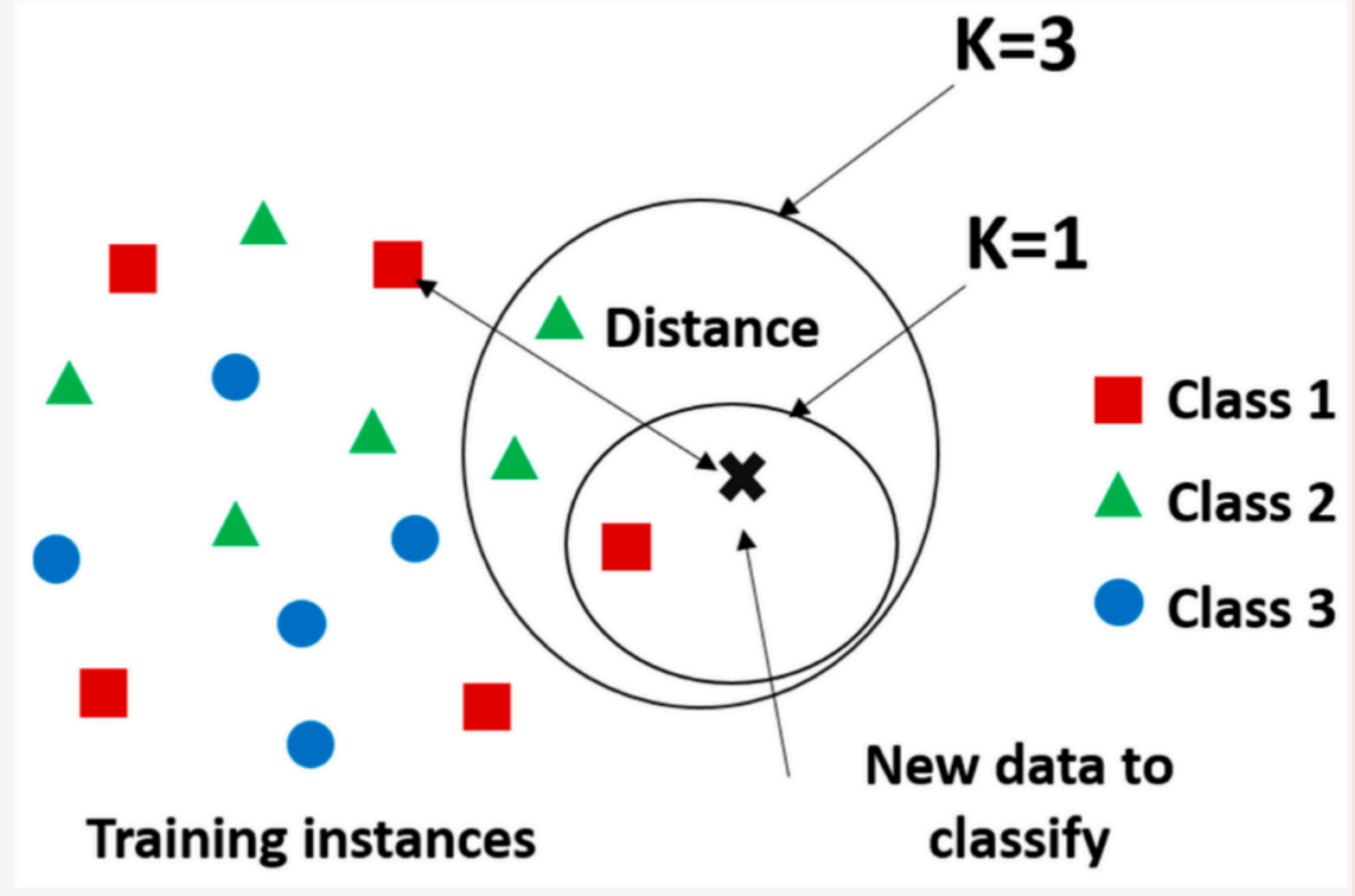
K值的選擇 -->過大不OK！
-->過小不OK！
K值只能奇數呦

2

樣本之間的距離 ：
資料中加入新點x時，KNN會先計算x和原先資料中所有點的距離。

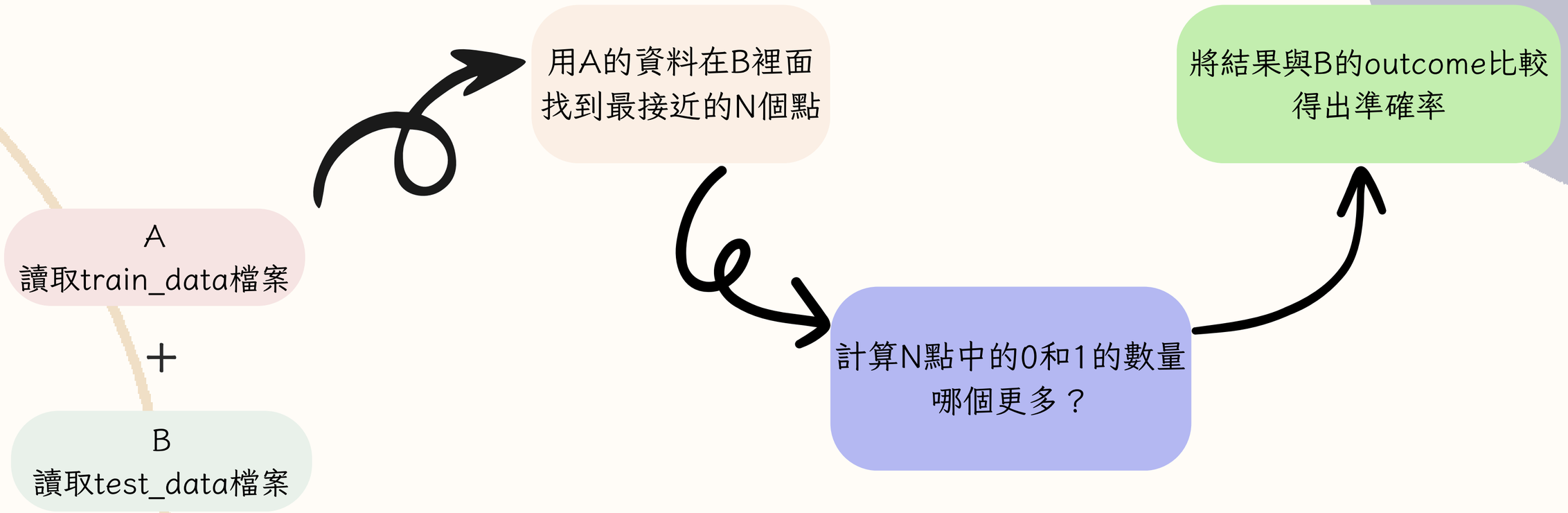
3

常見的距離計算方式有：
曼哈頓距離，歐幾裡得距離等。



圖片參考：https://www.mdpi.com/1424-8220/19/19/4058

KNN流程圖


A
讀取train_data檔案
＋
B
讀取test_data檔案
用A的資料在B裡面
找到最接近的N個點
計算N點中的0和1的數量
哪個更多？
將結果與B的outcome比較
得出準確率

邏輯斯回歸
(Logistic Regression)

基本的二元線性分類器

將機率P > 0.5 分類到 class A

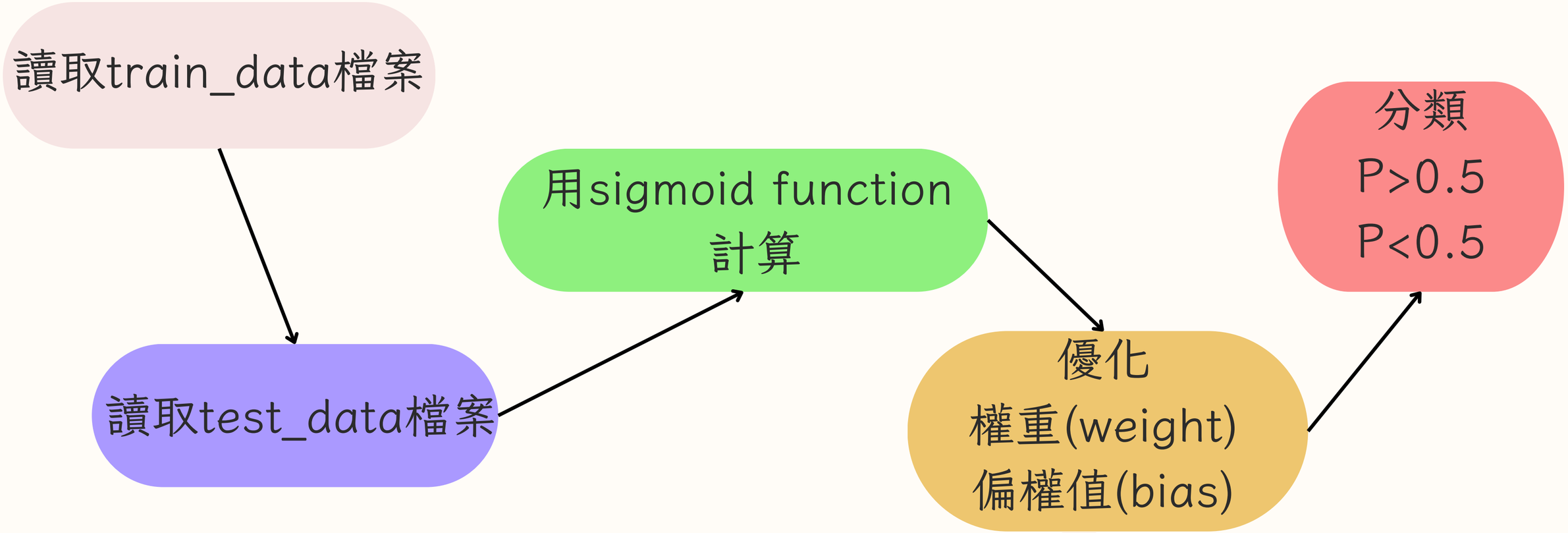
將機率P < 0.5 分類到 class B

偏權值(bias)
調整輸出決策邊界的位置

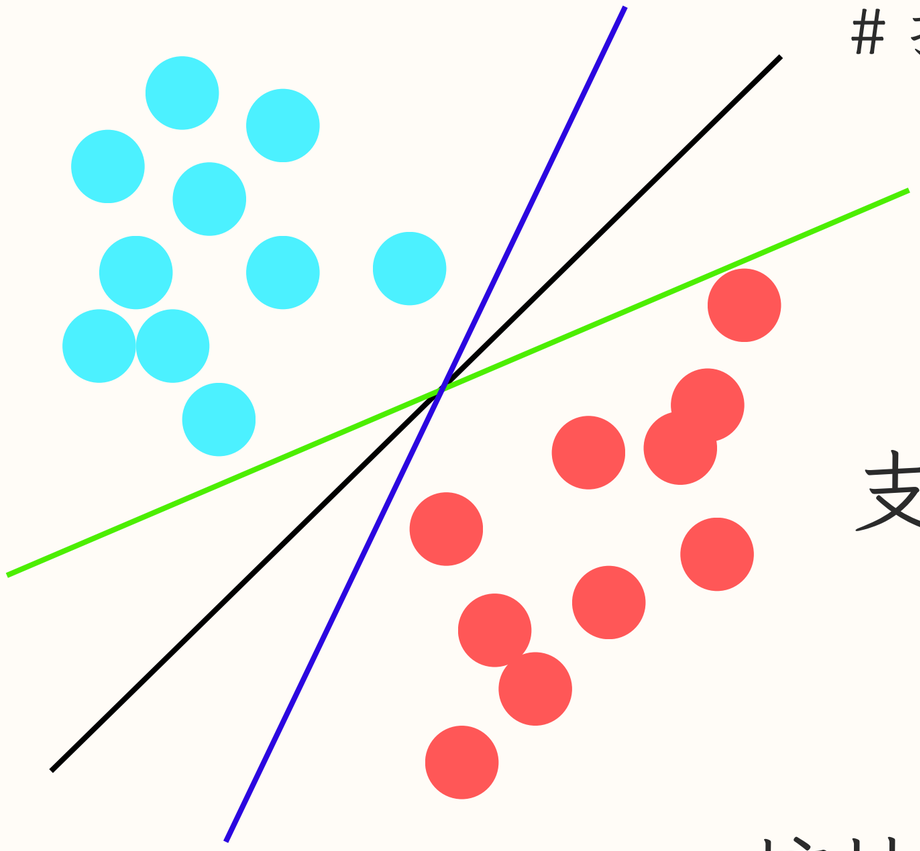
sigmoid function
將任意實數映射到（0,1）

權重(weight)
特徵對結果的影響強度

邏輯斯回歸 --> 流程圖


讀取train_data檔案
讀取test_data檔案
用sigmoid function
計算
優化
權重(weight)
偏權值(bias)
分類
P>0.5
P<0.5

SVM --支援向量機
(Support Vector Machine)



找出最適合的(N-1) 的超平面來進行分割。

超平面（hyperplane）

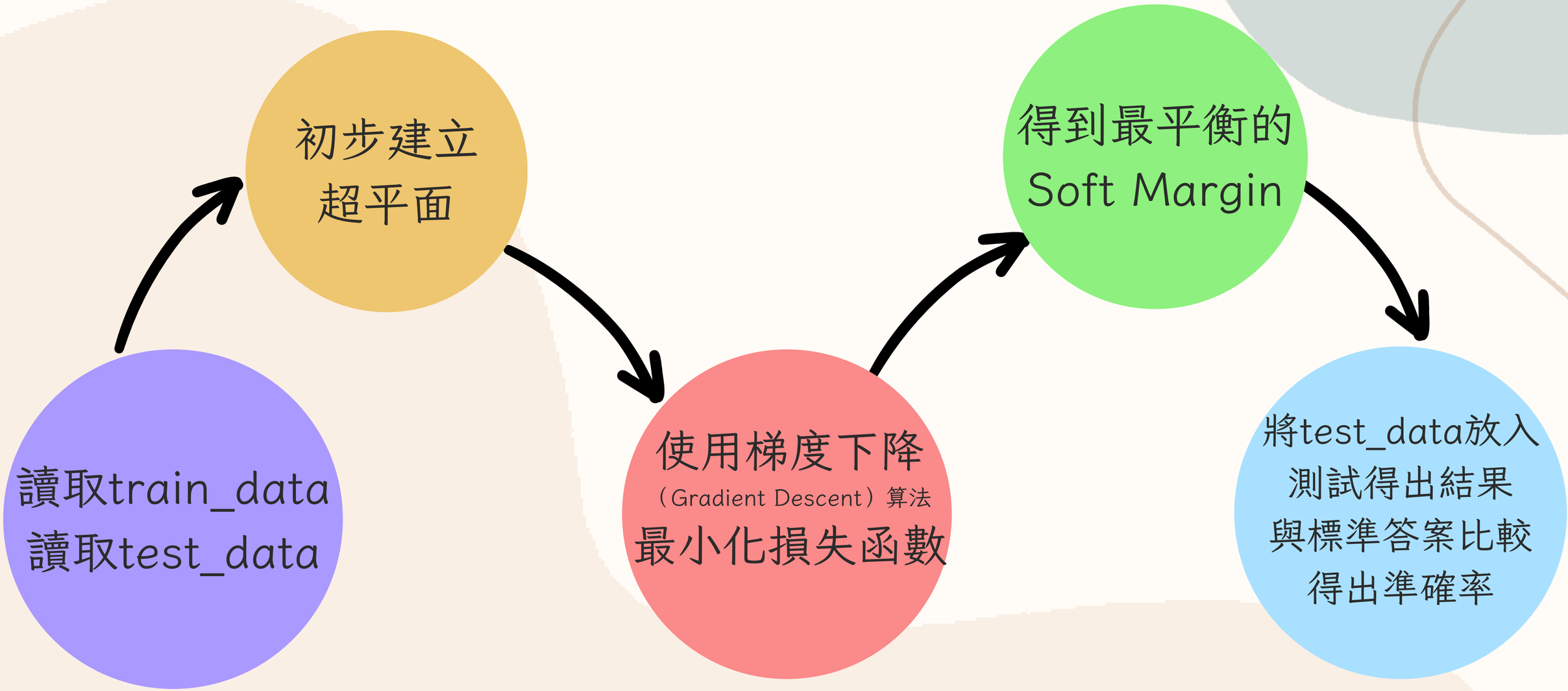
間隔（Margin）

支援向量（Support Vector）

軟/硬間隔（Hard / Soft Margin）

核技巧（Kernel Trick）

SVM 流程圖


讀取train_data
讀取test_data
初步建立
超平面
使用梯度下降
（Gradient Descent）算法
最小化損失函數
得到最平衡的
Soft Margin
將test_data放入
測試得出結果
與標準答案比較
得出準確率

決策樹
(Decision Tree)

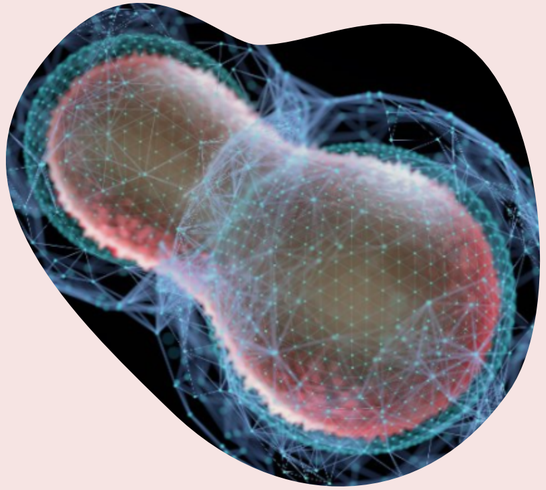
每一節點
代表一特征

樹狀結構

每一分支
代表一結果

最終的葉節點
代表最終分類

決策樹
の
產生



節點分裂

當一節點的特征
無法做出判斷時
分出新的子節點.

閾值確定

以二叉樹為例：
分出子節點後-------->
X節點的值將大於閾值
Y節點的值將小於閾值

圖片來源:https://geneonline.news/how-the-genome-is-packed-into-chromosomes-that-can-be-faithfully-moved-during-cell-division/

圖片來源https://img.christiantimes.cn/img/assets/media/post/202102/34309/www.christiantimes.cn-photo

切分時的維度選擇

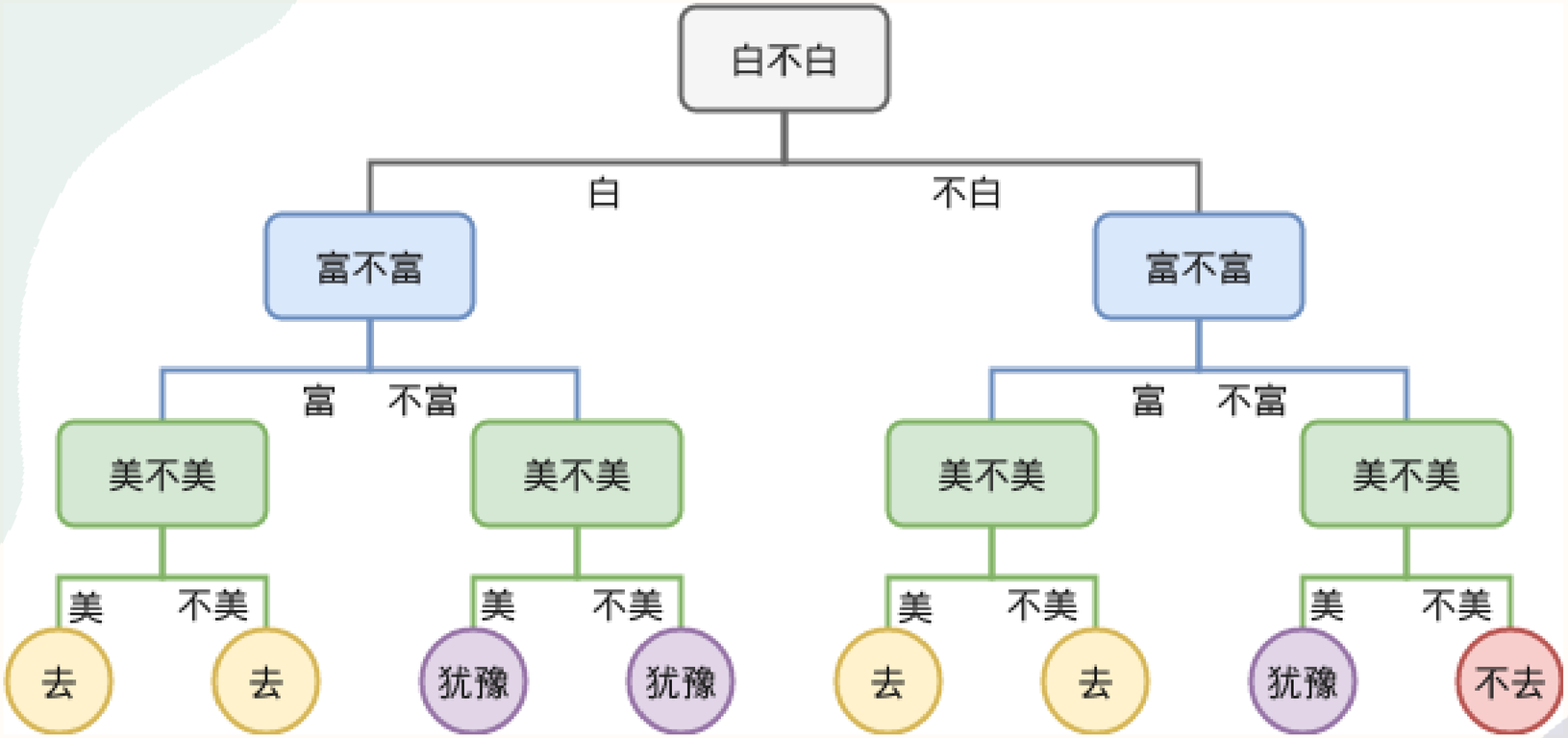
基尼不純度Gini impurity

一個從0到接近1的值，提供了一個度量標準來評估分類問題中資料集的混亂程度。
得到的值越小 --> 分類的效果越好。

信息熵增益 Information Gain

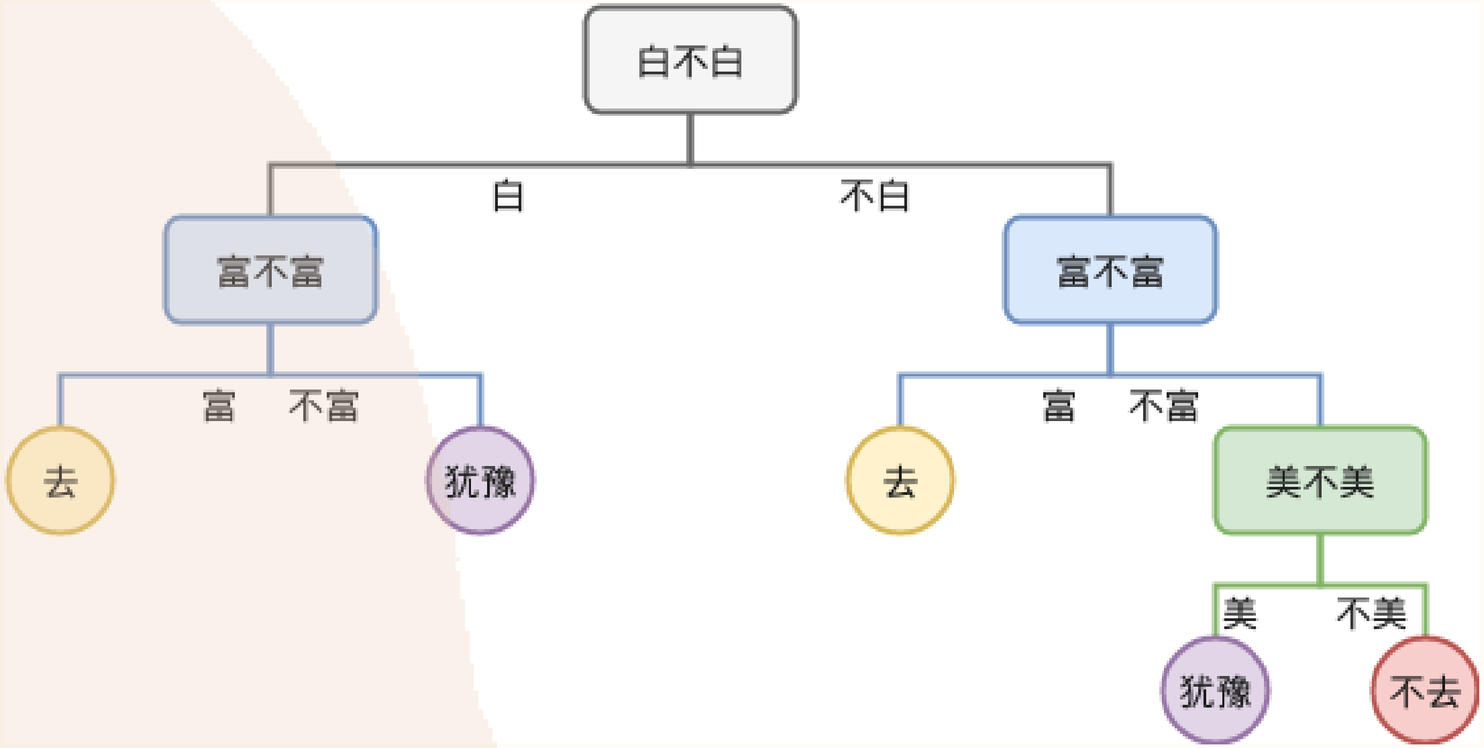
如果一个集合的熵越大，則集合越無序；熵越小，則集合越有序。

例子：基礎的樹



圖片來源：https://lotabout.me/2018/decision-tree/

合併後



提升泛化能力

圖片來源：https://lotabout.me/2018/decision-tree/

隨機森林 (Random forest)

從n筆資料中隨機抽出k個特徵作為樣本

從訓練的資料集中抽出n筆資料

重複執行m次
產生m顆決策樹

使用分類
多數投票機制
進行預測

Random forest 流程圖

讀取train_data檔案
讀取test_data檔案

設定計算次數n
（樹的數量）

設定抽取的
樣本數量

隨機抽取特征數
（4~8）

執行n次

結果為
0和1
誰多就是誰

結果和標準答案
比較得出準確率

Demo 時間

演算法分析

	KNN	Logistic Regression (No Data Standardization)	Logistic Regression (With Data Standardization)	SVM (No Data Standardization)	SVM (With Data Standardization)	Decision Tree	Random Forest
Average Accuracy Data A(10x)	75.622	70.647	80.597	70.647	78.109	73.114	76.122
Average Accuracy Data B(10x)	79.000	77.000	79.000	38.000	78.000	71.000	77.000