|  |  |  |  |
| --- | --- | --- | --- |
| Algorithm | 概念步驟 | 優點/缺點 | 備註 |
| K means | 1. 物以類聚的概念(**分群**) 2. 隨機在資料組中選擇中心點 3. 計算每一個點與中心點的距離，並找出最近的那一個 4. 利用當下的每個分群分別重新尋找中心點 5. 重複以上步驟直到收斂 | 1. 速度快，結果不差 2. 初始點設定不佳時結果會不好 3. 每次的結果都會不一樣 4. 用十字交叉驗證法(K-fold cross-validation)補足不同結果造成的問題 | 非監督 |
| Back Propagation Neural Network | 1. 需要輸入訓練資料進行學習 2. 分為輸入層、隱藏層(隱藏層不限數目)、輸出層(中間的節點稱為神經元) 3. 神經元之間的連結都會有不同的權重(通常稱w) 4. 通常會分為兩個階段:學習階段以及回想階段 5. 學習階段:已知的輸入層→未知的權重→已知的輸出層 6. 回想階段: 已知的輸入層→已知的權重→未知的輸出層 7. 利用最小誤差平方進行權重訓練 8. 計算**輸出結果**和**目標結果**的**誤差**再重新調整權重進行計算直至收斂 | 1. 可**建構非線性的模型** 2. 計算量大、耗費資源 3. 未知的輸入層也可以得到結果 4. 不知最佳解為何 5. 神經元數的多寡會影響結果，無法絕對掌握神經元的過程 6. 部分改良之人工類神經網路屬於非監督學習 | 監督 |
| k Nearest Neighbor(kNN) | 1. K=最近的數量 2. 當k=8，有一個黑點尚未分類別，找出離自己最近的八個點，範圍內有6黃2綠，該黑點自然的會變成黃(往多的靠) | 1. k的大小會是難題，有機會會造成不同的結果 2. 相對簡單的演算法 3. 要算出距離，記憶體消耗高 4. 利用已經分類好的資料對未分類的資料分類 5. 選擇特徵問題會有限制，不是每個特徵都可以量化成距離(例如頭髮顏色) | 監督 |
| Decision Tree | 1. 解決**分類**、回歸問題 2. 常見的有C4.5 與 ID3分類法 3. 利用Entropy，值越小越好代表類清楚 4. 再利用Entropy計算資訊獲利(Information Gain)，主要是利用父節點的Entropy減去子節點的Entropy結果是正的，代表該分類是有效的 5. 也可利用Gini計算，判斷分類有效與否與上者相同 | 1. 貪婪策略 Greedy Strategy有overfiting問題，要適當的修剪(pruning) 2. 小樣本的正確性會較高 3. 詳細計算方法可參考:http://123android.blogspot.tw/2011/11/111027-data-mining.html | 監督 |
| SVM(Support Vector Machine) | 1. 線性**分類**的一種，將資料映射到高維的空間當中，讓資料有向量空間可以計算 2. 可訓練許多分類器，找出可以最大間隔超平面(Maximum-margin hyperplane)的分類器，簡單來說就是要在兩群資料中盡可能地分出邊界，該邊界會盡可能地遠離資料點 3. 劃出中心線後會在兩側再畫出邊界，兩個邊界內的資料點就分屬於該邊界的不同群族 | 1. 資料不需要特定格式皆可分析 2. 計算消耗大，時間久 3. 介紹影片: https://youtu.be/3liCbRZPrZA | 監督 |
| 羅吉斯回歸 | 1. 傳統的回歸模型 2. 屬於**分類**的一種 3. 簡單的以解釋變數(X)和反應變數(Y)構成 | 1. 可建構模型 2. 易受資料結構影響，複雜資料上分類錯誤率高 | 監督 |
| Association rule - Apriori | 1. 利用資料的關聯程度進行計算 2. 其中的重要變數有Support(該資料出現在資料庫的機率)、Confidence(信心水準，檢測預設強度) 3. 開始會分別計算每個資料出現的次數(EX:A、B、C、D)，並利用minSupport進行篩選；再繼續組合存在的值(BC、BD、CD)再繼續篩選，直到最長組合出現(BCD) 4. 把過程中留下的組合(B、C、D、BC、CD、BCD)計算各自的Confidence，在與自己設下的門檻(通常會設定0.7)比對留下通過門檻的組合 | 1. 可能會遭遇feature過多，導致組合數量龐大，計算時間與記憶體消耗都相對龐大，而要考慮進行降維 2. 要對最後結果進行合理解釋，就算是預測結果是100%，也不一定是有用的(EX:買麵包的同時會買果醬) 3. 過程當中針對不同商品或資料最好要有專家介入討論，例如醫療領域、賣場，比較好掌握重要的feature或是解釋 | 非監督 |