D18 K-近鄰演算法

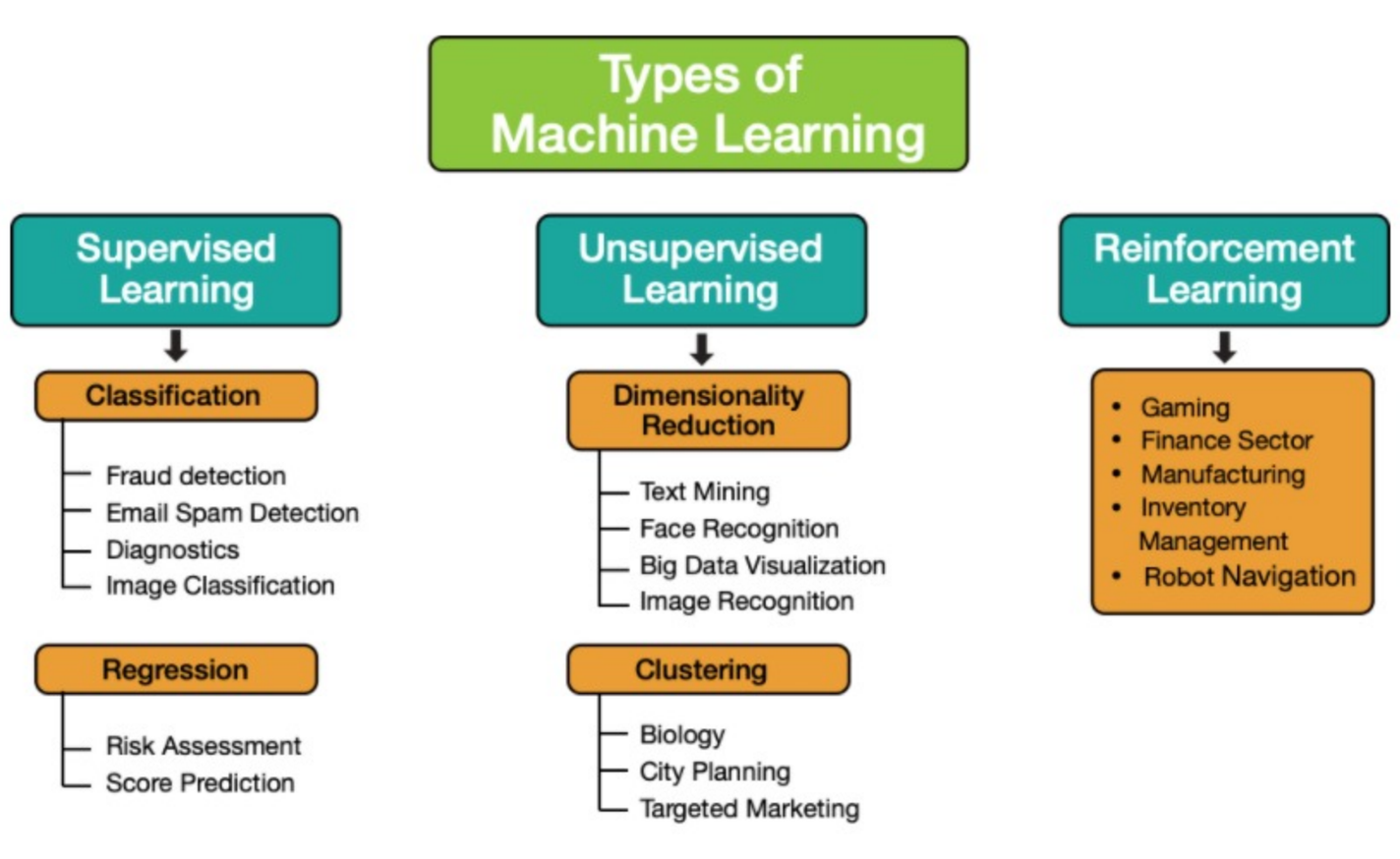
簡報閱讀 | 範例與作業 | 問題討論 | 學習心得(完成)

重要知識點

- 了解 Machine Learning 演算法分類
- 了解 k-nearest neighbors 演算法原理

演算法

由於 KNN 是本次課程接觸到的第一個**分類演算法**，在介紹 KNN 之前，先帶大家認識 Machine Learning 領域中演算法的基本分類。



資料來源：DataFlair

演算法(Supervised vs Unsupervised Learning)

Supervised Learning (監督式學習)

- 監督式學習得重點在於，需要 標注檔案(Labeling) 訓練。舉個例子，我們希望模型能夠認識貓和狗的圖片，首先我們需要做的就是 Data Annotation(資料標記)，我們標注每張照片所屬的類別，並且將照片與相對應的標注檔案送入模型學習。

- 監督式學習主要又可以分為：

1. 分類問題(classification)：預測有限的類別 ，像是預測貓狗。
2. 回歸問題(regression)： 預測連續的數值，像是房價、身高體重。

Unsupervised Learning (非監督式學習)

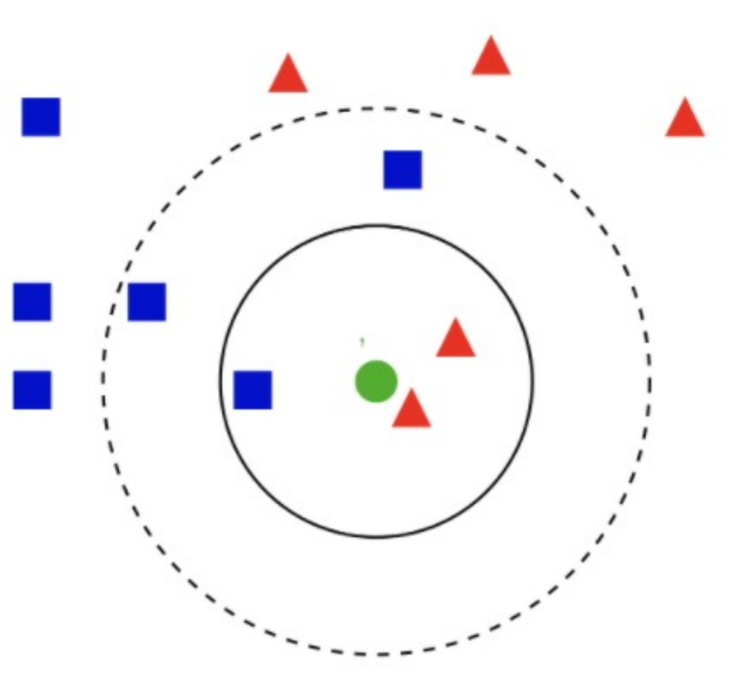
- 非監督式學習的重點在於，不需要 標注檔案(Labeling) ，而是特過資料本身的特徵來進行處理，舉個例子，一樣是貓狗分類，但這個時候我們不利用標注資料，而是用圖片本身資訊，計算不同圖片的相似度(如像素距離)，將相似度較高的歸為一類。

- 非監督式學習主要又可以分為：

1. 降維(Dimension Reduction)：將高維度特徵壓縮成低維度，ex. PCA
2. Clustering(集群)：利用資料本身特徵聚類，ex. K-mean
3. Anomaly detection (異常檢測)

K-近鄰演算法

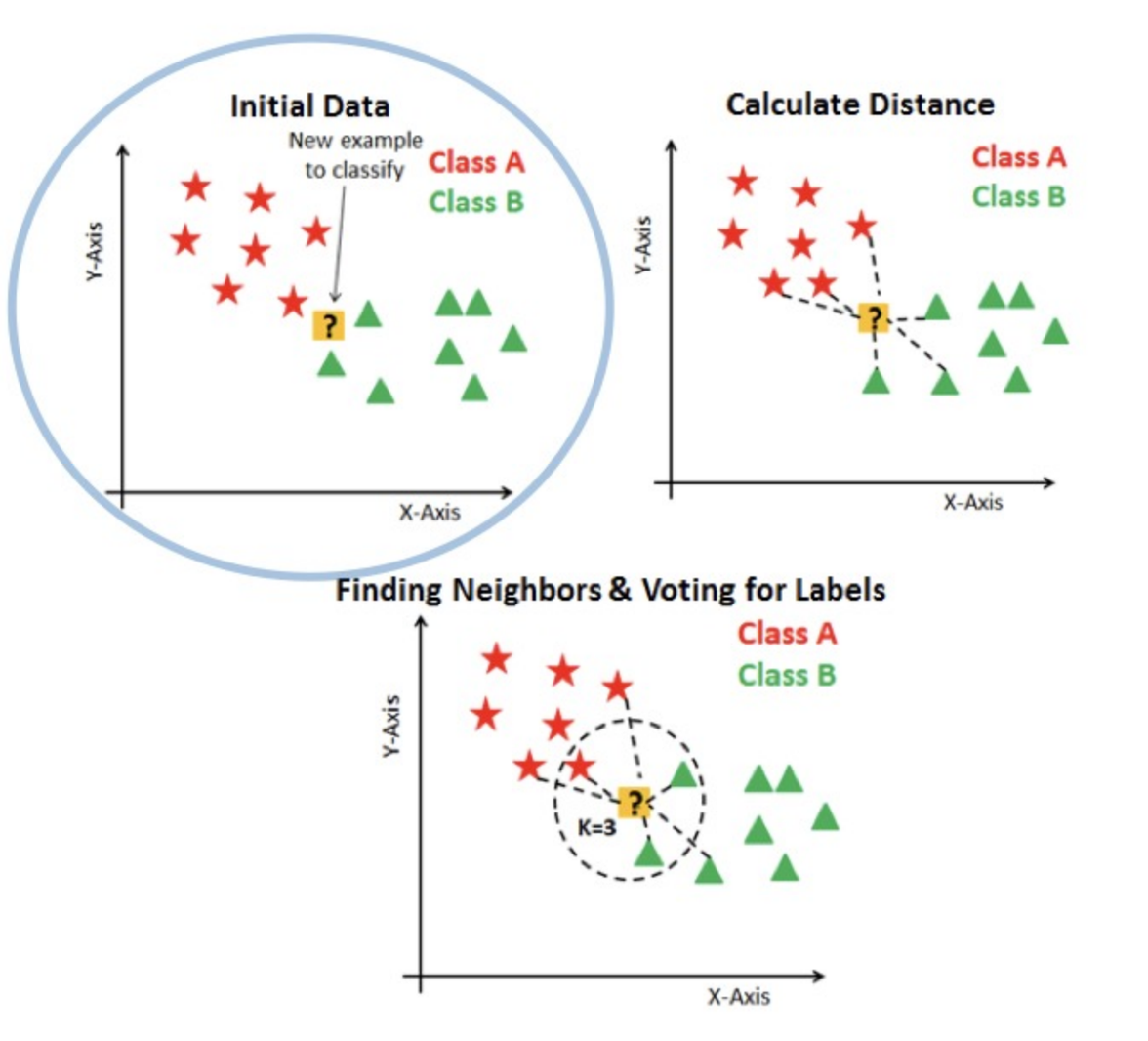
- k-nearest neighbors 演算法(KNN) 就是上述 Supervised Learning 底下的 分類演算法 (classification)。
- NLP領域中常見的文章分類，垃圾郵件分類等問題都可以藉由KNN解決。



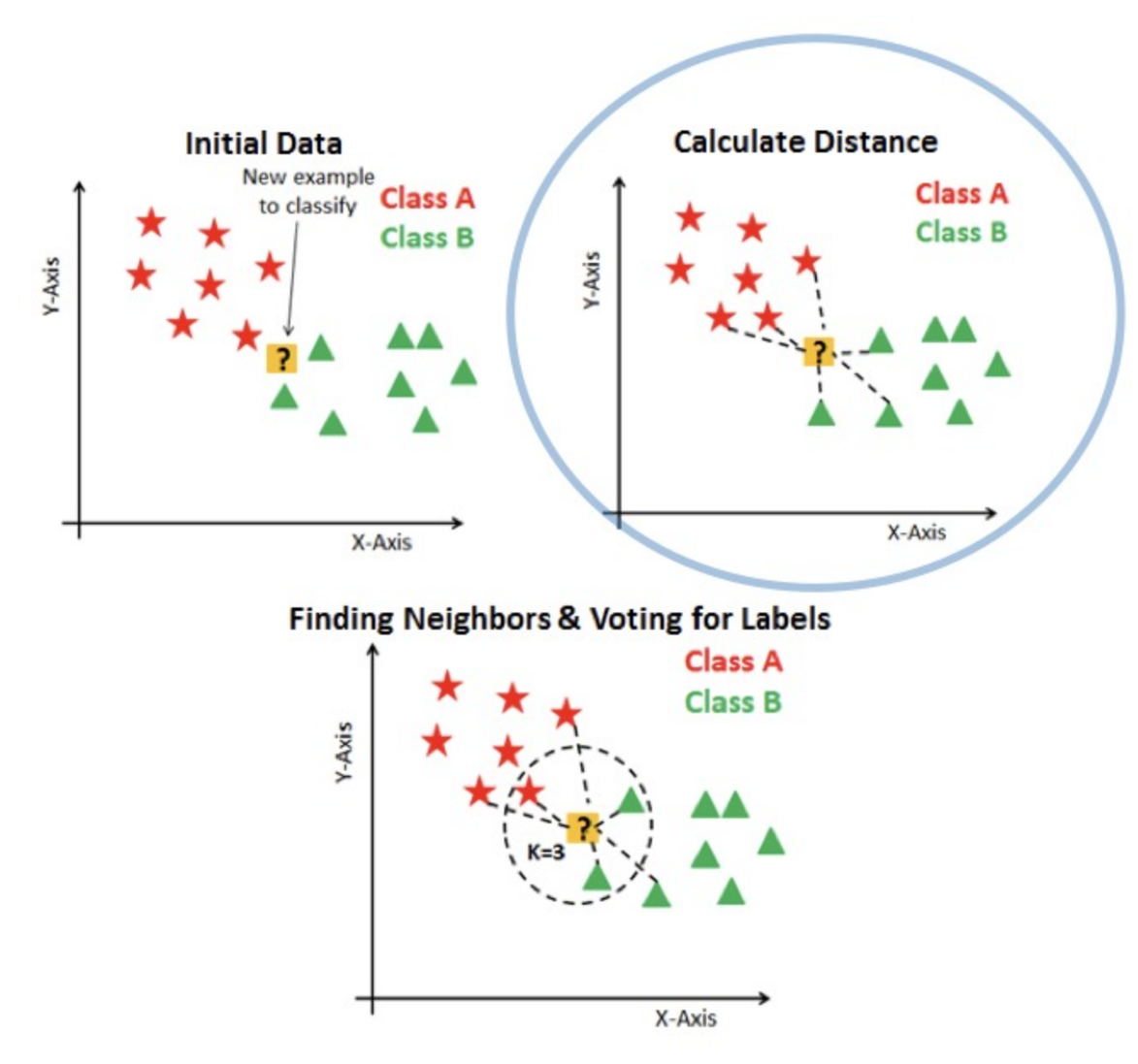
資料來源：AnalyticsVidhya

K-近鄰演算法原理

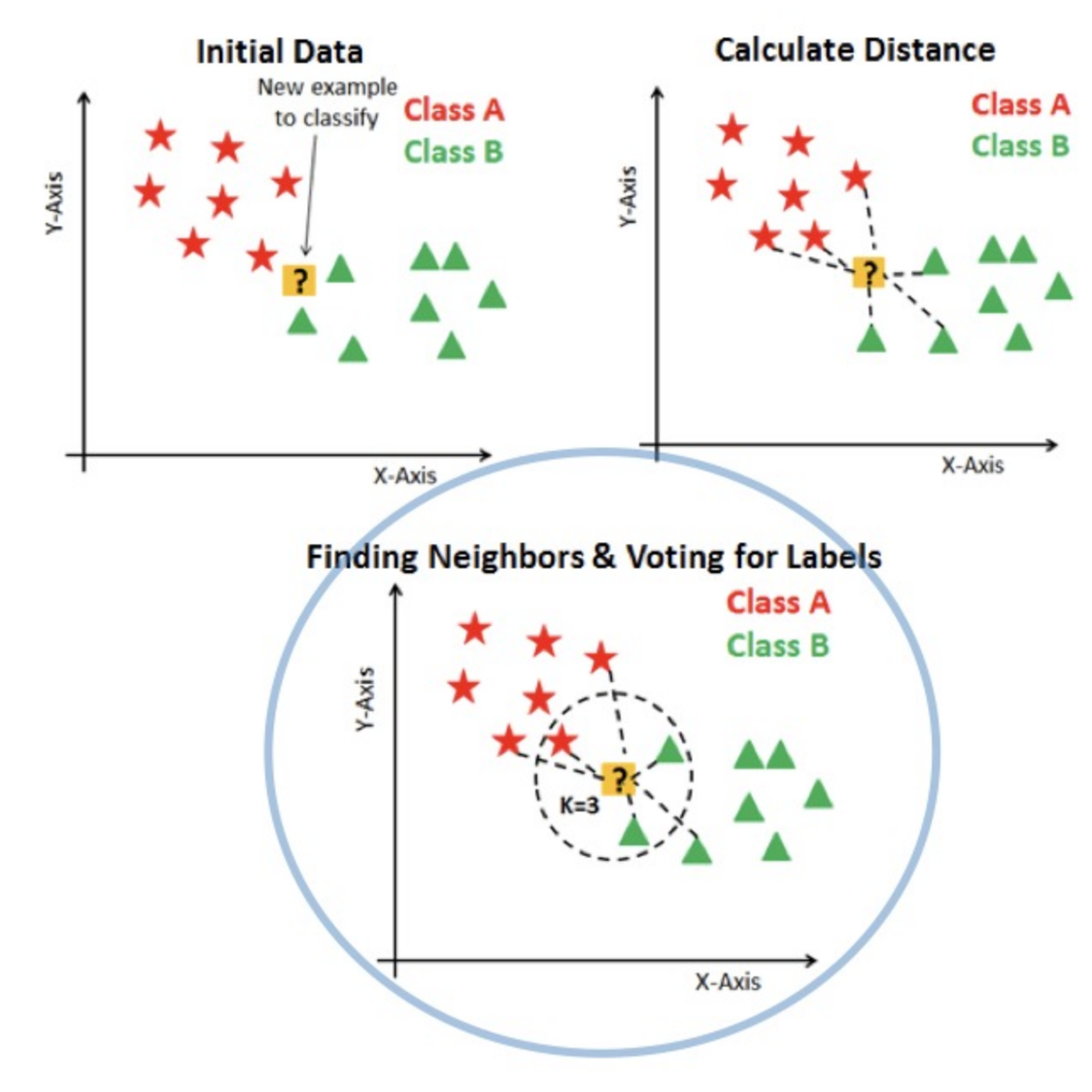
- **第一個步驟：**
  初始化資料，我們會將訓練集資料的資訊記住，包括每個 sample 的特徵，與相對應標記。



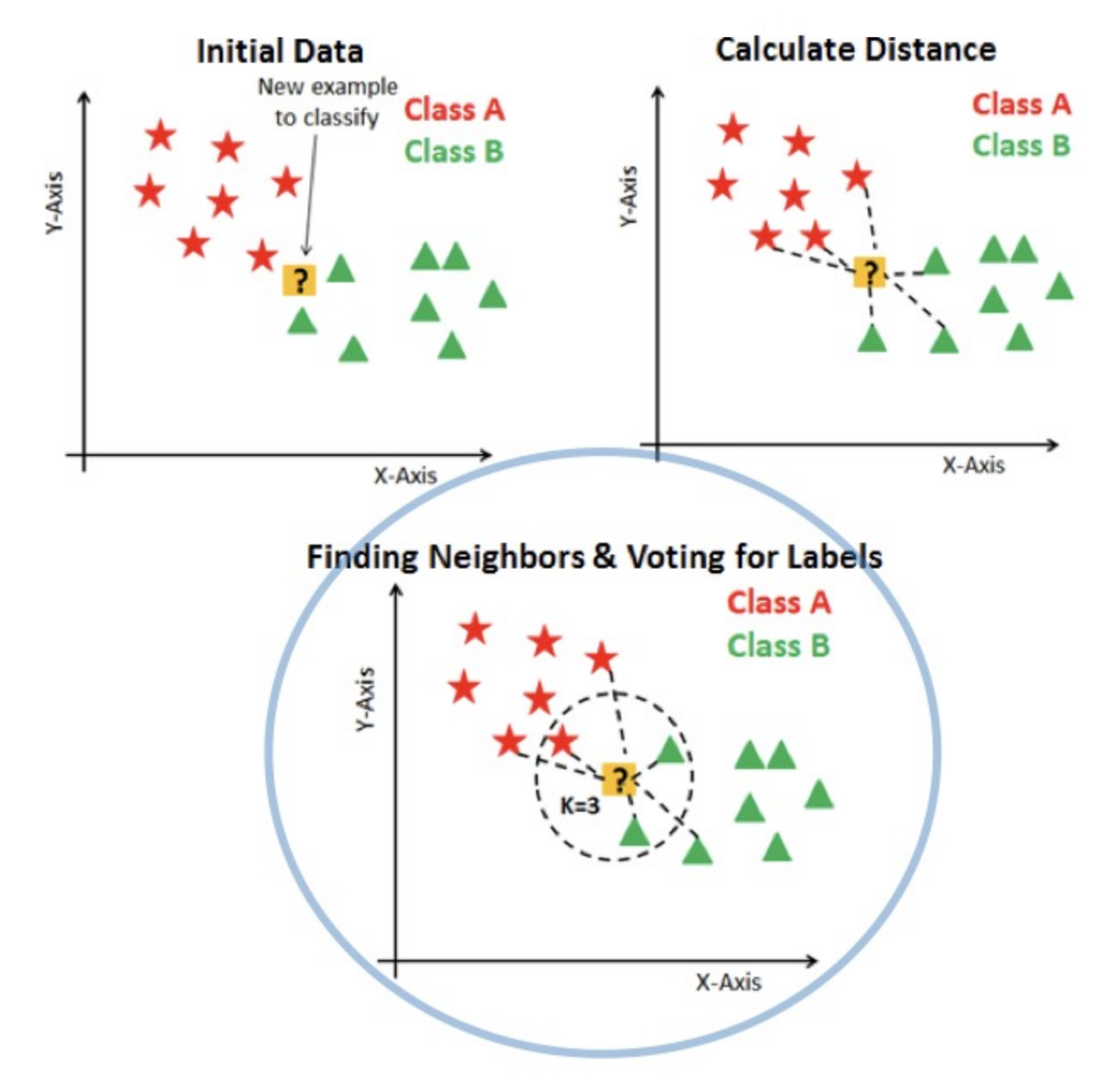
- **第二個步驟：**
  當今天我們要預測新的 sample 時，我們會計算新的 sample 與訓練集內 sample 的距離。



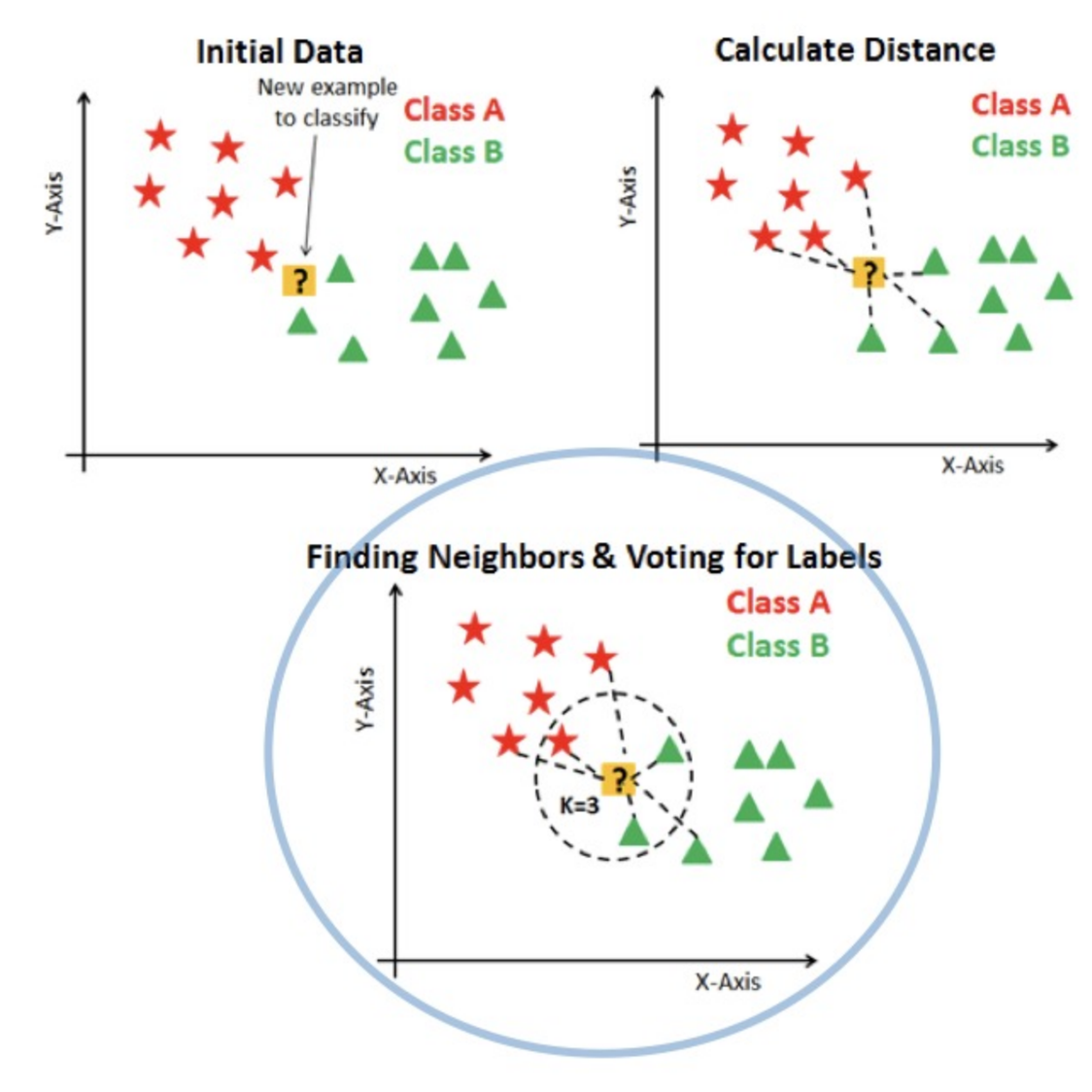
- **第三個步驟：**
  計算完距離後，來到 KNN 的核心觀念，選定 K 個最近的鄰居，並查看其標注的類別。



- **第四個步驟：**
  以下圖為例，我們選定了3個鄰居，而其中包含兩個Class B 以及 一個 Class A。



- **第五個步驟：**
  此時我們就可以說新的sample屬於類別B。



- 儘管 KNN 看似簡單，然而在許多的分類問題上，仍有不俗的表現。
- 本日課程先介紹了 KNN 的基本概念，更多的細節會在明天課程講解，ex. 該如何選擇 K 值？KNN 演算法的優缺點？

參考資料：AnalyticsVidhya

參考資料

網站：PCA：經典降維演算法，也會用在 image augmentation (whitening)

淺談降維方法中的 PCA 與 t-SNE

愷開 Follow
Jul 19, 2017 · 7 min read

在機器學習當中，如果特徵數過多時，有可能會造成一些問題，像是：

- 過擬合 (overfitting)
- 處理速度較慢
- 如果超過三個特徵以上不好視覺化

所以這時候就需要對特徵做降維，在實務上，一個幾百幾千個的特徵當中，手動挑選特徵顯然不是一個明智的方法，所以以下來介紹兩個在機器學習中常常使用的兩種降維方法。

PCA（principal component analysis）主成份分析

在介紹 PCA 之前，我們先來定義一下我們的目標是什麼：

> 將一個具有 n 個特徵空間的樣本，轉換為具有 k 個特徵空間的樣本，其中 k < n

網站：PCA 與 LDA

機器學習: 降維(Dimension Reduction)- 線性區別分析(Linear Discriminant Analysis)

Tommy Huang Follow
May 15, 2018 · 8 min read

線性區別分析(Linear Discriminant Analysis，LDA）是一種supervised learning，這個方法名稱會讓很人confuse，因為有些人拿來做降維(dimension reduction)，有些人拿來做分類(Classification)。如果用來做降維，此方法LDA會有個別稱區別分析特徵萃取(Discriminant Analysis Feature Extraction, DAFE)，多個名字也比較好區隔。

此篇主要是要講降維(dimension reduction)部份。如果有看過PCA的介紹，再來看這篇會比較有感覺，也比較容易上手。

在降維度的方法上，LDA是PCA延伸的一種方法，怎麼說哩。PCA目標是希望找到投影軸讓資料投影下去後分散量最大化，但PCA不需要知道資料的類別。而LDA也是希望資料投影下去後分散量最大，但不同的是這個分散量是希望「不同類別之間的分散量」越大越好。所以LDA和PCA差異的部份，PCA是無監督式(unsupervised learning)方法，LDA多了這四個字(**「不同類別」**)是一種監督式(supervised learning)方法。

LDA怎麼實現投影後不同組之間的分散量越大越好?

下一步：閱讀範例與完成作業