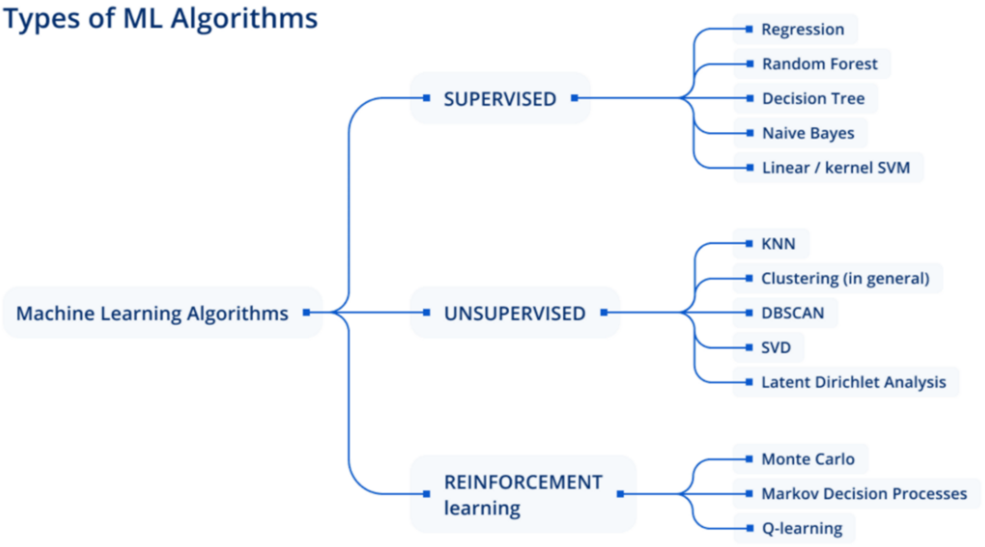
資料探勘期中考定義題彙整



The three measures, in general, return good results but

Information gain:

> 偏向**多值屬性**biased towards **multivalued attributes**

Gain ratio:

> 傾向於**不平衡的分割**，其中一個分區比其他分區小得多
> tends to prefer **unbalanced splits** in which one partition is much smaller than the others

Gini index:

> 適用於**多值屬性**，當**class數量很大計算會變的很困難**
> 較**適用在兩個分區大小相同且純度相同**的測試

> biased to multivalued attributes has difficulty when **# of classes is large tends to favor tests** that result in equal-sized partitions and purity in both partition

Give the definition of data mining. (5%)

> 資料探勘可以說是**從資料中發現知識**，從大量的資料中提取有趣（**重要的、隱含的、以前未知的和潛在有用的**）模式或知識，資料探勘是對**大型資料集進行排序**，以**識別可透過資料分析幫助解決業務問題的模式和關係的過程**。

> 資料探勘技術和工具使企業能夠**預測未來**趨勢並做出更明智的業務決策。

▼ Data mining (knowledge discovery from data)

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis.Data mining techniques and tools enable enterprises to predict future trends and make more-informed business decisions.

Data Science

> 資料科學是透過**發現、假設和分析假設**的過程直接從數據中提取可使用的知識。

What is SVM? (5%)

> SVM 是一種**監督學習**演算法，可用於**分類和迴歸任務**。

> SVM原理是**在高維度空間**中找到**最大限度地分隔不同類別的超平面**。並根據**資料點落在超平面的哪一側對資料點進行分類**。

> **SVM 的主要功能包括：**
>
> - 可以處理**高維度數據**

> - 在**類別之間有明顯分隔的情況下有效**
> - 可以**核化以處理非線性**邊界
>
> 例如SVM 應用在人臉辨識，根據五官特徵分類。

▼ SVMs

are a type of supervised learning algorithm that can be used for both classification and regression tasks. They work by finding the hyperplane in a high-dimensional space that maximally separates the different classes. Data points are then classified based on which side of the hyperplane they fall on.

Key features of SVMs include:

1. Can handle high-dimensional data
2. Effective in cases where there is a clear margin of separation between classes
3. Can be kernelized to handle nonlinear boundaries

A real-world example of SVMs in action is in face recognition, where they can be used to classify different faces based on features such as the shape of the eyes and nose.

What are support vectors in an SVM? (5%)

> 在支援向量機（Support Vector Machine，SVM）中，**支援向量**是來自訓練數據集中**定義不同類別之間的決策邊界關鍵的數據點**。

> SVM的主要目標是找到一個**最佳的超平面**，它能夠**最好地將數據分為不同的類別**，超平面**最大化了間隔（超平面與每個類別數據點之間的距離），同時最小化分類錯誤**。

> 也就是說，演算法會識別支援向量，並且最終的**超平面是基於這些支援向量的位置確定**的，支援向量在**定義決策邊界的位置和方向**相當重要。

> SVM多用於在**具有複雜、非線性數據分佈**的情況下，且通常與**核函數等技術**一起使用。

What is the difference between a hard margin SVM and a soft margin SVM? (5%)

> **hard margin SVM 和soft margin SVM** 兩者都是**利用超平面將資料分類**。

> 主要區別在於**它們對錯誤分類數據點的容忍度以及處理非線性可分離數據的能力**。

> **hard margin SVM** 能完美的將資料分類，**確保分類的正確性**，缺點是對輸入的數據要求較高，**需要完全可分類的數據**，若數據不能分類時就會產生錯誤，hard margin SVM**不允許任何錯誤分類**。

> 而**soft margin SVM** 更彈性、是較靈活的版本，當**資料不能完全分類或存在異常值**可以使用，soft margin SVM接受**數據中一定程度的錯誤分類和雜訊**。

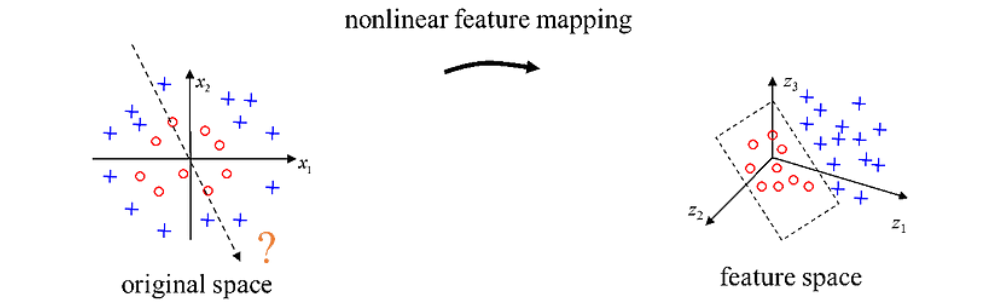
> 兩者之間的選擇取決於**數據性質以及最大化margin和容錯分類的接受度**。

How kernel functions are useful in training an SVM? (5%)

> 在機器學習內，一般說到kernel函數都是在SVM中去介紹，主要原因是**SVM必須搭配kernel l函數才能讓SVM可以在分類問題中得到非常好的效果**。

> Kernel trick主要作用於資料在原始空間中無法被**線性分類器區**隔開來時，**利用非線性投影(*φ*)，將資料進行轉換到更高維度的空間，**促使分類能順利進行。

> 主要是透過**kernel函數**來設計出一個好的**非線性投影(*φ*)公式**



It's difficult to find the linear separating relationship between this data set.

It's easier to find complete linear separating relationship in the feature space

Tell me about Backpropagation Algorithm

> Deep Learning的演算法—**反向傳播法**

> 反向傳播是一種**監督式學習的方法**，是深度學習中的演算法，利用反向傳播**找到損失函數對於權重的梯度，進而利用梯度下降法 (Gradient Descent) 來對每一個權重進行優化**。

> 並且反向傳播要求**神經元的Activation Function 是可微分的**

> 最主要的概念，**就是將誤差值往回傳遞資訊，使權重可以利用這樣的資訊進行梯度下降法來更新權重，進一步的降低誤差**。

What is the definition of gradient descent? gradient

> **梯度下降法(gradient descent)**是**最佳化理論**裡面的**找最佳解的一種方法**，主要**用梯度下降法找到函數局部最小值**。

> 算一個**函數*f*(*x*)的梯度函數要是任意可微分函數**，這也是深度學習為什麼都要**找可微分函數**出來**當激活函數(activation function)**。

> 我們需要Training的就是這些**權重Wij(ℓ)**，Gradient Descent的流程：
>
> 1. 定義出**Error函數**
> 2. Error函數讓我們可以去評估**Ein**

3. 算出梯度**∇Ein**
4. 朝著的**∇Ein反方向更新參數W**，而每次只跨出**η大小**的一步
5. **反覆計算新參數的梯度**，並一再的**更新參數W**

What is decision Tree how it works

決策樹是一種**監督學習演算法**，常用於**分類和回歸任務**。主要為**創建樹狀結構**，而**結構則根據某些規則或條件**將數據**拆分為越來越小的子集**。進而**達到每個數據點的預測或分類**。

決策樹的主要功能包括：

1. 易於理解和解釋
2. 可處理**數位和分類數據**
3. 可**處理多個輸入**要素

例如透過決策樹做醫學診斷，可根據患者的病史和測試結果確定患者癥狀的最可能原因。

▼ Decision trees

are a type of supervised learning algorithm that can be used for both classification and regression tasks. They work by creating a tree-like structure that splits the data into smaller and smaller subsets based on certain rules or conditions. The final splits result in predictions or classifications for each data point.

Key features of decision trees include:

Easy to understand and interpret

Can handle both numerical and categorical data

Can handle multiple input features

A real-world example of decision trees in action is in medical diagnosis, where they can be used to determine the most likely cause of a patient’s symptoms based on their medical history and test results.

Naive Bayes

> Naive Bayes是一種**簡單但功能強大**的分類演算法，它使用Naive Bayes定理進行預測。它**假設所有輸入特徵彼此獨立**，這使得它“Naive”(天真、幼稚)，但也能做出**快速準確的預測**。

> **Naive Bayes的主要特點包括：**
>
> 1. **簡單**且易於實施
> 2. **快**速高效 / 速度快
> 3. **可以處理大量的輸入要素**

> 例如透過Naive Bayes進行垃圾郵件檢測，可根據發件者、主題行和電子郵件內容**等特徵將電子郵件分類為垃圾郵件**。

▼ Naive Bayes

is a simple but powerful classification algorithm that uses the Bayes theorem to make predictions. It assumes that all input features are independent of each other, which makes it “naive” but also allows it to make fast and accurate predictions.

Key features of Naive Bayes include:

Simple and easy to implement

Fast and efficient

Can handle a large number of input features

A real-world example of Naive Bayes in action is in spam detection, where it can be used to classify emails as spam or not based on features such as the sender, subject line, and content of the email.