資料探勘（Data Mining）

> 參考資料：
>
> 資料探勘（Data Mining）Dr. Tun-Wen Pai
>
> Business Intelligence and Data Mining Anil K. Maheshwari, Ph.D.

資料探勘的介紹

資料探勘是一種技術，為了探索大量經過組織的資料中，得到有利的資訊、見解與模式。

資料探勘利用了統計學、人工智慧領域、機器學習的知識，挖掘資料中未知的資訊，例如發現資料的類別與結構、分類資料等等。

資料蒐集與資料選擇

從大量支離破碎的資料中很難探勘出有用的資訊，所以我們需要先做資料蒐集與資料選擇。

資料從大量的資料來源中被蒐集，此時資料是支離破碎的，所以我們需要經過一系列的處理（例如資料倉儲的 ETL）

在建立資料倉儲之前，我們可以使用企業模式的資料模型（Enterprise Data Model, EDM），為這些資料打造一個統一的框架。

經過資料倉儲一系列的處理之後得到組織過的資料，再從這些資料進行探勘。

這也是為什麼資料倉儲被用來幫助資料探勘，因為資料探勘需要整理過的資料，此時即完成了資料蒐集與資料選擇。

資料淨化

資料的品質會嚴重影響到資料探勘，所以我們會希望資料都是高品質的，也因此我們需要把資料淨化。

我們可以利用一些手段進行資料淨化，例如：填充遺失的欄位、處理異常值、劃分連續型變數等等。

經過資料淨化，就可以確保資料都是高品質，確保不會造成 Garbage in Garbage out 的問題。

以下列舉一些資料淨化的手段：

1. 重複的資料需要被移除，從各方資料來源蒐集資料可能會導致出現重複的資料，所以需要被移除。
2. 欄位若遺失值則需要被填上去，如果不該被填上去，則這個欄位應該被移除。

3. 資料元素從單一單位轉換成另一個單位。

   例如透過總病人數量與總花費較難得到有利的資訊，但如果資料是病人與花費的對應關係，那麼我們可以從這邊得到更有利的資訊。

4. 連續型變數可以被劃分以利於資料探勘更佳。

   例如工作經驗可以被劃分成低、中、高。

5. 資料元素可能需要經過調整，讓他能夠隨著時間的推移產生可比性。

   例如大量不同的貨幣可以被調整成通用貨幣，用來評估通膨的情況。

6. 極值需要被移除。
7. 偏差數值需要經過矯正，來確保分析結果是正常的。
8. 資料應該保持同樣的顆粒度（Granularity），來讓資料能夠比較。

   例如：櫃台銷售所產生的資料通常都是以日為單位，而銷售員的資料通常都是以月為單位。

   為此我們應該將櫃台銷售調整成月為單位，兩個資料才做比較。

9. 資料需要足夠密集。

Confusion Matrix

Confusion Matrix 可用來監督學習、可以確定在機器學習中是否將兩個不同類型的資料混淆了。

以採檢病人是否為陽性反應為例，我們可以獲得以下的表格。

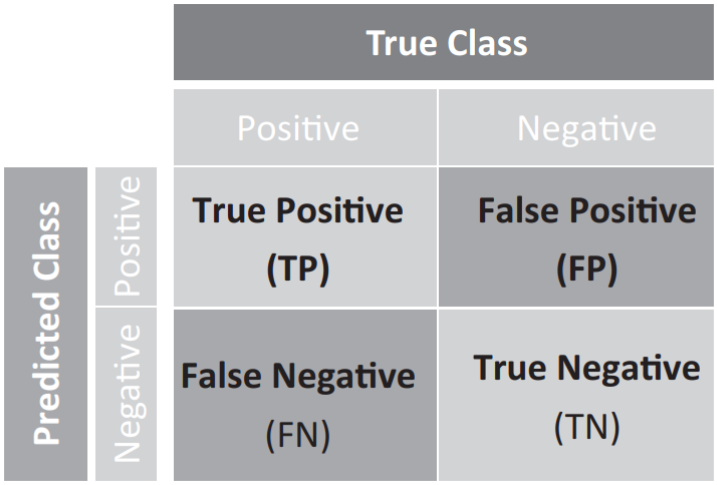


Figure 4.1 Confusion matrix

以理想的情況來說，採檢試劑要能夠分辨出陽性（True Positive）與陰性（True Negative）。

但必定會出現偽陰（False Negative）與偽陽（False Positive）的情況出現。

將這些數值填入表格，即可得到一個 Confusion Matrix。

我們可以擴展表格，得到更多的資訊。

		Predicted condition			
	Total population = P + N	Positive (PP)	Negative (PN)	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P}$ = 1 − FNR	False negative rate (FNR), miss rate = $\frac{FN}{P}$ = 1 − TPR
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N}$ = 1 − TNR	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N}$ = 1 − FPR
	Prevalence = $\frac{P}{P+N}$	Positive predictive value (PPV), precision = $\frac{TP}{PP}$ = 1 − FDR	False omission rate (FOR) = $\frac{FN}{PN}$ = 1 − NPV	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$
	Accuracy (ACC) = $\frac{TP+TN}{P+N}$	False discovery rate (FDR) = $\frac{FP}{PP}$ = 1 − PPV	Negative predictive value (NPV) = $\frac{TN}{PN}$ = 1 − FOR	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
	Balanced accuracy (BA) = $\frac{TPR+TNR}{2}$	F_1 score = $\frac{2PPV \times TPR}{PPV+TPR}$ = $\frac{2TP}{2TP+FP+FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP+FN+FP}$

Sources: [20][21][22][23][24][25][26][27] view · talk · edit

將同樣的概念，可以套用到機器分類數字、機器分類花...等情況，所以可以適用於績效評估機器學習。

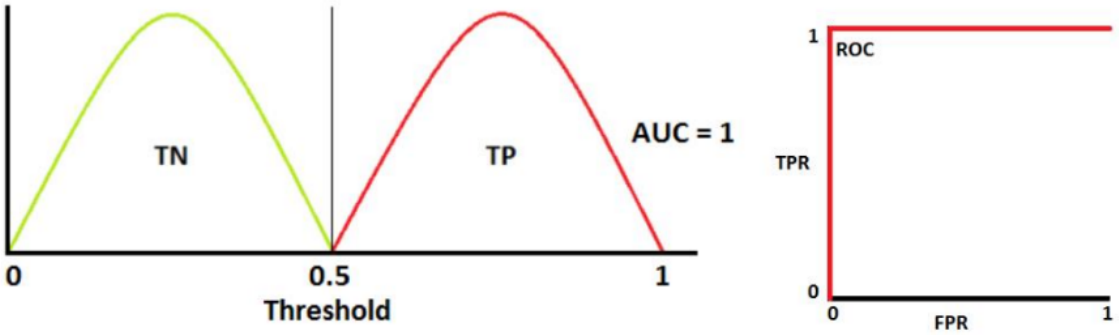
AUC - ROC Curve

我們可以使用 AUC - ROC Curve 來進行機器學習的績效評估。

對於上面表格，我們可以設定特定的閥值，用來得到在這個閥值中的真陽、真陰、偽陽、偽陰數量，進而劃出由兩個曲線組成的圖片。

例如我們預期在 TPR 閥值低於 0.5 時則為陰性，在閥值高於 0.5 時則為陽性。

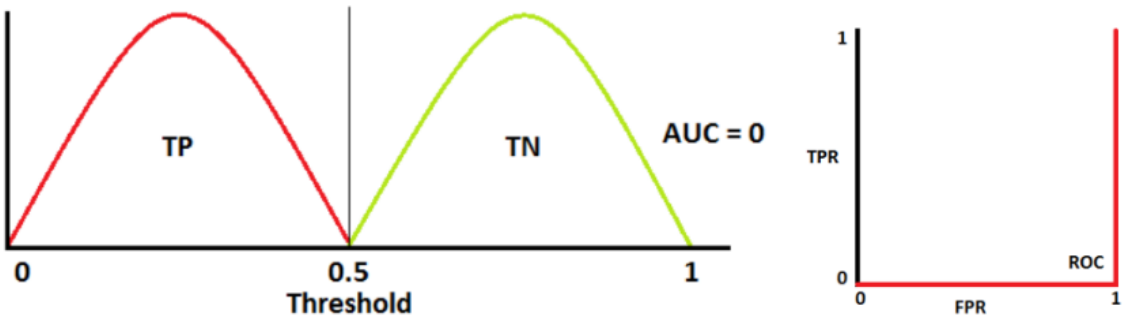
此時我們可以預期沒有任何偽陰、偽陽的情況，所以可以得到以下這張圖。



這時無論到哪點都不會出現偽陽性，故 ROC 曲線呈現直角，幾乎是完美分類。

另一個例子，若我們期望 TPR 閥值低於 0.5 時為陰性，高於 0.5 時則陽性。

然而結果不如預期，得到了以下這張圖。

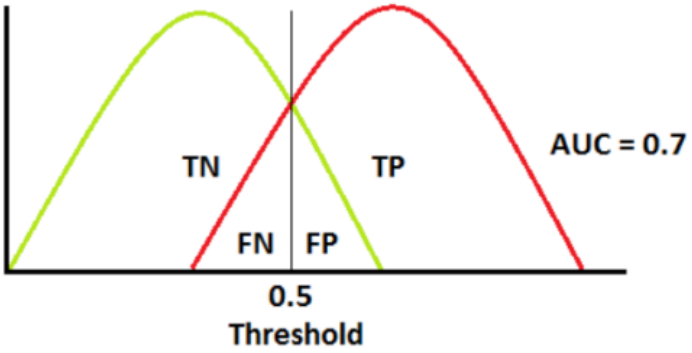


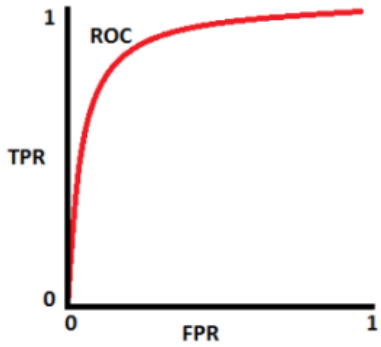
這時我們只需要反預測即可校正回上例。

換另一個不理想的例子，若可能出現偽陰偽陽的情況，

例如閥值是 0.46 時，陰性與陽性都出現，這時就會出真陽、偽陽的情況，得到以下這張圖。

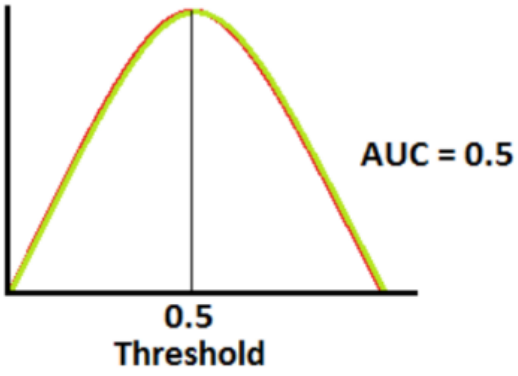
可以發現，在出現 FPR 時，TPR 的數值稍微下跌，而在之後幾乎回到 1 的位置。

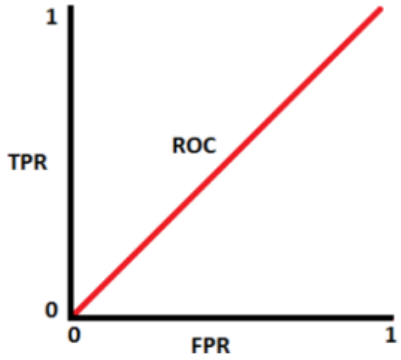




另一個例子，若無論在任何閥值出現的情況，真陰、真陽的人數都一樣，

則這個模型視同無效，因為檢驗一個人是否偽陰偽陽的機率相等於隨機。





我們可以透過描點的方式畫出 ROC 曲線，就可以從 ROC 曲線中得出 AUC 值，若 AUC 值越大則模型越好。

機器學習的種類

1. 學習問題
    1. 監督式學習：所有資料都被標註，給機器去學習與分類，通常來說對機器最簡單，對人類來說最累。
    2. 非監督式學習：所有資料都沒有被標註，透過機器去尋找特徵的學習分式，對機器最困難，誤差較大。
    3. 強化式學習：不標註任何資料，但告訴機器哪步正確，哪步錯誤，讓機器逐步自我修正。
2. 混合學習問題
    1. 半監督式學習：對少部分資料進行標註，機器透過有標註的資料找出特徵並進行分類，能使非監督式學習的準確率提升。

2. 自我監督式學習：從一堆沒有 label 標註的資料，訓練出一個監督式模型，然後造出更多 label。
3. 多實例學習：輸入許多「包」，這些包都含有許多實例，若所有實例都是負例時則包即為負包，若有至少一個實例為正例時，包即為正包。

3. 推論統計學
    1. 歸納學習：用來辨識汽車的知識可以用來提升辨識卡車的能力，以現有問題的解決模型利用在其他不同但相關的問題上。
    2. 演繹學習：利用廣泛的前提去推論較具體的結論，通常依賴前提是否正確。
    3. 轉導學習：通過觀察特定的訓練樣本，來預測特定測試樣本的方法。
4. 學習技術
    1. 多任務學習：利用單一一個模型，解決多個問題
    2. 主動學習：從每輪學習迭代中，尋找出一個最不確定的一個或一組樣本，來讓外部反饋者給予回饋。
    3. 線上學習：利用當前的資料來直接更新模型，進而在預測前根據先前的資料給予預測。
    4. 遷移學習：一個模型先訓練一個例子，接著所有模型用這個模型的訓練例子當成起始點，訓練其他的例子。
    5. 集成學習：集成兩個以上合適的模型，接著從這個集成的模型上訓練。。
    6. 聯盟式學習：從各式各樣的自訓練模型中，在不用給自己數據的情況下，也可以進行訓練得到模型。

> Reference Website:
>
>
> 1. Marketing. (n.d.). *你知道機器學習(Machine Learning)，有幾種學習方式嗎？伊雲谷 eCloudvalley*. Retrieved April 5, 2022, from https://www.ecloudvalley.com/zh-hant/machine-learning/
> 2. *自監督學習 self-supervised learning 介紹*. (2021, June 11). 藏字閣. https://jigfopsda.com/zh/posts/2021/self_supervised_learning/
> 3. *多實例學習*. (2021, August 31). 維基百科，自由的百科全書. https://zh.wikipedia.org/wiki/%E5%A4%9A%E7%A4%BA%E4%BE%8B%E5%AD%A6%E4%B9%A0
> 4. *遷移學習*. (2020, September 19). 維基百科，自由的百科全書. https://zh.wikipedia.org/wiki/%E8%BF%81%E7%A7%BB%E5%AD%A6%E4%B9%A0
> 5. *簡要介紹Active Learning(主動學習)思想框架，以及從IF（isolation forest）衍生出來的演算法：FBIF（Feedback-Guided Anomaly Discovery）*. (n.d.). IT人. Retrieved April 5, 2022, from https://iter01.com/420911.html
> 6. Su, S. (2021, December 13). *聯盟式學習 (Federated Learning) - Sherry.AI*. Medium. https://medium.com/sherry-ai/%E8%81%AF%E7%9B%9F%E5%BC%8F%E5%AD%B8%E7%BF%92-federated-learning-b4cc5af7a9c0
>