ROC Curve, AUC, PR Curve and F1-Score

Introduction

> 在處理分類問題時，最後的輸出往往是樣本被歸在某類的"機率"，最後將樣本歸類於機率最大值的類別。

但統計有一句很有名的格言

> All models are wrong, but some are useful.

也就是很難找到一個完美的模型能夠為樣本完美分類，畢竟樣本本身就會有變異數在

那麼如何對一個模型"評分"呢?又或者有甚麼指標可以參考?

底下介紹常見的評估方法

ROC Curve

> ROC的全名叫做Receiver Operating Characteristic，是以偽陽性率(FPR)為橫軸，真陽性率(TPR)為縱軸所繪製的曲線， 曲線上的點代表不同閾值(threshold)下，該模型FPR和TPR的對應關係，底下介紹觀念及算法

混淆矩陣(Confusion Matrix):

	實際YES	實際NO
預測YES	True Positive(TP)	False Positive(FP)
預測NO	False Negative(FN)	True Negative(TN)

真陽性率(TPR),召回率(Recall),敏感度(Sensitivity):

意思:在所有正樣本中，有多少比例預測正確

$$TPR = \frac{TP}{TP + FN}$$

偽陽性率(FPR):

意思:在所有負樣本中，有多少比例被預測為正，也等於1-特異度(Sensitivity)

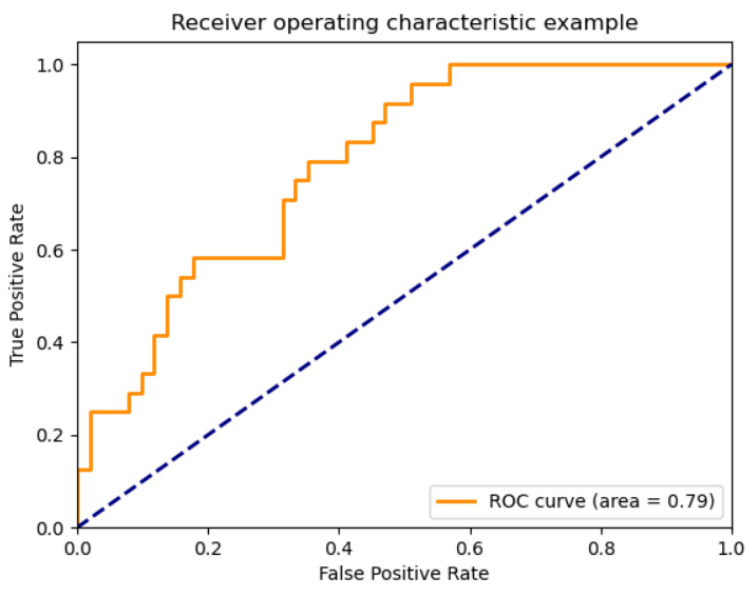
$$FPR = \frac{FP}{FP + TN} = 1 - \frac{TN}{TN + FP} = 1 - Specificity$$

Remark:特異度(Sensitivity)的意思為，在所有負樣本中，預測正確的比例

ROC and AUC

ROC:將各個Threshold情況下的FPR和TPR對應關係描繪出來

AUC:全名叫做Area Under the Curve(AUC)，就是ROC曲線下的面積，下圖為0.79



圖片來源:https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html#sphx-glr-auto-examples-model-selection-plot-roc-p

討論:

由圖可發現，當TPR越大，那麼FPT也會越大，這是因為當你"放寬"了Threshold時，雖然會讓正樣本預測正確的比例增加，但錯把負樣本分類成正(Positive)的機會也同時增加了。

PR Curve

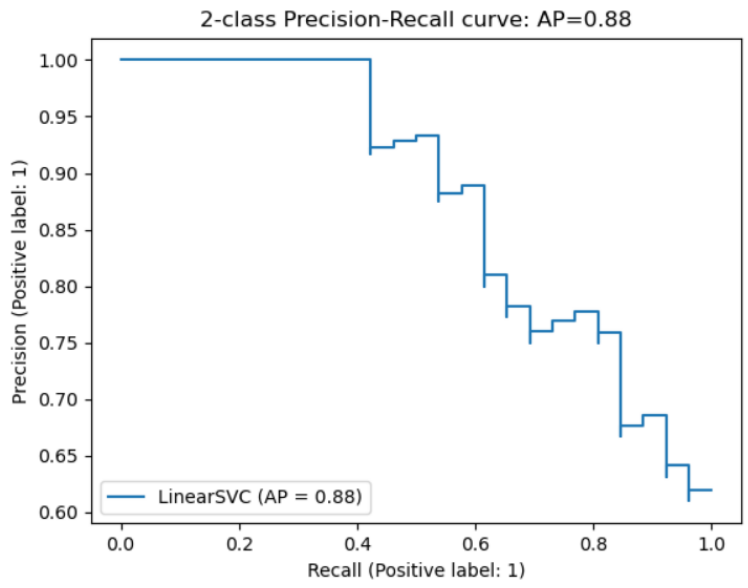
> 以準確率(Precision)縱軸，召回率(Recall)為橫軸，所繪製的曲線圖

準確率(Precision)

意思:在預測為正的樣本中，確實為正樣本的比例。

$$Precision = \frac{TP}{TP + FP}$$

Precision- Recall Curve



圖片來源:https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html#sphx-glr-auto-examples-model-selection-plot-precision-recall-py

Average Precision(AP):

不同threshold下，Recall差異比上Precision的平均

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

R_n=Recall at the nth threshold

P_n=Precision at the nth threshold

圖片來源:https://scikit-learn.org/stable/modules/generated/sklearn.metrics.average_precision_score.html#sklearn.metrics.average_precision_score

討論:

Precision vs. Recall

Precision探討的範圍是在預測為正樣本的情況下，有多少比例為真的正樣本

而Recall探討的是範圍是在真實為正樣本的情況下，有多少比例預測正確

所以根據不同目標要關注的指標就會不同

使用場景

在資料不平衡的情況下，例如疫苗的正確率，通常使用ROC，因為ROC會將正樣本和負樣本分開討論，所以比較可以克服正負樣本不平衡的問題

在資料平衡的情況下，PR更專注於那些預測為正樣本的分類正確率，比方說信用卡詐欺檢測。

一般情況下，調高threshold能提高Precision，但Recall會降低

F1-Score

> 將Precision和Recall做調和平均的指標。

由上面討論可發現，不同情況下所關注的指標不同，也可能要同時考慮到兩者，這時候將兩者做調和平均做為新的指標，就是F1-Score。

F1-Score是F-measure通式的一個特例

F-Measure

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}$$

圖片來源:Wiki

F1-Score

Let $\beta = 1$

$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

圖片來源:Wiki

討論:

由F-measure的計算式可發現，當β趨近於無限大時，F-measure就是Recall；當β等於零，則代表了Precision。