機器學習\統計方法: 模型評估-驗證指標(validation index)

Tommy Huang Follow
Jul 10, 2018 · 11 min read

這篇主要是說「怎麼評估我們訓練出來的模型，成效(performance)好不好」。
這時候就會牽扯到我們這個定義所謂的成效，所以這篇介紹一些驗證指標(validation index)來當成效指標，依據應用分為**「分類指標」**和**「回歸指標」**。

「分類指標」: 二元相關(二元混淆矩陣和相對應驗證指標、**ROC曲線**、**AUC)**和**多元相關(多元混淆矩陣**和相對應驗證指標)。
Note: 二元指標內有比較多diagnosis index算法和介紹。

「回歸指標」: 平均均方誤差(Mean Squared Error, MSE)、**平均絕對誤差(Mean Absolute Error, MAE)**和**平均均方對數誤差(Mean Squared Logarithmic Error, MSLE)**

· · ·

分類指標(Classification metrics)

分類指標這邊會是大宗，主要原因是除了機器學習之外，很多臨床研究或是統計研究也會用到這邊的指標，不一定是機器學習才會用到，後續會繼續說明為什麼。

分類這邊我們可以很直接知道，分類大概可以分成**二元分類(binary case)**和**多元分類(multiclass case)**，所有的分類問題都可以先產生出一個稱為混淆矩陣(Confusion matrix)的東西，然後從這個矩陣在去算出一些成效指標。

二元分類這邊介紹會比較多，主要原因是醫學臨床和統計學用比較多，所以會有很多名詞，如果只是想要看分類的指標可以直接看多元指標。

二元分類(binary case)指標

二元混淆矩陣(Confusion matrix)

在二元分類基本上就是分「有」和「沒有」、「真」跟「假」、「正」和「負」。下表是二元分類的混淆矩陣，True condition就是你資料的答案，Predicted outcome就是模型預測出來的結果。
Positive就是「有」、「真」或是「正」，在醫學上通常用「有發病」；
Negative就是「沒有」、「假」或是「負」，在醫學上通常用「沒有發病」。

True Positive (TP)「真陽性」:真實情況是「有」，模型說「有」的個數。
True Negative(TN)「真陰性」:真實情況是「沒有」，模型說「沒有」的個數。
False Positive (FP)「偽陽性」:真實情況是「沒有」，模型說「有」的個數。
False Negative(FN)「偽陰性」:真實情況是「有」，模型說「沒有」的個數。

		True Condition	
	Total Population (T)	Positive	Negative
Predicted outcome	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

混淆矩陣(Confusion matrix)

這邊舉個例子說confusion怎麼算的，基本上多元的也是用同樣的方式算。
假設我有一組資料是看有沒有生病，然後醫師做診斷和電腦去診斷，結果如下:

	真實狀況	醫師診斷	電腦診斷
S1	生病	生病	生病
S2	生病	生病	生病
S3	生病	生病	沒生病
S4	生病	生病	生病
S5	沒生病	沒生病	沒生病
S6	沒生病	沒生病	生病
S7	沒生病	沒生病	沒生病
S8	沒生病	沒生病	沒生病
S9	生病	生病	沒生病
S10	生病	生病	生病

先轉成二元指標

	真實狀況	醫師診斷	電腦診斷
S1	1	1	1
S2	1	1	1
S3	1	1	0
S4	1	1	1
S5	0	0	0
S6	0	0	1
S7	0	0	0
S8	0	0	0
S9	1	1	0
S10	1	1	1

我們去算「真實狀況和醫生判斷的混淆矩陣」和「真實狀況和電腦判斷的混淆矩陣」，如下表:

		True Condition	
	T=10個人	生病	沒生病
醫師診斷	生病	TP = 6	FP = 0
	生病	FN = 0	TN = 4

		True Condition	
	T=10個人	生病	沒生病
電腦診斷	生病	TP = 4	FP = 1
	生病	FN = 2	TN = 3

這時候雖然明眼看就知道診斷結果是醫生比較好，但還是需要換成指標來說明誰比誰好吧。

下圖基本上二元分類會用到的所有指標名稱和計算方式，基本上我列了所有會用到的指標:

		True Condition			
	Total Population (T)	Positive	Negative		
Predicted outcome	Positive	True Positive (TP)	False Positive (FP) Type I error	Positive predictive value (PPV), Precision $\frac{TP}{TP+FP}$	False discovery rate (FDR) $\frac{FP}{TP+FP}$
	Negative	False Negative (FN) Type II error	True Negative (TN)	False omission rate (FOR) $\frac{FN}{FN+TN}$	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
		True Positive Rate (TPR) Sensitivity, Recall $\frac{TP}{TP+FN}$	False Positive Rate (FPR) Fall-out $\frac{FP}{FP+TN}$	Positive likelihood ratio (LR+) =TPR/FPR Negative likelihood ratio (LR−) =FNR/TNR Diagnostic odds ratio (DOR) =LR+/LR− F_1-score=$2\frac{Precision \times Recall}{Precision+Recall}$ F_β-score $=(1+\beta^2)\frac{Precision \times Recall}{(\beta^2 Precision)+Recall}$ $=\frac{TP}{(1+\beta^2)TP+(\beta^2 FN)+FP}$ G-measure=$\sqrt{Precision \times Recall}$	
	Accuracy $\frac{TP+TN}{T}$	False Negative Rate (FNR) Miss rate $\frac{FN}{TP+FN}$	True negative rate (TNR) Specificity $\frac{TN}{FP+TN}$		

這邊我大概說一下常用的指標
Sensitivity 靈敏性: 也稱為True Positive Rate (TPR), Recall，「有病的偵測率」，所以是越高越好。
Specificity 特異性: 也稱為True negative rate (TNR)，「沒病的偵測率」，也是越高越好。
但基本上這兩個指標是trade off，這兩個指標跟等會要介紹的ROC有關係，也是臨床上非常長看的兩個指標。

Accuracy正確率: 基本上就是模型的整體判斷的正確率，所以有時候也稱為overall accuracy，越高越好。

False Negative Rate 偽陰性率: 預測模型判成沒病，但實際上有病的比率，越小越好。

False Positive Rate (FPR) 偽陽性率: 預測模型判成有病，但實際上沒有病的比率，越小越好。

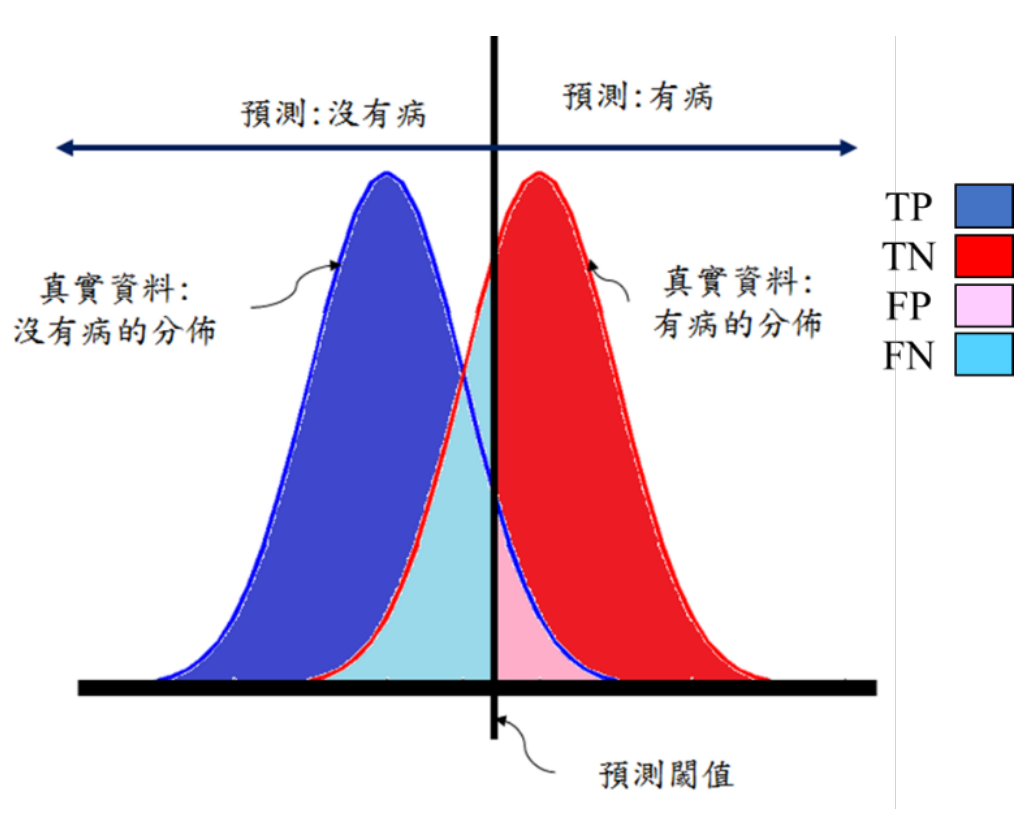
Positive predictive value (PPV) 陽性預測值: 也稱為Precision，在臨床上也是很常用的指標，模型診斷結果呈現有病且確實有病者的比率，越高越好。

Negative predictive value (NPV) 陰性預測值: 模型診斷結果呈沒病且實際上也沒有病的比率，越高越好。

剛剛的例子可以算出所有指標，最後得到下表，所以從大部份指標都可以知道醫師比較好。

	醫師診斷	電腦診斷
Sensitivity	100%	67%
Specificity	100%	75%
FNR	0%	33%
FPR	0%	25%
PPV	100%	80%
NPV	100%	60%
FDR	0%	20%
FOR	0%	40%
LR+	not computability	2.67
LR-	0.00	0.44
DOR	not computability	13.50
F1score	100%	73%
G-measure	100%	73%
accuracy	100%	70%

回到混淆矩陣我們可以知道，誤判的地方是FP和FN，下圖是用來解釋在混淆矩陣的東西。



在統計學上會將FP稱為「型一錯誤 (Type I error)，上圖淺藍色那塊」，FN稱為「型二錯誤 (Type II error)，上圖粉紅色那塊」

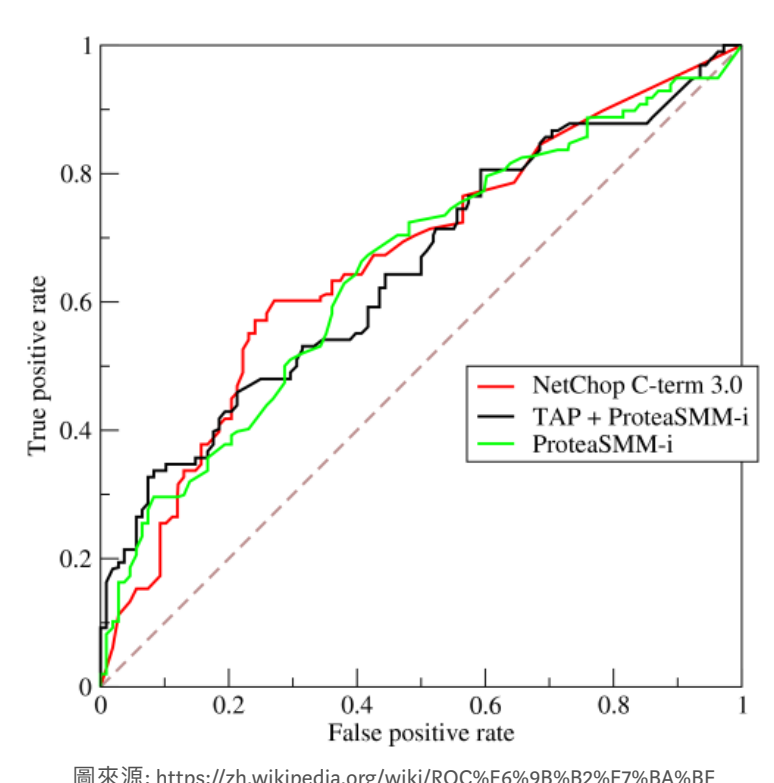
在假設檢定中(假設檢定的敘述可參考「統計學:大家都喜歡問的系列-p值是什麼」)，做的事情就是利用統計方法推測虛無假設(H0)是否成立，也就是拒絕或是不拒絕虛無假設。
若是虛無假設事實上成立，但檢定結果拒絕虛無假設時，這個錯誤就稱做為型一錯誤。
若是虛無假設事實上不成立，但檢定結果不拒絕虛無假設時，這個錯誤就稱做為型二錯誤。

> Note: *我念書的時候老師特別強調，統計檢定沒有「接受」這個選項，只有「不拒絕」，因為做檢定的目的是有沒有證據去「拒絕」一個假設，我們只能下結論是我們沒有統計上的證據去拒絕虛無假設。*

ROC曲線 (Receiver operating characteristic curve) & AUC (Area Under Curve)

ROC曲線也是用來評估二元分類問題的一個方法，基於上面的圖，我們可以知道閾值(threshold)在往左往右變化的時候代表的是，FP和FN是會變化的，這時候就會有trade off的問題，我前面有提到靈敏性和特異性兩個指標也是trade off的問題，原因是這兩個指標跟FP和FN也是直接相關，所以FP和FN變化基本上靈敏性和特異性也是跟著變，且FP和FN說的是個數，靈敏性和特異性才是百分比的指標。

簡單說一下ROC曲線怎麼來的，閾值(threshold)變化可以得到靈敏性和特異性指標，所以我們將所有可能的閾值(threshold)都去設定，然後可以跑出很多組靈敏性和特異性，一個靈敏性會對上一個特異性，因此把所有可能的連起來得到的就是ROC曲線了。

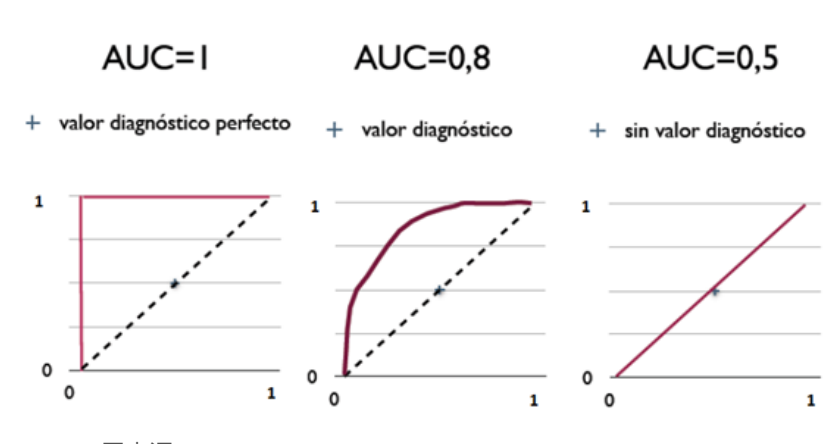


圖來源: https://zh.wikipedia.org/wiki/ROC%E6%9B%B2%E7%BA%BF
X軸為1-特異度(1-Specificity)，Y軸為敏感度(Sensitivity)

ROC曲線解讀方式，會以對角線為基準，若是算出來的ROC曲線等於對角線的話，代表你的模型完全沒有鑑別性(簡單說你的模型就沒啥屁用，回去重新training)，若ROC往左上角移動，代表模型對疾病的敏感度越高(偽陽性率(1-Specificity)越低)，代表模型的鑑別力越好。

一般在判別檢驗工具的好壞時，除了看曲線的圖形之外，也可以利用**曲線下的面積(Area Under Curve; AUC)**來判別ROC曲線的鑑別力，AUC數值的範圍從0到1，數值愈大愈好。以下為AUC數值一般的判別規則：
AUC=0.5 (no discrimination 無鑑別力)，**ROC剛好是對角線**。
0.7≦AUC≦0.8 (acceptable discrimination可接受的鑑別力)
0.8≦AUC≦0.9 (excellent discrimination優良的鑑別力)
0.9≦AUC≦1.0 (outstanding discrimination極佳的鑑別力)

所以AUC也可以當作比較不同模型之間比較的指標，AUC越大代表那個模型越好。



圖來源: https://zh.wikipedia.org/wiki/ROC%E6%9B%B2%E7%BA%BF

多元分類(multiclass case)指標

多元混淆矩陣(Confusion matrix)

基本上在二元的混淆矩陣以經將該說的都說了，這邊還有比較特殊的指標。

我們先舉個三個類別的混淆矩陣，如果有玩過UCI資料庫的/機器學習課程的人，應該都知道鳶尾花分類的問題，假設我的預測模型是SVM，所以我得到下面這個這麼好的結果，只有在Iris-versicolor這類別分錯了兩個樣本。

		Actual			
		Iris-setoda	Irir-versicolor	Iris-virginica	total
Predicted	Iris-setoda	50	0	0	50
	Irir-versicolor	0	48	0	48
	Iris-virginica	0	2	50	52
	error rate	0.00%	4.00%	0.00%	1.33%

多類別指標，基本上大概看三種，
第一種整體正確率/錯誤率
第二種單一類別的正確率/錯誤率
第三種看Cohen's kappa coefficient (Kappa)。
從上表可以看到整體錯誤率(2/150=1.33%)和單一類別錯誤率怎麼算的，但我這邊沒有提到Cohen's kappa coefficient。

Cohen's kappa coefficient是一種統計量化指標，平衡類別之間正確性的一種指標，簡單說就是要把**大者恆大的影響消除掉的指標**。什麼意思呢，假設我們今天有分類結果如下:

		Actual			
		1	2	3	total
Predicted	1	9999	20	50	10069
	2	0	0	0	0
	3	1	0	0	1
	error rate	0.01%	100.00%	0.00%	0.21%

第1類有1000個樣本，第二類只有20個，第三類只有50個，這時候只需要將資料都判給第1類，整體的正確率都很高/錯誤率很低，如果我們只看整體評估的指標，這時候只看整理正確率就會有問題，所以Kappa就是用來解決這件事情。

Kappa計算方式如下:

$$\kappa = \frac{(p_0 - p_c)}{(1 - p_c)}$$

首先我們要先將混淆矩陣的元素除上整體的樣本數，得到的是百分比的混淆矩陣。
$p0$就是百分比混淆矩陣的對角線元素相加。
pc就是把把每類別(實際第i類別的總和乘上預測第i類別的總和)做加總。

		Actual			
		1	2	3	sum
Predicted	1	99.29%	0.20%	0.50%	99.99%
	2	0.00%	0.00%	0.00%	0.00%
	3	0.01%	0.00%	0.00%	0.01%
	sum	99.30%	0.20%	0.50%	

$$p_0 = 99.29\% + 0.00\% + 0.00\% = 99.29\%$$

$$p_c = 99.30\% * 99.99\% + 0.00\% * 0.20\% + 0.01\% * 0.50\% = 99.30\%$$

$$\kappa = \frac{(p_0 - p_c)}{(1 - p_c)} = -0.00016787$$

所以這個時候Kappa只剩下-0.00016787，非常的差。

> *Note: Kappa是介於-1~1之間的數字，值正越大，代表模型越好。*

· · ·

回歸指標(Regression metrics)

回歸的部份比較沒有什麼好講的，因為回歸的指標通常都只看平均均方誤差(mean square error)，但如果是回歸的損失函數(loss function)那就很有趣，會有更多東西可以說，但這篇主要是說validation index，所以就不提損失函數的部份。

為什麼我說回歸沒什麼好講，主要原因是
「回歸是做預測一個連續的值，這時候我們只希望預測的值跟實際上的值越接近越好」
這是什麼意思呢?
假設我們做出一個模型預測小明的身高是180公分($\hat{y}$)，實際上小明是160公分(y)，這時候的誤差是20公分，我們都會希望**誤差越小越好**，所以回歸基本上評估的指標都是基於「y-$\hat{y}$」
比如
1. 平均均方誤差(Mean Squared Error, MSE)

$$MSE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

2. 平均絕對誤差(Mean Absolute Error, MAE)

$$MAE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

3. 平均均方對數誤差(Mean Squared Logarithmic Error, MSLE)

$$MSLE(y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left(\ln(1+y_i) - \ln(1+\hat{y}_i)\right)^2$$

所以在回歸基本上只要評估時，評估指標只要用一樣的，**哪個模型的評估指標越小的就代表那個模型越好**。

Tommy Huang Follow
怕老了忘記這些吃飯的知識，開始寫文章記錄機器/深度學習相關內容。
Chih-Sheng Huang, mail: chih.sheng.huang821@gmail.com