

機器學習秘訣和技巧參考手冊

Afshine AMIDI 和 Shervine AMIDI

December 15, 2019

翻譯: *kevingo*. 審閱: *kentropy*.

分類器的評估指標

在二元分類的問題上，底下是主要用來衡量模型表現的指標

□ **混淆矩陣** – 混淆矩陣是用來衡量模型整體表現的指標

		預測類別	
		+	-
真實類別	+	TP True Positives	FN False Negatives Type II error
	-	FP False Positives Type I error	TN True Negatives

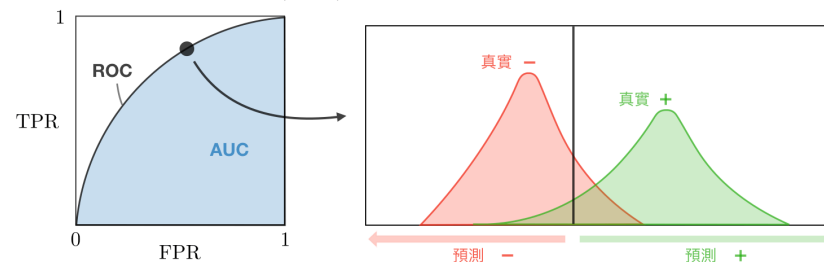
□ **主要的衡量指標** – 底下的指標經常用在評估分類模型的表現

指標	公式	解釋
準確度	$\frac{TP + TN}{TP + TN + FP + FN}$	模型的整體表現
Precision	$\frac{TP}{TP + FP}$	預測的類別有多精準的比例
Recall Sensitivity	$\frac{TP}{TP + FN}$	實際正的樣本的覆蓋率有多少
Specificity	$\frac{TN}{TN + FP}$	實際負的樣本的覆蓋率
F1分數	$\frac{2TP}{2TP + FP + FN}$	對於非平衡類別相當有用的混合指標

□ **ROC** – 接收者操作特徵曲線(ROC Curve)，又被稱為ROC，是透過改變閾值來表示TPR 和FPR 之間關係的圖形。這些指標總結如下：

衡量指標	公式	等同於
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ **AUC** – 在接收者操作特徵曲線(ROC) 底下的面積，也稱為AUC 或AUROC：



回歸器的評估指標

□ **基本的指標** – 給定一個迴歸模型 f ，底下是經常用來評估此模型的指標：

總平方和	被解釋平方和	殘差平方和
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ **決定係數** – 決定係數又被稱為 R^2 or r^2 ，它提供了模型是否具備復現觀測結果的能力。定義如下：

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

□ **主要的衡量指標** – 藉由考量變數 n 的數量，我們經常用使用底下的指標來衡量迴歸模型的表現：

Mallow's Cp	AIC	BIC	Adjusted R^2
$\frac{SS_{\text{res}} + 2(n+1)\hat{\sigma}^2}{m}$	$2[(n+2) - \log(L)]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1-R^2)(m-1)}{m-n-1}$

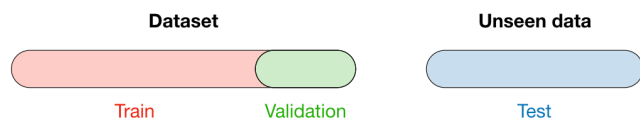
當中， L 代表的是概似估計， $\hat{\sigma}^2$ 則是變異數的估計

模型選擇

□ **詞彙** – 當進行模型選擇時，我們會針對資料進行以下區分：

訓練資料集	驗證資料集	測試資料集
<ul style="list-style-type: none"> - 用來訓練模型 - 通常是80 	<ul style="list-style-type: none"> - 用來評估模型 - 又被稱為hold-out 資料集或開發資料集 	<ul style="list-style-type: none"> - 模型用來預測用的資料集

當模型被選擇後，就會使用整個資料集來做訓練，並且在沒看過的資料集上做測試。你可以參考以下的圖表：



□ **交叉驗證** – 交叉驗證，又稱之為CV，它是一種不特別依賴初始訓練集來挑選模型的方法。幾種不同的方法如下：

k -fold	Leave- p -out
<ul style="list-style-type: none"> - 把資料分成k份，利用$k-1$份資料來訓練，剩下的一份用來評估模型效能 - 一般來說$k=5$或10 	<ul style="list-style-type: none"> - 在$n-p$份資料上進行訓練，剩下的p份資料用來評估模型效能 - 當$p=1$時，又稱為leave-one-out

最常用到的方法叫做 k -fold 交叉驗證。它將訓練資料切成 k 份，在 $k-1$ 份資料上進行訓練，而剩下的一份用來評估模型的效能，這樣的流程會重複 k 次。最後計算出來的模型損失是 k 次結果的平均，又稱為交叉驗證損失值。

Fold	Dataset	Validation error	Cross-validation error
1		ϵ_1	$\frac{\epsilon_1 + \dots + \epsilon_k}{k}$
2		ϵ_2	
\vdots	\vdots	\vdots	
k		ϵ_k	
	Train Validation		

□ **正規化** – 正歸化的目的是為了避免模型對於訓練資料過擬合，進而導致高方差。底下的表格整理了常見的正規化技巧：

LASSO	Ridge	Elastic Net
<ul style="list-style-type: none"> - 將係數縮減為0 - 有利變數的選擇 	將係數變得更小	在變數的選擇和小係數之間作權衡
$\dots + \lambda \ \theta\ _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda \ \theta\ _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \left[(1-\alpha) \ \theta\ _1 + \alpha \ \theta\ _2^2 \right]$ $\lambda \in \mathbb{R}, \alpha \in [0,1]$

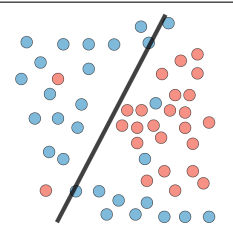
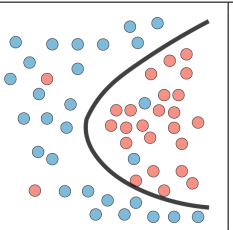
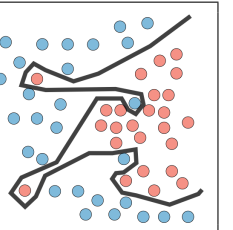
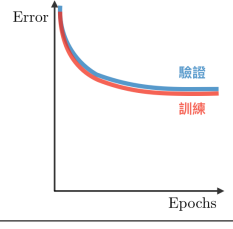
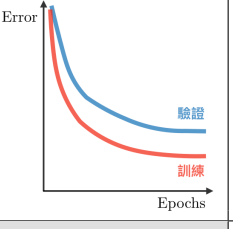
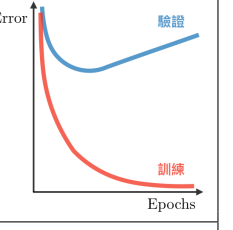
診斷

□ **偏差** – 模型的偏差指的是模型預測值與實際值之間的差異

□ **變異** – 變異指的是模型在預測資料時的變異程度

□ **偏差/變異的權衡** – 越簡單的模型，偏差就越大。而越複雜的模型，變異就越大

	Underfitting	Just right	Overfitting
現象	<ul style="list-style-type: none"> - 訓練錯誤較高 - 訓練錯誤和測試錯誤接近 - 高偏差 	<ul style="list-style-type: none"> - 訓練誤差會稍微比測試誤差低 	<ul style="list-style-type: none"> - 訓練誤差很低 - 訓練誤差比測試誤差低很多 - 高變異
迴歸圖示			

分類圖示			
深度學習圖示			
可能的解法	<ul style="list-style-type: none"> - 使用較複雜的模型 - 增加更多特徵 - 訓練更久 		<ul style="list-style-type: none"> - 採用正規化化的方法 - 取得更多資料

□ **誤差分析(Error analysis)** – 誤差分析指的是分析目前使用的模型和最佳模型之間差距的根本原因

□ **銷蝕分析(Ablative analysis)** – 銷蝕分析指的是分析目前模型和基準模型之間差異的根本原因