

VIP Cheatsheet: Mẹo và thủ thuật

Afshine AMIDI và Shervine AMIDI

Ngày 17 tháng 5 năm 2020

Dịch bởi Trần Tuấn Anh, Nguyễn Trí Minh, Vinh Pham và Đàm Minh Tiến

Độ đo phân loại

Đối với phân loại nhị phân (binary classification) là các độ đo chính, chúng khá quan trọng để theo dõi (track), qua đó đánh giá hiệu năng của mô hình (model).

□ **Ma trận nhầm lẫn (Confusion matrix)** – Confusion matrix được sử dụng để có kết quả hoàn chỉnh hơn khi đánh giá hiệu năng của model. Nó được định nghĩa như sau:

		Lớp dự đoán	
		+	-
lớp thực sự	+	TP True Positives	FN False Negatives Type II error
	-	FP False Positives Type I error	TN True Negatives

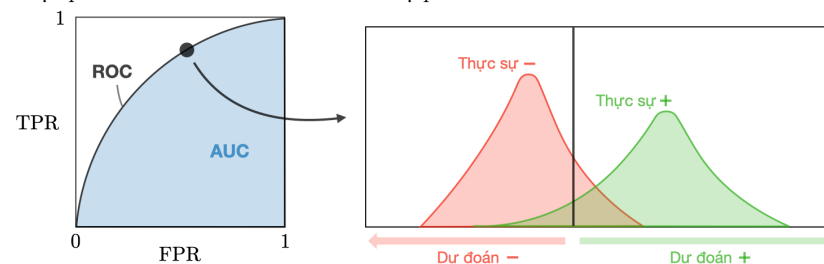
□ **Độ đo chính** – Các độ đo sau thường được sử dụng để đánh giá hiệu năng của mô hình phân loại:

Độ đo	Công thức	Diễn giải
chính xác	$\frac{TP + TN}{TP + TN + FP + FN}$	Hiệu năng tổng thể của mô hình
Precision	$\frac{TP}{TP + FP}$	Độ chính xác của các dự đoán positive
Recall Sensitivity	$\frac{TP}{TP + FN}$	Bao phủ các mẫu thử chính xác (positive) thực sự
Specificity	$\frac{TN}{TN + FP}$	Bao phủ các mẫu thử sai (negative) thực sự
Điểm F1	$\frac{2TP}{2TP + FP + FN}$	Độ đo Hybrid hữu ích cho các lớp không cân bằng (unbalanced classes)

□ **ROC** – Đường cong thao tác nhận, được kí hiệu là ROC, là minh họa của TPR với FPR bằng việc thay đổi ngưỡng (threshold). Các độ đo này được tổng kết ở bảng bên dưới:

Độ đo	Công thức	Tương đương
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ **AUC** – Khu vực phía dưới đường cong thao tác nhận, còn được gọi tắt là AUC hoặc AUROC, là khu vực phía dưới ROC như hình minh họa phía dưới:



Độ đo hồi quy

□ **Độ đo cơ bản** – Cho trước mô hình hồi quy f , độ đo sau được sử dụng phổ biến để đánh giá hiệu năng của mô hình:

Tổng của tổng các bình phương	Mô hình tổng bình phương	Tổng bình phương dư
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ **Hệ số quyết định** – Hệ số quyết định, thường được kí hiệu là R^2 hoặc r^2 , cung cấp độ đo mức độ tốt của kết quả quan sát đầu ra (được nhân rộng bởi mô hình), và được định nghĩa như sau:

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

□ **Độ đo chính** – Độ đo sau đây thường được sử dụng để đánh giá hiệu năng của mô hình hồi quy, bằng cách tính số lượng các biến n mà độ đo đó sẽ cân nhắc:

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

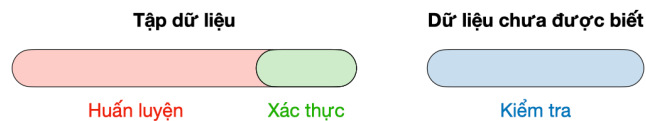
trong đó L là khả năng và $\hat{\sigma}^2$ là giá trị ước tính của phương sai tương ứng với mỗi response (hồi đáp).

Lựa chọn model (mô hình)

□ **Vocabulary** – Khi lựa chọn mô hình, chúng ta chia tập dữ liệu thành 3 tập con như sau:

Tập huấn luyện	Tập xác thực	Tập kiểm tra (testing)
<ul style="list-style-type: none"> Mô hình được huấn luyện Thường là 80% tập dữ liệu 	<ul style="list-style-type: none"> mô hình được xác thực Thường là 20% tập dữ liệu Cũng được gọi là hold-out hoặc development set (tập phát triển) 	<ul style="list-style-type: none"> mô hình đưa ra dự đoán Dữ liệu chưa được biết

Khi mô hình đã được chọn, nó sẽ được huấn luyện trên tập dữ liệu đầu vào và được test trên tập dữ liệu test hoàn toàn khác. Tất cả được minh hoạ ở hình bên dưới:



□ **Cross-validation** – Cross-validation, còn được gọi là CV, một phương thức được sử dụng để chọn ra một mô hình không dựa quá nhiều vào tập dữ liệu huấn luyện ban đầu. Các loại khác nhau được tổng kết ở bảng bên dưới:

<i>k</i> -fold	Leave- <i>p</i> -out
<ul style="list-style-type: none"> Huấn luyện trên $k - 1$ phần và đánh giá trên 1 phần còn lại Thường thì $k = 5$ hoặc 10 	<ul style="list-style-type: none"> Huấn luyện trên $n - p$ phần và đánh giá trên p phần còn lại Trường hợp $p = 1$ được gọi là leave-one-out

Phương thức hay được sử dụng được gọi là *k*-fold cross-validation và chia dữ liệu huấn luyện thành k phần, đánh giá mô hình trên 1 phần trong khi huấn luyện mô hình trên $k - 1$ phần còn lại, tất cả k lần. Lỗi sau đó được tính trung bình trên k phần và được đặt tên là cross-validation error.

Phần	Tập dữ liệu	Lỗi xác thực	Lỗi cross-xác thực
1		ϵ_1	$\frac{\epsilon_1 + \dots + \epsilon_k}{k}$
2		ϵ_2	
\vdots	\vdots	\vdots	
k		ϵ_k	
	Huấn luyện Xác thực		

□ **Chuẩn hoá** – Mục đích của thủ tục chuẩn hoá là tránh cho mô hình bị overfit với dữ liệu, do đó gặp phải vấn đề phương sai lớn. Bảng sau đây sẽ tổng kết các loại kĩ thuật chuẩn hoá khác nhau hay được sử dụng:

LASSO	Ridge	Elastic Net
<ul style="list-style-type: none"> Giảm hệ số xuống còn 0 Tốt cho việc lựa chọn biến 	Làm cho hệ số nhỏ hơn	Thay đổi giữa chọn biến và hệ số nhỏ hơn
$\dots + \lambda \theta _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda \theta _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \left[(1 - \alpha) \theta _1 + \alpha \theta _2^2 \right]$ $\lambda \in \mathbb{R}, \alpha \in [0, 1]$

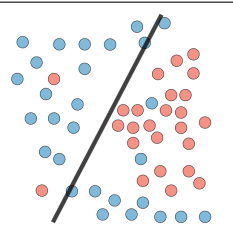
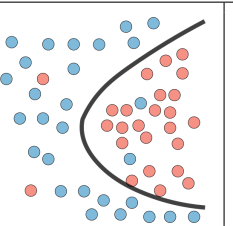
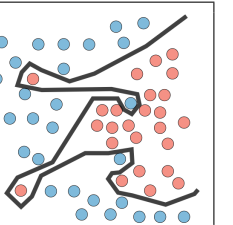
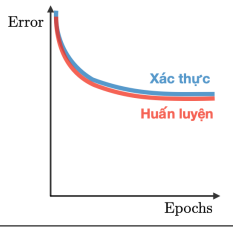
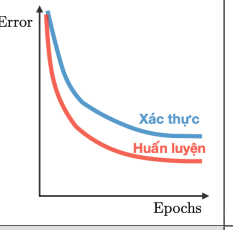
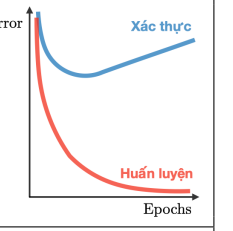
Dự đoán

□ **Bias** – Bias của mô hình là sai số giữa dự đoán mong đợi và dự đoán của mô hình trên các điểm dữ liệu cho trước.

□ **Phương sai** – Phương sai của một mô hình là sự thay đổi dự đoán của mô hình trên các điểm dữ liệu cho trước.

□ **Sự đánh đổi bias/phương sai** – Mô hình càng đơn giản bias càng lớn, mô hình càng phức tạp phương sai càng cao.

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"> Lỗi huấn luyện cao Lỗi huấn luyện tiến gần tới lỗi test Bias cao 	<ul style="list-style-type: none"> Lỗi huấn luyện thấp hơn một chút so với lỗi test 	<ul style="list-style-type: none"> Lỗi huấn luyện rất thấp Lỗi huấn luyện thấp hơn lỗi test rất nhiều Phương sai cao
Minh hoạ hồi quy			

Mình hoạ phân loại			
Mình hoạ deep learning (học sâu)			
Biện pháp khắc phục có thể dùng	<ul style="list-style-type: none"> - Mô hình phức tạp - Thêm nhiều đặc trưng - Huấn luyện lâu hơn 		<ul style="list-style-type: none"> - Thực hiện chuẩn hóa - Lấy nhiều dữ liệu hơn

□ **Phân tích lỗi** – Phân tích lỗi là phân tích nguyên nhân của sự khác biệt trong hiệu năng giữa mô hình hiện tại và mô hình lý tưởng.

□ **Phân tích Ablative** – Phân tích Ablative là phân tích nguyên nhân của sự khác biệt giữa hiệu năng của mô hình hiện tại và mô hình cơ sở.