

VIP Cheatsheet: Machine Learning Tips

Afshine AMIDI and Shervine AMIDI

October 13, 2018

翻译: *hujinsen*. 由 *spin6lock* 审阅.

分类问题的度量

在二分类问题中, 下面这些主要度量标准对于评估模型的性能非常重要。

□ **混淆矩阵** – 混淆矩阵可以用来评估模型的整体性能情况。它的定义如下:

		预测类别	
		+	-
实际类别	+	TP True Positives	FN False Negatives Type II error
	-	FP False Positives Type I error	TN True Negatives

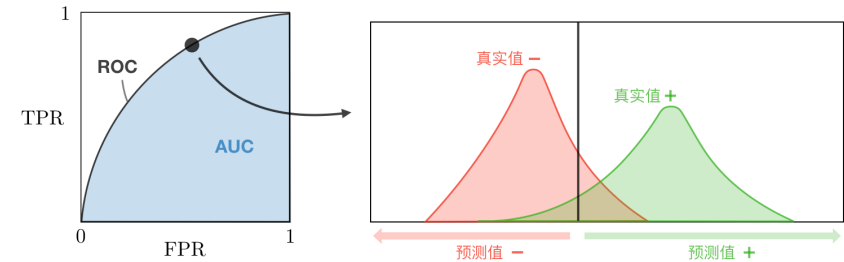
□ **主要度量标准** – 通常用下面的度量标准来评估分类模型的性能:

性能度量	公式	说明
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	模型总体性能
Precision	$\frac{TP}{TP + FP}$	预测为正样本的准确度
Recall Sensitivity	$\frac{TP}{TP + FN}$	真正样本的覆盖度
Specificity	$\frac{TN}{TN + FP}$	真负样本的覆盖度
F1 score	$\frac{2TP}{2TP + FP + FN}$	混合度量, 对于不平衡类别非常有效

□ **ROC** – 受试者工作曲线, 又叫做ROC曲线, 它使用真正例率和假正例率分别作为纵轴和横轴并且经过调整阈值绘制出来。下表汇总了这些度量标准:

性能度量	公式	等价形式
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ **AUC** – 受试者工作曲线的之下的部分, 又叫做AUC或者AUROC, 如下图所示ROC曲线下的部分:



回归指标

□ **基本性能度量** – 给定一个回归模型 f , 下面的度量标准通常用来评估模型的性能

全部平方和	解释平方和	残差平方和
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ **确定性系数** – 确定性系数, 记作 R^2 或 r^2 , 提供了模型复现观测结果的能力, 定义如下:

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

□ **主要性能度量** – 以下性能度量通过考虑变量 n 的数量, 常用于评估回归模型的性能:

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

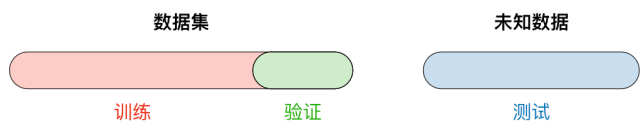
L 代表似然, $\hat{\sigma}^2$ 代表方差

模型选择

□ **词汇** – 在选择模型时，我们将数据分为3个不同部分：

训练集	验证集	测试集
<ul style="list-style-type: none"> - 模型训练 - 一般数据集中的80 	<ul style="list-style-type: none"> - 模型评估 - 一般数据集中的20 - 又叫做留出集或者开发集 	<ul style="list-style-type: none"> - 模型预测 - 未知数据

一旦选择了模型，就会在整个数据集上进行训练，并在测试集上进行测试。如下图所示：



□ **交叉验证** – 交叉验证，记为CV，是一种不必特别依赖于初始训练集的模型选择方法。下表汇总了几种不同的方式：

k -fold	Leave- p -out
<ul style="list-style-type: none"> - 在 $k - 1$ 个子集上训练，在剩余的一个子集中评估 - 通常 $k = 5$ 或 10 	<ul style="list-style-type: none"> - 在 $n - p$ 个子集上训练，在剩余的 p 个子集评估模型 - $p = 1$ 时又叫做留一法

最常用的模型选择方法是 k 折交叉验证，将训练集划分为 k 个子集，在 $k - 1$ 个子集上训练模型，在剩余的一个子集上评估模型，用这种划分方式重复训练 k 次。交叉验证损失是 k 次 k 折交叉验证的损失均值。

子集	数据集	验证错误	交叉验证错误
1		ϵ_1	$\frac{\epsilon_1 + \dots + \epsilon_k}{k}$
2		ϵ_2	
\vdots	\vdots	\vdots	
k		ϵ_k	
	训练 验证		

□ **正则化** – 正则化方法可以解决高方差问题，避免模型对于训练数据产生过拟合。下表展示了常用的正则化方法：

LASSO	Ridge	Elastic Net
<ul style="list-style-type: none"> - 将系数收缩为0 - 有利于变量选择 	使系数更小	在变量选择和小系数之间进行权衡
$\dots + \lambda \ \theta\ _1$ $\lambda \in \mathbb{R}$	$\dots + \lambda \ \theta\ _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \left[(1 - \alpha) \ \theta\ _1 + \alpha \ \theta\ _2^2 \right]$ $\lambda \in \mathbb{R}, \alpha \in [0, 1]$

诊断

□ **偏差** – 模型的偏差是模型预测值和真实值之间的差距

□ **方差** – 模型的方差是给定数据点的模型预测的可变性

□ **偏差/方差权衡** – 模型越简单，偏差越高，模型越复杂，方差越高。

	Underfitting	Just right	Overfitting
症状	<ul style="list-style-type: none"> - 高训练误差 训练误差接近测试误差 - 高偏差 	<ul style="list-style-type: none"> - 训练误差略低于测试误差 	<ul style="list-style-type: none"> - 极低训练误差 训练误差远低于测试误差 - 高方差
回归图			

分类图			
深度学习插图			
可能的补救措施	<ul style="list-style-type: none"> - 模型复杂性 - 添加更多特征 - 训练更长时间 		<ul style="list-style-type: none"> - 实施正则化 - 获得更多数据

❑ **错误分析** – 错误分析分析当前模型和完美模型之间性能差异的根本原因

❑ **烧蚀分析** – 烧蚀分析可以分析当前和基线模型之间性能差异的根本原因