机器学习模型性能评估（三）：代价曲线

原创 石头 机器学习算法那些事 2018-09-24

上文详细介绍了P-R曲线与ROC曲线的性能评估方法，P-R曲线与ROC曲线是基于均等损失代价的模型性能评估方法，本文承接上文，详细介绍基于非均等损失代价的模型性能评估方法，并对已介绍的性能评估方法进行了总结。

2.4 代价敏感错误率

前面介绍的性能评估方法都默认了一种情况，即学习模型把正样本误分为负样本的损失代价与负样本误分为正样本的损失代价相同，实际情况可能是不同类型的分类错误造成的损失代价不同。

比如医疗诊断中，若一个正常人误诊断为癌症病人，那么会给测试者造成了很大的心理压力以及花费很长时间去做进一步的检查；若一个癌症病人误诊断为正常人，那么病人很可能因此丧失生命，显然，癌症病人误诊断为正常人的损失代价要大的多。

若在评价学习模型性能时考虑不同类分类错误所造成不同损失代价的因素时，称为代价敏感错误率评估方法。

假设训练数据集D包含正例子集D + 和负例子集D - 。若分类错误的损失代价相同，即均等代价。代价矩阵如下：

真实类别	预测类别	
	正类	负类
正类	0	1
负类	1	0

模型的代价敏感错误率：

$$E(f;D)=\frac{1}{m}(\sum_{x_i\in D^+} II(f(x_i)\neq y_i)\times 1+\sum_{x_i\in D^-} II(f(x_i)\neq y_i)\times 1)$$

$II(\bullet)$是指示函数，若$\bullet$为真则为1，否则为0

分类错误的损失代价不相同，即非均等代价。如下图，其中costij表示将第i类样本预测为第j类样本的代价。

若预测正确，则代价为0，即costii = 0。一般用1代表正类，0代表负类。

代价矩阵如下：

真实类别	预测类别	
	正类	负类
正类	0	cost10
负类	cost01	0

模型的代价敏感错误率：

$$E(f;D;\cos t)=\frac{1}{m}(\sum_{x_i\in D^+} II(f(x_i)\neq y_i)\times \cos t_{10}+\sum_{x_i\in D^-} II(f(x_i)\neq y_i)\times \cos t_{01})$$

$II(\bullet)$是指示函数，若$\bullet$为真则为1，否则为0

该定义是计算模型损失代价的一般化，若分类错误的损失代价相同，令cost10 = cost01 = 1，则与之前的代价敏感错误率表达式一致。

不管是使用何种损失代价函数，构建模型最优化都等价于最小化代价敏感错误率。

2.5 代价曲线

ROC曲线在均等代价（分类错误的损失代价相同）的前提下反映学习模型的泛化能力，“代价曲线”是在非均等代价的前提下反映学习模型的期望总体代价。期望总体代价越小，则模型的泛化能力越强。

代价曲线的横坐标是样例为归一化的正例概率代价，正例概率为p，给定的正例概率为先验概率，范围为0~1，纵轴是归一化的损失代价。代价曲线研究的是正样本先验概率和损失代价的关系。

归一化的正例概率代价：

$$P(+)\cos t = \frac{p \times \cos t_{10}}{p \times \cos t_{10} + (1-p) \times \cos t_{01}}$$

这个表达式与周老师《机器学习》稍微有点不同，但含义是相同。习惯用1代表正样本，0代表负样本。

正例概率代价：

$$p \times \cos t_{10}$$

总概率代价：

$$p \times \cos t_{10} + (1-p) \times \cos t_{01}$$

因此，经过归一化的正例概率代价的表达式，归一化的作用是使正例概率代价的取值范围为[0,1]。

模型归一化代价：

$$\cos t_{norm} = \frac{FNR \times p \times \cos t_{10} + FPR \times (1-p) \times \cos t_{01}}{p \times \cos t_{10} + (1-p) \times \cos t_{01}}$$

模型总概率代价：

$$FNR \times p \times \cos t_{10} + FPR \times (1-p) \times \cos t_{01}$$

模型总概率代价最大值：

$$p \times \cos t_{10} + (1-p) \times \cos t_{01}$$

$$\text{其中} FNR = 1,\ FPR = 1$$

周老师《机器学习》对模型归一化代价描述的比较精炼，可能很多童鞋不是很清楚具体推导细节，本节详细的解释了模型代价和总体期望代价的推导过程。

模型代价和期望总体代价

模型归一化代价的分母项是归一化项，分析模型归一化代价的时候，只考虑分子项，归一化项可理解为常数，不影响模型的理解。

假设正样本的先验概率：p

模型代价：

$$FNR \times p \times \cos t_{10} + FPR \times (1-p) \times \cos t_{01}$$

二分类问题中，ROC曲线的每个点对应一个分类器，分类器可用阈值表示为：

$$\{x \mid f(x) \geq \eta\} \in H1,\ \{x \mid f(x) < \eta\} \in H0$$

x为测试样本，学习模型为f，当学习模型预测测试样本为正样例的概率大于等于η，则测试样本的预测结果为正例，即H1；反之，预测结果为负样本，即H0。

FNR：假负例率，即已知测试样本是正样本，测试结果是负样本的概率。

$$\{x \mid f(x) \geq \eta\} \in H1,\ \{x \mid f(x) < \eta\} \in H0$$

FNR为条件概率。

TNR，FPR，TPR的表达式也可以通过条件概率的形式给出：

$$TNR = \Pr[H0 \mid H0],\ TNR为正负例率$$
$$FPR = \Pr[H1 \mid H0],\ FPR为假正例率$$
$$TPR = \Pr[H1 \mid H1],\ TPR为真正例率$$

正例先验概率p，负例的先验概率（1-p）。结合正例先验概率，那么当输入为正例，测试结果为负例的联合概率分布（joint probability distribution）。

$$\Pr(H1, H0) = \Pr(H1) * \Pr(H0 \mid H1) = p * FNR$$

其他联合概率分布也可推导：

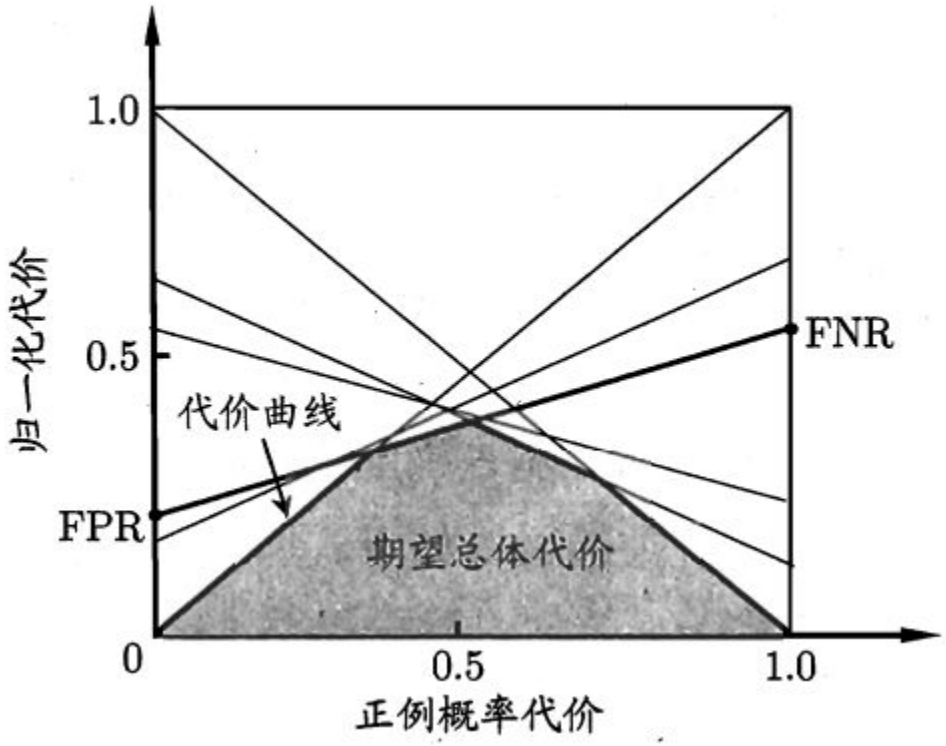
$$\Pr(H1, H1) = \Pr(H1) * \Pr(H1 \mid H1) = p * TPR$$
$$\Pr(H0, H1) = \Pr(H0) * \Pr(H1 \mid H0) = (1-p) * FPR$$
$$\Pr(H0, H0) = \Pr(H0) * \Pr(H0 \mid H0) = (1-p) * TNR$$

所以模型的期望损失代价：

$$E[Cost] = \sum_{i=0}^{1}\sum_{j=0}^{1} c_{ij} \Pr(Hi, Hj) = \sum_{i=0}^{1}\sum_{j=0}^{1} c_{ij} \Pr(Hj \mid Hi) * \Pr(Hi)$$
$$E[Cost] = c_{00} \Pr(H0 \mid H0) * \Pr(H0) + c_{01} \Pr(H1 \mid H0) * \Pr(H0) + c_{10} \Pr(H0 \mid H1) * \Pr(H1) + c_{11} \Pr(H1 \mid H1) * \Pr(H1)$$
$$E[Cost] = c_{00} * TNR * (1-p) + c_{01} * FPR * (1-p) + c_{10} * FNR * p + c_{11} * TPR * p$$
$$\because c_{00} = 0, c_{11} = 0$$
$$\therefore E[Cost] = c_{01} * FPR * (1-p) + c_{10} * FNR * p$$
$$c_{01} = cost_{01}, c_{10} = cost_{10}$$
$$\therefore E[Cost] = cost_{01} * FPR * (1-p) + cost_{10} * FNR * p$$

期望损失代价是由正例先验概率和混淆矩阵决定的，当ROC曲线的阈值确定时，FPR和FNR也相应的确定。因此，代价曲线是一条直线，不同的代价曲对应不同的分类器。

因此，代价曲线如下图：



其中，灰色的阴影部分为模型的期望总体代价，期望总体代价越小，则模型的泛化性能越好；反之，模型泛化性能越差。

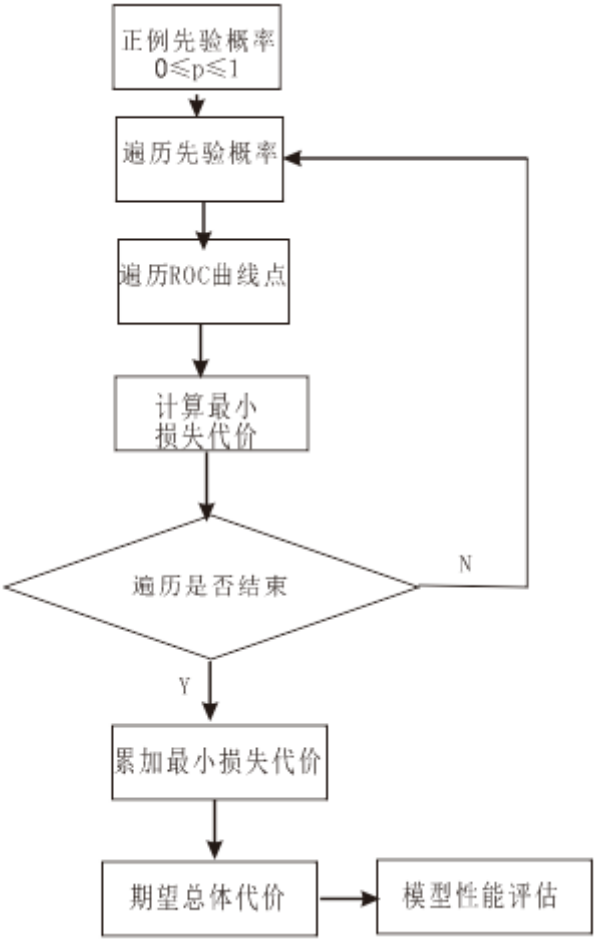
期望总体代价的意义：正例先验概率下学习模型的最小损失代价，并对所有正例先验概率下的最小损失代价求和。

期望总体代价计算步骤：

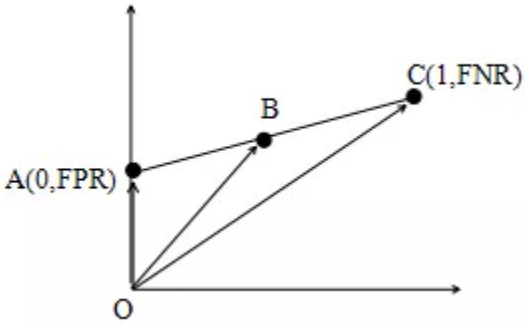
- 取ROC曲线上的每一点，计算相应的FNR，FPR。FNR和FPR在已知的情况下，归一化代价是一条随正例概率变化的直线。因此，代价曲线是一条直线，ROC曲线每个点都对应一条代价曲线。
- 对于给定的正例概率p，取所有代价曲线在该频率下的最小值，然后求该最小值与正例概率所围成的面积，得到如上图所示的阴影曲线。

$$\text{期望总体代价} = \sum_{p=0}^{p=1} \mathrm{p} * \cos t_{(\min, p)}$$

期望总体代价流程图：



证明：在FPR和FRN已知的情况下，代价曲线是一条直线



代价曲线方程：

$$E[Cost] = \cos t_{01} * FPR * (1-p) + \cos t_{10} * FNR * p$$

$\cos t_{01}, \cos t_{10}$可认为是常数

等价于：

$$\overline{OC} = \overline{OA} * (1-\mathrm{p}) + \overline{OB} * p, \quad p \neq 0$$

证明：A, B, C三点共线

$$\overline{OC} = \overline{OA} * (1-\mathrm{p}) + \overline{OB} * p = \overline{OA} - \mathrm{p} * \overline{OA} + \overline{OB} * p$$

$$\Rightarrow \dot{OC} - \dot{OA} = p * (\dot{OB} - \dot{OA})$$

$$\Rightarrow \dot{AC} = p * \dot{AB}$$

$$\because p \neq 0$$

$$\therefore A, B, C\text{三点共线}$$

3、 总结

本文首先介绍了期望和均值的概念，均值是抽样数据的统计量，期望是总体分布的统计量，由大数定律可知，当抽样数据无穷大时，样本均值等于总体期望。

其次介绍了机器学习模型性能评估方法，评价机器学习模型性能的金标准是模型的泛化能力。

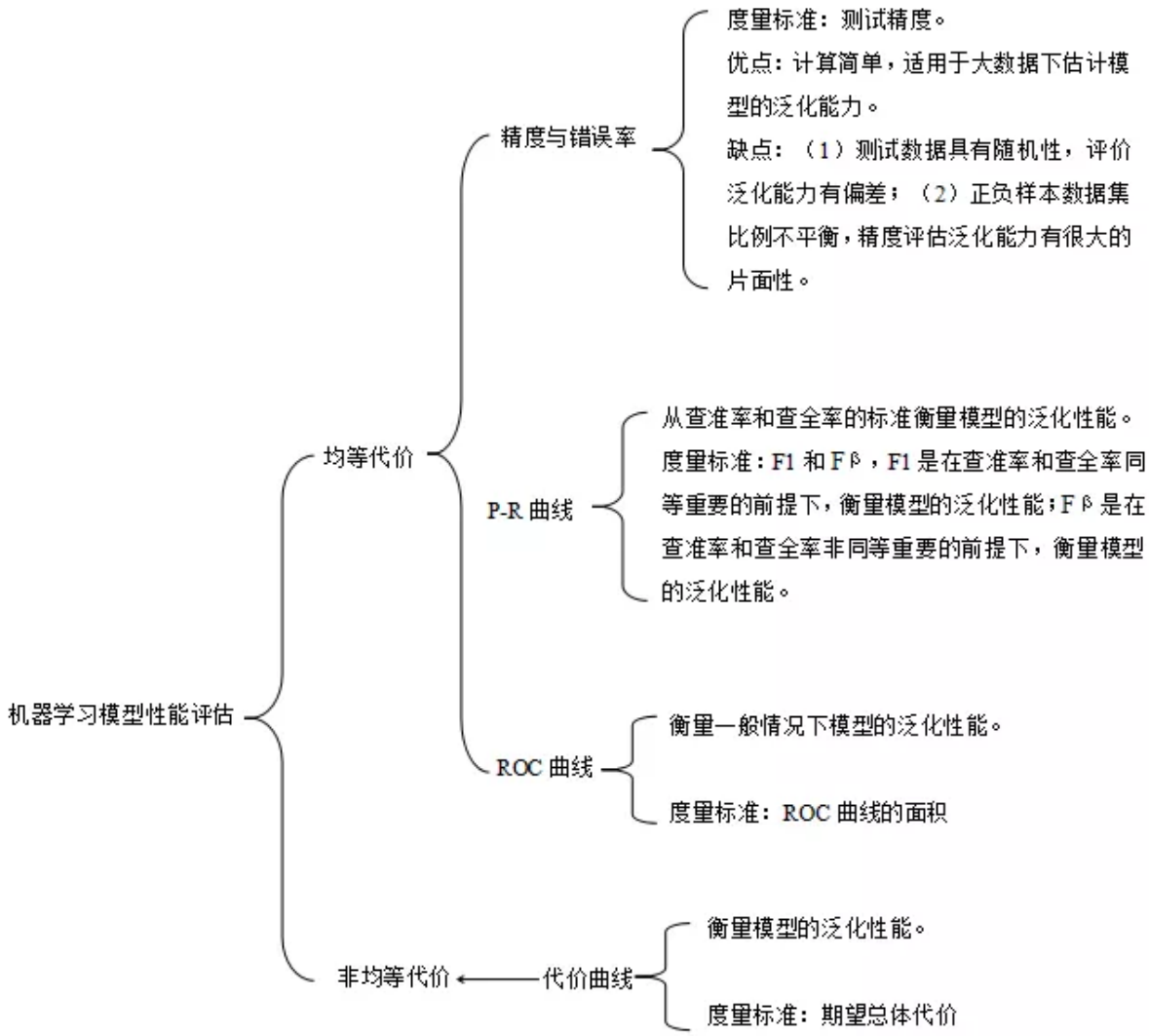
常用测试样本的精度来评价模型的泛化能力，这样做的缺点在于：

（1）测试样本具有随机性，不同测试样本的精度很可能不一样，评价泛化能力存在偏差；
（2）若测试样本的正负样本数据集比例不平衡，用精度评价泛化能力存在很大的缺陷。

若模型的实际场景应用需要重点考虑查准率和查全率的因素，P-R曲线是通过查准率和查全率的角度来评价模型的性能，可通过Fβ来度量模型性能；

ROC曲线是无任何限制条件下评价模型泛化性能的好坏，度量标准是面积大小；若需要考虑分类模型产生的非均等损失代价问题，可通过代价曲线的期望总体代价作为度量标准。

机器学习模型性能评估框架：



END