第二章 模型评估与选择

王照国

2019 年 3 月 21 日

1 NP问题

指的是其解可以在多项式时间内被验证的问题集合，P是否等于NP是指“如果一个问题能够在多项式时间内被验证，那么是否可以在多项式时间内找到这个问题的解”

NP问题的典型例子：一个物流配送公司欲将N个客户的订货沿最短路线全部送到，那么它应该如何确定最短路线？对于这一问题，P=NP意味着这样的物流分配可以很快地进行，但反之则意味着当物流规模逐渐扩大时，我们将无法在有效时间内找到最短路线。

2 模型评估方法

2.1 留出法

可以直接将数据集D划分为两个互斥的集合，其中一个作为训练集S，另外一个作为测试集T，即$\mathrm{D} = \mathrm{S} \cup \mathrm{T}, \mathrm{S} \cap \mathrm{T} = \emptyset$

2.2 交叉验证法

交叉验证法是将数据集D划分为k个大小相等的互斥子集，即$D_1 \cup D_2 ... \cup D_k, D_i \cap D_{i+1} = \emptyset$。每个子集尽可能保持数据分布的一致性，然后每次用k-1个子集的并集作为训练集，剩下的一个子集作为测试集，最终返回的是这k个训练结果的平均值

2.3 自助法(bootstrapping)

给定包含m个样本的数据集D，我们对它进行采样产生数据集D'，每次我们从数据集D中抽取一个数据样本放入到D'中，然后放回到D中，重复m次之后，D'中有m个数据样本

如果我们对数据无限次抽样，那么一个样本不被抽到的概率为：

$$\lim_{m \to \infty} (1 - \frac{1}{m}) \mapsto \frac{1}{e} \approx 0.368$$

也就是说，通过自主采样，原数据集中大约有36.8%的数据没有被提取出去，所以可以使用这一部分数据作为测试集，提取出去的D'作为训练集

自助法主要适用于数据集较小，难以有效划分训练/测试数据集的时候，而且由于是从初始的训练集中提取出来不同的训练集，所以对集成学习有好处

3 性能度量

3.1 混淆矩阵

TP,FP,TN,FN组成的矩阵称为混淆矩阵

3.2 均方误差

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

更加一般的情况是，对于数据分布D和概率密度分布函数p(.)均方误差描述为:

$$E(f:D)=\int_{x\sim D}(f(x_i)-y_i)^2p(x)dx$$

3.3 错误率与精度

错误率定义为：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}(f(x_i)\neq y_i)$$

精度定义为:

$$acc(f;D)=1-E(f;D)$$

3.4 查准率，查全率，F1

查准率P(precision):

$$P=\frac{TP}{TP+FP}$$

查全率R(recall)：

$$R=\frac{TP}{TP+FN}$$

3.5 P-R图

很多情况下，可以根据学习器对样本的预测结果对样本按照score由大到小进行排序，这样我们顺序把每个样本排序，依次将样本设为正例，计算此时的P和R，绘制图像，这样就可以做出来一个横轴为R纵轴为P的P-R图，当对两个模型的优良进行对比时，如果一个模型预测结果的P-R图完全包含另外一个P-R图的结果，那么就可以断言其性能较优，但是P-R图比较难算。

3.6 F1

F1是根据P与R的调和平均值来确定的

$$F1=\frac{2\times P\times R}{P+R}$$

3.7 F_β

F_β 是加权调和平均。在一些应用中，我们对于查全率和查准率的重视程度不一样，这种情况下F_β是更加合适的：

$$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$$

β大于1的时候对查准率影响更大，β小于1的时候对查全率影响更大

3.8 ROC与AUC

一般来说，对样本进行预测一个score，会有一个threshold对其进行截断，将样本按照score(判断为正例的概率)进行由大到小排序，当我们对P更加关心时，threshold可以适当加大，当我们对R更加关心时，可以适当减小threshold

ROC称为受试者工作特征(Receiver Operating Characteristic)，与P-R曲线类似，我们首先将数据按照scrore进行由大到小排序，真正例率TPR(True Positive Rate)作为纵轴，假正例率FPR(False Positive Rate作为横轴)

$$TPR=\frac{TP}{TP+FN}$$
$$FPR=\frac{FP}{TN+FP}$$
绘制ROC曲线：假设有m^+个正例和m^-个反例，按照判断为正例的分值对结果进行由大到小排序，把分类的阈值设为最大，即所有的例子都是反例，此时TPR和FPR都是0，从原点开始，然后依次将每个样本划分为正例，假设前一个点的坐标为(x,y)，如果样本为真正例，点为$(x,y+\frac{1}{m^+})$，如果是假正例，点的坐标为$(x+\frac{1}{m^-},y)$

AUC(Area Under Roc Curve)也就是ROC曲线下面的面积。如果两个模型的ROC曲线一方包含另一方，那么在外层的模型性能更优，如果交叉，可以进一步比较AOC的大小，大者更优

3.9 代价敏感错误率与代价曲线

在现实中经常会遇到不同类型的错误所造成的后果不同，也就是对于FP和FN的惩罚程度不一样，而不仅仅是分类的错误，在非均等代价下，ROC曲线不能反映出学习期的期望总体cost，而代价曲线可以，代价曲线的横轴为正例概率代价：
$$P(+)_{cost}=\frac{p\times cost_{01}}{p\times cost_{01}+(1-p)\times cost_{10}}$$
纵轴为归一化代价：
$$cost_{norm}=\frac{FNR\times p\times cost_{01}+FPR\times(1-p)\times cost_{10}}{p\times cost_{01}+(1-p)\times cost_{10}}$$
其中$FNR=1-TPR$,ROC曲线上面每个点对应的是代价曲线上面的一条线，给定ROC上面的(FPR,TPR)，可以计算出FNR，代价平面上面绘制一条从(0,FPR)到(1,FNR)的线段，线段下面的面积即表示该条件下的期望总体代价。