机器学习第二章课后习题

162050127 颜劭铭

2022 年 5 月 9 日

1 课后习题

1.1 数据集包含 100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用 10 折交叉验证法和留一法分别对错误率进行评估所得的结果.

使用 10 折交叉验证法：

首先将数据集 D 中 100 个样本划分为 10 个大小相同的互斥子集，记为 $D_1, D_2, \cdots, D_{10}$，其中每个子集包含 10 个样本，5 个正例 5 个反例.

接着随机挑选 9 个子集的并集作为训练集，剩下那个子集作为测试集，因此训练集中有 45 个正例 45 个反例，测试集中有 5 个正例 5 个反例.

由于学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），因此对于测试集的样本进行随机猜测，有 50% 的几率猜测正确.

因此易得，对 10 折交叉验证法用错误率进行评估的结果为 50%.

使用留一法：

首先将数据集 D 中 100 个样本划分为 100 个大小相同的互斥子集，记为 $D_1, D_2, \cdots, D_{100}$.

随机选择 99 个子集作为训练集，1 个子集作为测试集.

当选择的测试集子集为正例时，意味着训练集中有 50 个反例 49 个正例，由于学习算法所产生的模型是将新样本预测为训练样本数较多的类别，因此对于测试集将会预测为反例.

同理，当选择的测试集子集为反例时，意味着训练集中有 50 个正例 49 个反例，由于学习算法所产生的模型是将新样本预测为训练样本数较多的类别，因此对于测试集将会预测为正例.

以上两种情况预测的结果都是错误的，因此易得：对留一法使用错误率进行评估的结果为 100%.

1.2 试述真正例率 (TPR)、假正例率 (FPR) 与查准率 (P)、查全率 (R) 之间的联系.

对于二分类问题，分类结果的混淆矩阵如下表所示：

真实情况 (标签)	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

首先给出题目中四个性能度量的定义：

真正例率，指的是标签为正例，被正确分为正例的样本数目，公式如下：

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

假正例率，指的是标签为反例，被错误分为正例的样本数目，公式如下：

$$FPR = \frac{FP}{TN + FP}$$

查准率，指的是预测结果为正例中标签为正例的样本数目，公式如下：

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

查全率，指的是标签为正例的样本中被正确分为正例的样本数目，公式如下：

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

经过对比我们可以发现，真正例率与查全率是相等的，而假正例率与查准率暂时没有发现什么联系.

1.3 试述错误率与 ROC 曲线的联系.

ROC 曲线使用的是假正例率 (FPR) 为横轴，真正例率 (TPR) 为纵轴进行绘制所得的图形.

而错误率指的是分类错误的样本数占样本总数的比例，基于表 1 的符号，可以给出以下公式：

$$E = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i) \neq y_i)$$

根据 ROC 曲线的绘图我们可以发现，它是先调整阈值为最大，将所有样本均预测为反例，绘制第一个点 (0,0)，然后不停地调整分类阈值，计算 FPR 和 TPR 进行绘制得到一条曲线，并且点 (0,1) 代表所有正例都排在反例之前的理想模型.

而错误率则是给定一个固定的分类阈值计算判断分类效果.

我们可以得出以下公式：

$$E = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(f(x_i) \neq y_i)$$

由 TPR 和 FPR 的定义将所有样本拆成正负例分别进行计算

$$= \frac{1}{m}\mathbb{I}\left(\sum_{x_i \in D^+}(f(x_i) \neq y_i) + \sum_{x_i \in D^-}(f(x_i) \neq y_i)\right)$$

$$= \frac{1}{m}(m^+ \times (1 - TPR) + m^- \times FPR)$$

因此，我们可以很轻松地得到 ROC 曲线与错误率的联系：ROC 曲线上绘制的每个点实际上就是确定好不同分类阈值时的真正例率和假正例率，而通过上式我们可以用 TPR 和 FPR 计算出所对应的错误率，因此我觉得可以认为 ROC 曲线就是一条对于同一个学习器，按照不同的分类阈值，得到的所有错误率连接形成的一条曲线，只不过曲线上每个点代表的并不是错误率，而是该分类阈值下的假正例率和真正例率.

2 附加题

2.1 对于有限样例，请证明

$$AUC = \frac{1}{m^+m^-}\sum_{x^+ \in D^+}\sum_{x^- \in D^-}\left(\mathbb{I}\left(f(x^+) > f(x^-)\right) + \frac{1}{2}\mathbb{I}\left(f(x^+) = f(x^-)\right)\right)$$

根据书上定义可得：

AUC 就是 ROC 曲线与 x 轴围成的面积之和，因此可以利用面积公式推导出 AUC 的计算公式，下面通过 ROC 曲线图进行推导。

这边给出南瓜书上的 ROC 曲线图：

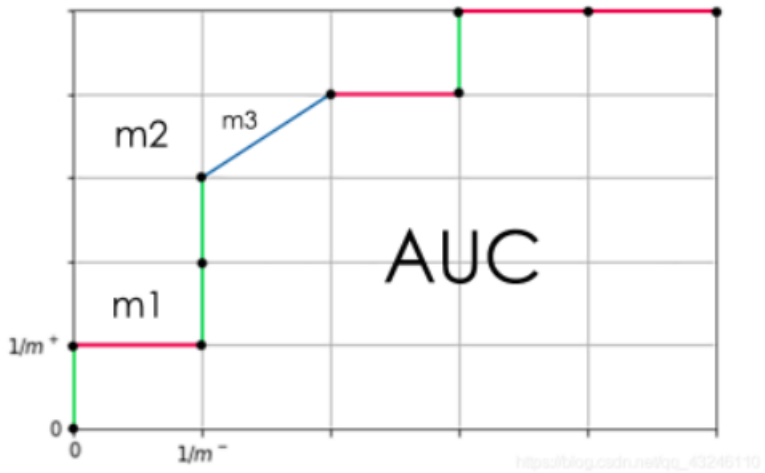

图 1: ROC 曲线示意图

先对于这张图给出解释：

将 x 轴（假正例率轴）的步长定为 $\frac{1}{m^-}$，y 轴（真正例率轴）的步长定为 $\frac{1}{m^+}$，因此每次变动分类阈值的时候，若新增 i 个假正例，那么相应的 x 轴坐标也就增加 $\frac{i}{m^-}$，若新增 j 个真正例，那么相应的 y 轴坐标也就增加 $\frac{i}{m^+}$.

对于不同颜色的线段，绿色线段表示在分类阈值变动的时候只新增了真正例，红色线段表示只新增了假正例，蓝色线段表示既新增了真正例也新增了假正例，因此此时的 AUC 值其实就是所有红色线段和蓝色线段与 x 轴围成的面积之和，并且观察图 1 可以发现，红色线段与 x 轴围成的图形恒为矩形，蓝色线段与 x 轴围成的图形恒为梯形，而梯形的面积公式既能算梯形面积，也可以算矩形面积，所以可以用梯形的面积公式 $\frac{1}{2} \times$ 高 $\times$ (上底 $+$ 下底) 进行证明该公式。

证明过程：

首先对于要证明的式子进行恒等变换：

$$
\begin{aligned}
AUC &= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{2}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right)\\
&= \frac{1}{m^+m^-}\sum_{x^+\in D^+}\left[\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{2}\cdot\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right]\\
&= \sum_{x^+\in D^+}\left[\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{2}\cdot\frac{1}{m^+}\cdot\frac{1}{m^-}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right]\\
&= \sum_{x^+\in D^+}\frac{1}{2}\cdot\frac{1}{m^-}\cdot\left[\frac{2}{m^+}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{m^+}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right]
\end{aligned}
$$

由图 1 及不同颜色的线段可得，当新增假正例的时候，图 1 会相应地增加红色线段或蓝色线段，因此 $\sum_{x^+\in D^+}$ 可以看作是累加所有红色和蓝色线段，因此 $\sum_{x^+\in D^+}$ 后面的内容便是在求红色线段或蓝色线段与 x 轴围成的面积，也就是：

$$
\frac{1}{2}\cdot\frac{1}{m^-}\cdot\left[\frac{2}{m^+}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{m^+}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right]
$$

上式中的 $\frac{1}{m^-}$ 即每个小梯形的高，中括号内便是“上底 + 下底”。

每新增一个真正例时，y 坐标就新增一个步长，因此对于上底，也就是红色线段或蓝色线段的下端点到 x 轴的距离，长度就等于 $\frac{1}{m^+}$ 乘以预测值大于 $f(x^-)$ 的真正例的个数，即：

$$
\frac{1}{m^+}\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)
$$

而对于下底，长度就等于 $\frac{1}{m^+}$ 乘以预测值大于等于 $f(x^-)$ 的真正例的个数，即：

$$
\frac{1}{m^+}\left(\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)>f(x^-)\right)+\sum_{x^-\in D^-}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right)
$$

因此将 (上底 + 下底) × 高 × $\frac{1}{2}$ 并求和就可以证明 AUC 的公式，即：

$$
AUC = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}\left(f(x^+)>f(x^-)\right)+\frac{1}{2}\mathbb{I}\left(f(x^+)=f(x^-)\right)\right)
$$

2.2 若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，此时的版本空间是什么？在此情形下，试设计一种归纳偏好用于假设选择.

版本空间：

当假设空间中不存在与训练样本一致的假设时，由于版本空间指的是与训练集一致的假设集合，因此我们可以得到此时的版本空间为空集。

当假设空间中存在与训练样本一致的假设时，我们可以得到他的版本空间是那些与训练集一致的假设集合。

归纳偏好：

对于假设空间有可能不存在与所有训练样本都一致的假设，我的理解是可能假设与训练样本特征属性相同但是标签不同引起的噪声。

① 第一种想法是去除噪声，忽略那些特征属性相同但标签不同的数据，但是感觉这样不仅会忽略掉噪声，也会忽略掉有效信息。

② 第二种想法 (因为第一章讲的是二分类问题) 是将那些特征向量相同，但是标签既有正例又有反例的样本都认为是正例，虽然样本中有一半是噪声，但是保证了 50% 的正确率。

③ 第三种想法是引入一个准确率的概念，准确率等于假设和样本相同的特征属性数/假设总的特征属性数，先保证假设和样本的标签是正确的，这个时候归纳偏好尽可能选择那些准确率高的假设。