周志华《机器学习》第2章课后习题

1. 数据集包含 1000 个样本，其中 500 个正例、500 个反例，将其划分为包含 70% 的训练集和 30% 样本的测试集用于留出法评估，试估算共有多少种划分方式。

> 1. 500 个正例，提取 70% 即为 350 个正例被作为训练集，剩余 30% 用作测试集；
> 2. 为保证划分后的训练集和测试集，其正（反）例比例与总体相当，反例也需要按相同的方式进行划分；
> 3. 综上，划分方式共有 $(C_{500}^{350})^2$ 种，数量级约为 2.986×10^{262} 。

2. 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。

> 对于 10 折交叉验证法，我们始终保证划分后的数据集正反例比例与总体相当，因此正反例仍各占一半。由于学习算法在不同类别的训练样本数相同时进行随机猜测，对于任意测试样例，其精度为 50% 。
>
> 对于留一法，当预留的测试样例为反例时，训练集中正例数量比反例数量多 1 个；当预留的测试样例为正例时，训练集中反例数量比正例数量多 1 个。由于学习算法将新样本预测为训练样本数量较多的类别，其精度将为 0% （错误率 100% ）。

3. 若学习器 A 的 F1 值比学习器 B 高，试析 A 的 BEP 值是否也比 B 高。

> F1 值由 P 值与 R 值的调和平均数定义，其中的 P 值与 R 值是基于给定阈值完成所有样本分类后的统计结果。
>
> BEP 值则来自于一个动态变化的阈值，使得 P 值与 R 值相等，此时 BEP = P = R 。
>
> F1 值与 BEP 值之间没有必然联系，F1 仅对数据集整体敏感，但 BEP 值对数据集内部的样本顺序敏感，之间不存在因果关系。

4. 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。

> 对于任意数据集，其混淆矩阵如下：
>
> | | 预测正例 | 预测反例 |
> | --- | --- | --- |
> | 真实正例 | TP | FN |
> | 真实反例 | FP | TN |
>
> 查准率描述所有真实正例被预测为正例的比例，公式为 $P = \frac{TP}{TP+FN}$ ；查全率描述所有预测正例中确实为正例的比例，公式为 $R = \frac{TP}{TP+FP}$ ；真正例率公式与查准率公式相同，假正例率描述所有真实反例被预测为正例的比例，公式为 $FRP = \frac{FP}{TN+FP} = 1 - R$ 。

5. 试证明公式 $\mathrm{AUC} = 1 - \ell_{rank}$ 。

> 详细证明可查阅知乎 @我是韩小琦 的推理过程。

6. 试述错误率与 ROC 曲线的联系。

> 错误率的计算基于固定的阈值，而 ROC 曲线是阈值随样本预测值变化的情况下绘制得到的。ROC 曲线上每个点都对应了相应状态下的错误率。

7. 试证明任意一条 ROC 曲线都有一条代价曲线与之对应，反之亦然。

> ROC 曲线上的每一个点，都对应了代价平面中的一条线段，取所有线段的下界连成折线，得到一条代价曲线。
>
> 反之，沿代价曲线从左向右遍历折线的每一组线段，即可得到 ROC 曲线上的每一个点。

8. 试析 Min-max 规范化和 z-score 规范化的优缺点。

> 根据 Min-max 规范化表达式 $x' = x'_{min} + \frac{x - x_{min}}{x_{max} - x_{min}} \times (x'_{max} - x'_{min})$，我们发现此种规范化方法计算简单、仅当新变量值超出规范域（最大或最小值）时需要更新域并重算所有值，但它对数据集中的极端异常值（离群点）敏感。
>
> 根据 Z-score 规范化表达式 $x' = \frac{x - \bar{x}}{\sigma_x}$，我们发现此种规范化方法计算稍复杂（获得 σ_x 值较为耗时），每次新样本的加入都要触发所有数据的重新规范化，但它对离群点不敏感，能够抹除离群点对规范化带来的影响。

9. 简述 $\mathcal{X}^2$ 检验的过程。

> 1. 对于二分类问题，使用留出法获得两种不同学习器的分类结果差别，制作『联列表』；
> 2. 假设学习器的学习性能相同，则变量 $|e_{01} - e_{10}|$ 服从正态分布，均值为 1，方差为 $e_{01} + e_{10}$，因此变量
>
> $$\tau_{\mathcal{X}^2} = \frac{(|e_{01} - e_{10}| - 1)^2}{e_{01} + e_{10}}$$
>
>   服从自由度为 1 的 $\mathcal{X}^2$ 分布；
> 3. 给定显著度 α，当以上变量值小于临界值 $\mathcal{X}^2_\alpha$ 时不能拒绝假设，则两个学习器性能没有明显差别，否则拒绝假设，两个学习器由显著差别，平均错误率更小的学习器性能更好。

10. 试述 Friedman 检验中使用式 (2.34) 和 (2.35) 的区别。

> τ_F 公式（式 (2.35) ）中的 $\tau_{\mathcal{X}^2}$ 由式 (2.34) 计算导出，可见 τ_F 相对 $\tau_{\mathcal{X}^2}$ 进行了进一步缩放，其数值显示效果更好。