## 第二章 模型评估与选择

2.1 结合对性能度量部分的阅读，简述错误率、精度、查准率与查全率的含义。

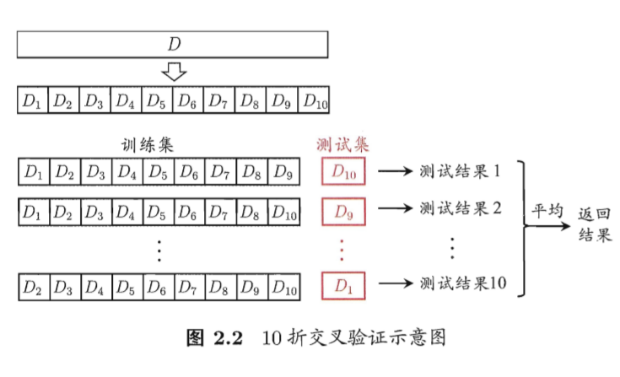
参考答案：错误率是酚类错误的样本数占样本总数的比例；精度是分类正确的样本数占总的样本数的比例；以文献检索为例，查准率是指检索出的相关文献量占检索出的文献总量的比例；查全率是指检索出的相关文献量占检索系统中相关文献总量的百分比。

2.2简述偏差与方差的含义，并结合方差与偏差理解过拟合和欠拟合问题。

参考答案：偏差度量了学习算法的期望预测和真实结果的偏离程度，刻画了学习算法本身的拟合能力，方差度量了同样大小的训练集的变动所导致的学习性能的变化，刻画了数据扰动所造成的影响，噪声表达了当前任务上任何学习算法所能达到的期望泛化误差下界，刻画了问题本身的难度。泛化误差可以分解成偏差的平方加上方差加上噪声，一般训练程度越强，偏差越小，方差越大，泛化误差一般在中间有一个最小值，如果偏差较大，方差较小，此时一般称为欠拟合，而偏差较小，方差较大称为过拟合。

2.3 试画出6折交叉验证法的示意图。

参考答案：形如下图：



2.4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%样本的训练集和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？

参考答案：这是一个组合问题，从500正反例中分别选出150正反例用于留出法评估，所以可能取法应该是(C150 500)2种。

2.5 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。

参考答案：10折交叉检验：由于每次训练样本中正反例数目一样，所以将结果判断为正反例的概率也是一样的，所以错误率的期望是50%。留一法：如果留下的是正例，训练样本中反例的数目比正例多一个，所以留出的样本会被判断是反例；同理，留出的是反例，则会被判断成正例，所以错误率是100%.

2.6 若学习器A的F1值比学习器B高，试分析A的BEP值是否也比B高。

参考答案：BEP是查准率P与查全率R相等时的取值，F1则为P与R的调和平均。若A的F1值要高于B，则在以下不等式中：

http://img.blog.csdn.net/20160301114839212?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

令PA=RA=BEPA,PB=RB=BEPB,得到BEPA>BEPB ,故假设成立。

2.7试述真正利率（TPR）、假正利率（FPR）与查准率（P）、查全率（R）之间的联系。

参考答案：真正例率和查全率都是指真实正例被预测为正例的比例，显然二者是相等的。查准率是指预测为正例的实例中真实正例的比例，假正例率是指真实反例被预测为正例的比例，两者没有直接的数值关系。

2.8 已知分类结果混淆矩阵如下，试计算查准率与查全率的值。

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | 100 | 300 |
| 反例 | 200 | 400 |

解：

查准率P=

查全率R= ，代入数据得P=0.33，R=0.25。

2.9试述错误率与ROC曲线的联系。

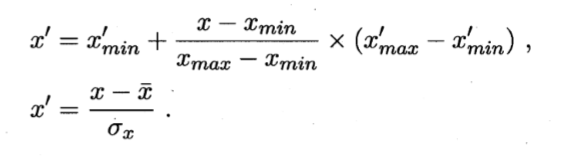
参考答案：错误率可以用下式表示：

http://img.blog.csdn.net/20160301115046073?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

参考真假正例率的定义，由于样本中正例和反例的比例是确定的。因此以某个样本的score为阈值时（即对应ROC曲线中的一个点），该点的真正例率越高，假正例率越低，则错误率越低，因此越靠近（0，1）点错误率越低。

2.10 Min-max规范化和z-score规范化是两种常用的规范化方法。令x和x’分别表示变量在规范前后的取值，相应的，令xmin和xmax表示规范化前的最小值和最大值，x min’和xmax’

表示规范化后的最小值和最大值,‾x和 σx分别表示规范化前的均值和标准差，则min-max规范化、z-score规范化分别如式(2.43)和（2.44）所示。试分析二者的优缺点。

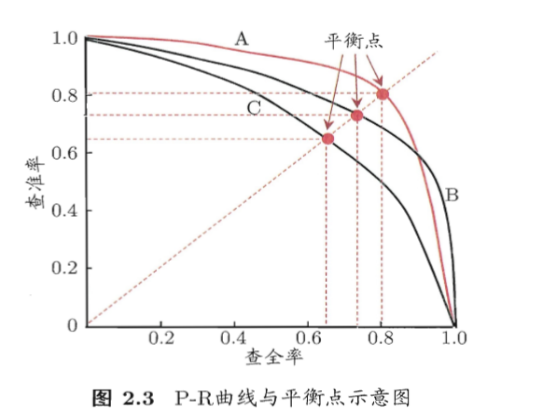


参考答案：Min-max规范化方法简单而且保证规范化后的所有元素都是正的，当有新数据输入时只有在该元素大于最大值或者小于最小值时才需要重新定义。z-course标准化对个别极端元素不敏感，适用于最大值最小值未知的情况，或者有超出取值范围的离群数据的情况，当有新数据输入时，需要重新计算方差与均值。

2.11下面的交叉验证方法 i.有放回的Bootstrap方法 ii.留一个测试样本的交叉验证 iii.5折交叉验证 iv.重复两次的5折交叉验证，当样本是1000时，上述交叉验证方法执行时间的顺序是?

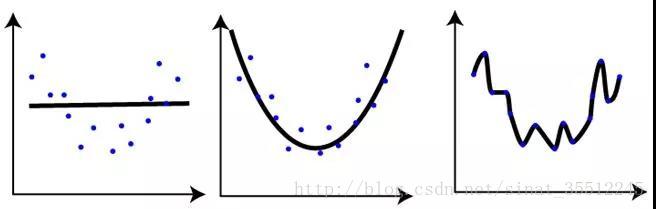
参考答案：ii>iv>iii>i

2.12 试比较下图中三个学习器的性能。



参考答案：学习器A的性能最优，B次之，C最差。

2.13 试比较下列三个模型的训练情况。



参考答案：第一张图的训练错误与其余两张图相比是最大的；第三张图相对于前两张图可能会出现过拟合；第二张图是三个图中表现最好的模型。

2.14 若要建立一个5000个特征，100万数据的机器学习模型，该怎么应对这样的大数据训练？

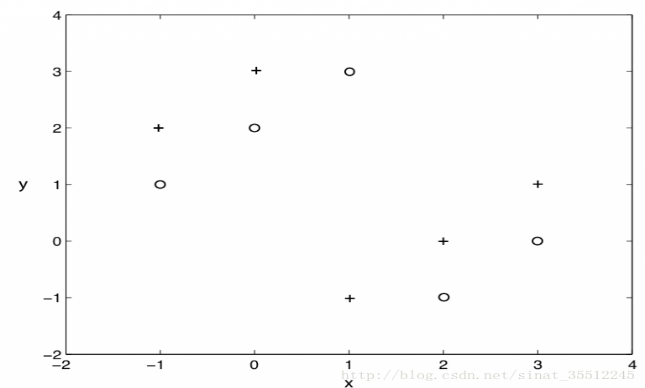
参考答案：

1.随机抽取一些样本，在这些少量样本之上训练

2.可以尝试使用在线机器学习算法

3.应用PCA降维，减少特征数

2.15 使用k=1的KNN算法，下图二类分类问题，“+”和“o”分别代表两个类，那么，用仅拿出一个测试样本的交叉验证方法，交叉验证的错误率是多少？



参考答案：100%