第一章 绪论 + 第二章 模型评估

2020年3月3日 17:39

第一章 绪论

1、机器学习的定义

利用经验来改善计算机系统自身的性能。经验主要是以数据的形式存储。即完成对数据的分析。

2、机器学习与数据挖掘的区别与联系

数据挖掘是识别出巨量数据种有效的、新颖的、潜在有用的、最终可理解的模式的非平凡过程。是机器学习和数据库的交叉，主要利用机器学习界提供的技术来分析海量数据。总体上机器学习更偏理论，数据挖掘偏向应用。

3、分类

分类的定义：

构建一个分类很熟或分类模型（即分类器），然后通过分类器对数据对象映射到某个给定的类别的过程。

分类过程：训练、测试、工作

第二章 模型评估

评估方法：

得到数据集后将数据集拆分成训练集和测试集。

训练集和测试集是两个互不相交的样本。

一般采用三种方法：留出法、交叉验证法、自助法

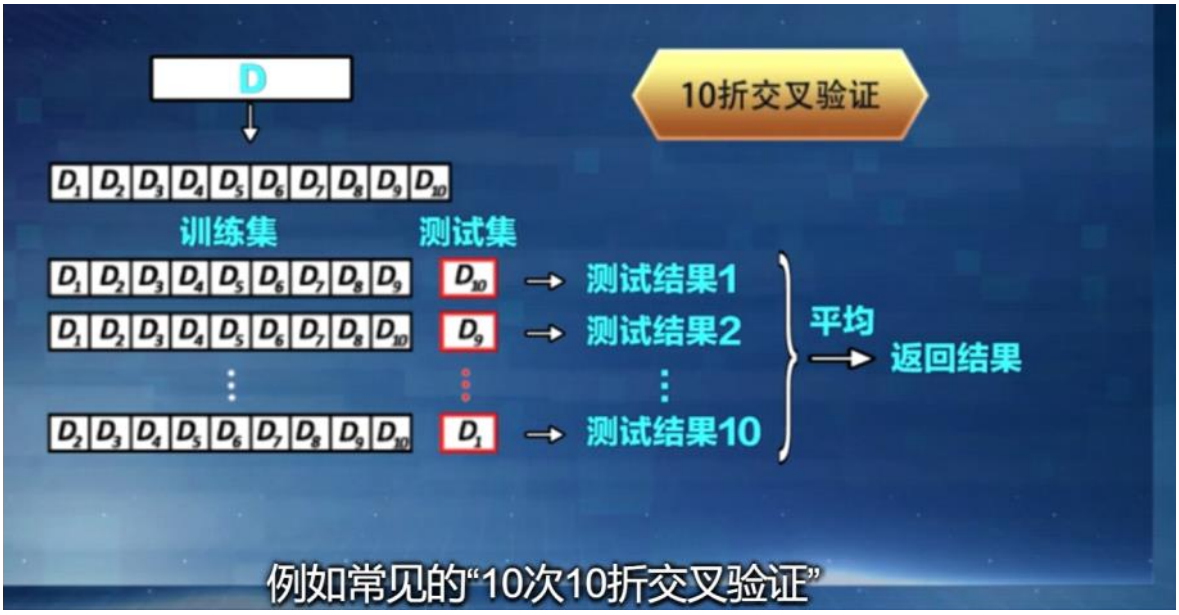
留出法：

直接将数据集分成两个互斥的集合，训练集和测试集尽可能保持数据分布的一致性。训练样本和测试样本的比例为2：1 、3：1、4：1

交叉验证法：

将数据集分层采样划分成k个大小相同或相似的互斥的子集，每次使用k-1个子集的并集为训练集，剩余的子集作

为测试集。最终返回k个测试结果的均值。k折交叉验证通常使用不同的划分重复p次。最终结果取均值



自助法：

以自助采样法为基础，对数据集D有放回采样n次得到训练集D'。为含有m个样本的数据集D，有放回的采样m次得到训练集，数据集中没有出现的为测试集

评估指标

准确率

分对的概率

错误率

分错的概率

查准率

被分为正类的样本中实际为正类的样本比例

$$P=TP/(TP+FP)$$

查全率

实际为正类的样本中被分为正类的样本比例

$$R=TP/(TP+FN)$$

F1度量：

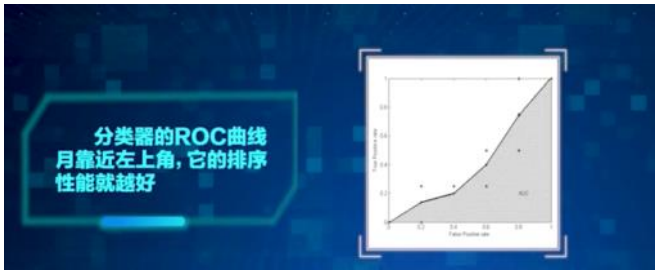
$$F1=\frac{2\times P\times R}{P+R}=\frac{2\times TP}{样例总数+TP-TN}$$

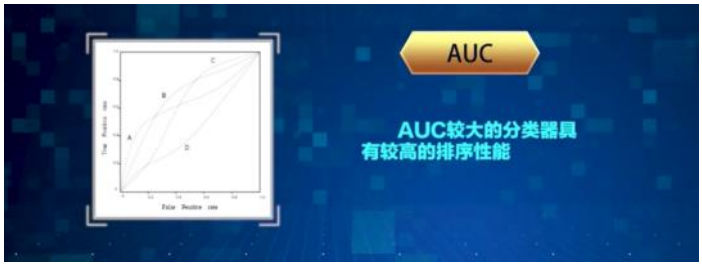
排序本身的质量好坏体现了分类器在不同任务下的泛化性能，度量性能的工具：ROC曲线

ROC曲线的绘制过程

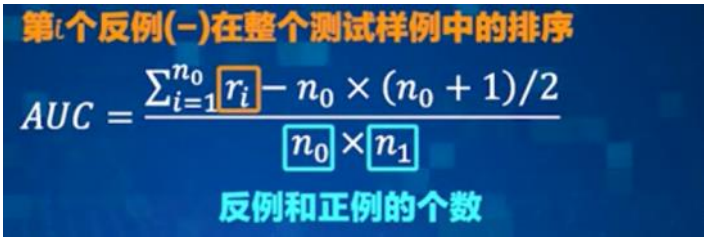
根据分类器的概率预测结果对样例排序，并按此顺序依次选择不同的“截断点”逐个把样例作为正例进行预测每次计算当前分类器的真正率和假正率。横轴纵轴作图

不同分类器的排序性能比较：若A分类器的ROC曲线包住了B分类器，则A更优





AUC：计算ROC曲线的面积。当曲线存在交叉时：



回归和分类的区别：

预测的目标函数是连续取值/离散取值

比较检验

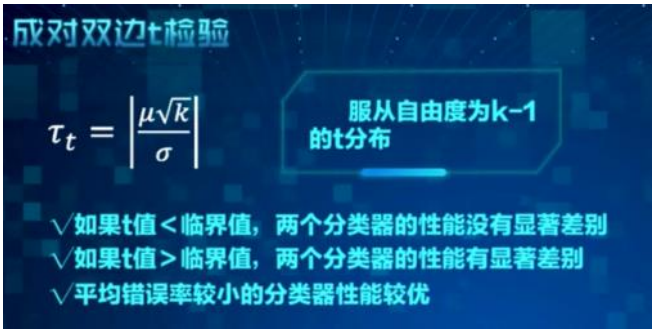
先使用某种实验评估方法测得分类器的某个评估指标结果，然后对这些结果进行比较

性能比较的方法：

成对双边t检验：

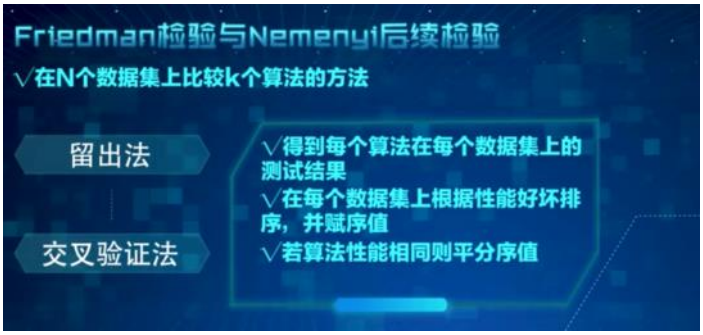
（一个数据集上比较两个分类器性能）

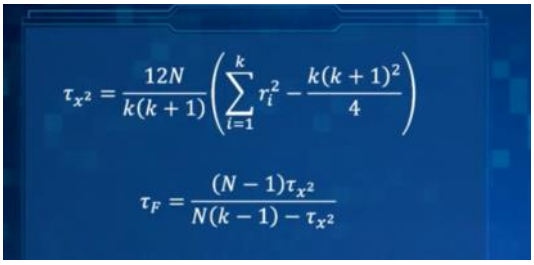
计算k组数据的误差统计量



Friedman检验与Nemenyi后续检验

N个数据集上比较K个算法的方法





依照上图计算F检验的统计量。TF满足：

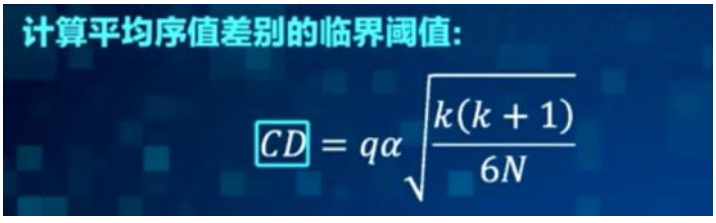
F 检验的常用临界值

$\alpha = 0.05$

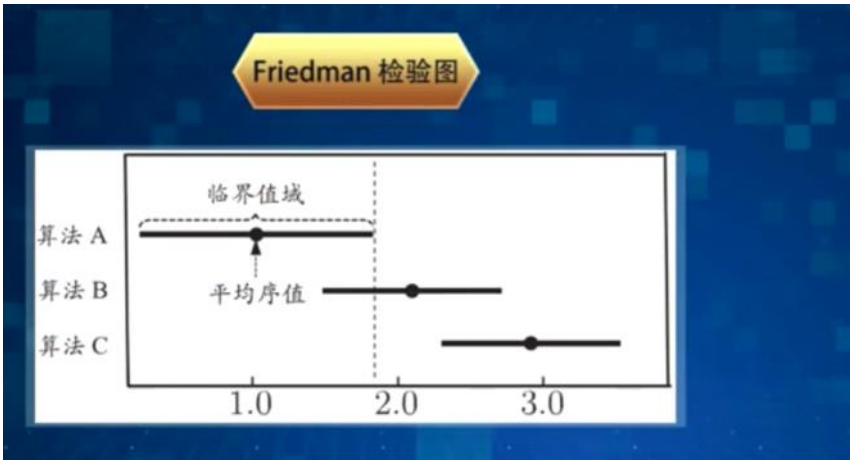
数据集个数 N	算法个数 k								
	2	3	4	5	6	7	8	9	10
4	10.128	5.143	3.863	3.259	2.901	2.661	2.488	2.355	2.250
5	7.709	4.459	3.490	3.007	2.711	2.508	2.359	2.244	2.153
8	5.591	3.739	3.072	2.714	2.485	2.324	2.203	2.109	2.032
10	5.117	3.555	2.960	2.634	2.422	2.272	2.159	2.070	1.998
15	4.600	3.340	2.827	2.537	2.346	2.209	2.104	2.022	1.955
20	4.381	3.245	2.766	2.492	2.310	2.179	2.079	2.000	1.935

只要Tf不大于临界值，则说明所比较的算法是没有显著不同的，否则还要进行后续检验。

如果比较的算法有显著不同，则要使用 Nemenyi 后续检验来进一步区分各 个算法的性能。Nemenyi 后续检验首先根据 Tukey 分布的临界值，这个临界值 在不同置信度下的临界值可以通过 Tueky 分布的临界值表查表得到，然后计算 平均序值差别的临界阈值，如果两个算法的平均序值<临界阈值，则两个算法的性能在相应的置信度下没有显著差别，反之则有显著差别，平均序值较小的算法较优。



举例为：



途中纵轴显示各个算法，横轴是平均序值。对每个算法用圆点表示其平均序值，以圆点为中心的横线段表示临界值域的大小。若两个算法有交叠，则说明没有明显差别，否则具有显著差别。AC之中，A显著优于C算法。