**<<机器学习>>整理和心得**

**第一、二章**

**机器学习:致力于如何通过计算的手段，利用经验即数据来改善系统性能，研究的主要内容是学习算法**

**属性(特征)张成的空间叫属性空间、样本空间或输入空间**

**样例指的是拥有了标记信息的示例**

**标记集合称为标记空间或输出空间**

**回归与分类，二分类或多分类**

**聚类:每组称为一个簇**

**书上的示例中，说浅色瓜、本地瓜这样的概念我们是事先不知道的，那是机器怎么自己得到的**

**监督学习(分类和回归)和无监督学习(聚类)**

**泛化能力:适应新样本的能力**

**版本空间:存在一个或多个与训练集一致的假设集合**

**看书上的举例，我是不是可以说所有可能性组成的空间也可称为一个版本空间**

**归纳偏好:机器学习算法在学习过程中对某种类型假设的偏好**

**奥卡姆剃刀原则:若有多个假设与观察一致，选最简单的那个，然而简单二字有时确实不好确定**

**所有算法期望性相同，总误差与学习算法无关**

**NFL定理告诉我们要结合具体问题谈算法的相对优劣**

**PR曲线:查全率为横坐标，查准率为纵坐标，平衡点(BEP)是横纵坐标相等时的取值，平衡点是学习器好坏的一个度量**

**F1度量(基于查准率和查全率的调和平均定义的)和Fβ度量(加权调和平均)**

**公式并没有搞懂，特别是F1那个带TP和TN的公式**

**在n个二分类混淆矩阵上综合考察查准率和查全率:宏-微**

**ROC曲线:纵轴是真正例率，横轴是假正例率，AUC是ROC曲线下的面积，越大越好，lrank=1-AUC,对应的是ROC曲线之上的面积**

**代价敏感错误率，代价曲线((0，FPR)到(1,FNR)的线段)，在横轴是正例概率代价，纵轴是归一化代价的图表上各线段的下界围成的面积为所有条件下学习器的期望总体代价**

**由测试错误率估推出泛化错误率的分布**

**假设检验:二项检验/t检验/交叉验证t检验/McNemar检验(列联表)/Friedman检验(通过平均序值相同推断算法性能相同)/Nemenyi后续检验(计算平均序值差别的临界值域，超出则以相应的置信度拒绝两个算法性能相同的假设)**

**Friedman检验图(纵轴表示各个算法，横轴表示平均序值，两个算法的横线段有交叠则说明两算法没有显著差别，否则就有)**

**偏差-方差分解:解释学习算法泛化性能的一种重要工具**

**泛化偏差可分解为偏差、方差、噪声之和**

**偏差度量了学习算法的期望预测和真实结果的偏离程度，即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，即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的下界，即刻画了学习问题本身的难度**

**偏差-方差窘境随着训练程度的增加，从偏差主导泛化错误率到方差主导泛化错误率，泛化误差先减后增**

**所以要控制训练程度**

1. **四章**

**线性模型试图学得一个通过属性的线性组合来进行预测的函数**

**非线性模型可在线性模型的基础上通过引入层级结构或高维映射而得**

**线性回归:基于均方误差最小化来进行模型求解的方法称为最小二乘法，就是试图找到一条直线使所有样本到直线上的欧氏距离之和最小**

**对数线性回归、对数几率回归(Sigmoid函数的代表)、线性判别分析(LDA是一种经典的线性学习方法，给定训练样例集，设法将样例投影到一条直线上，使得同类样例的投影点尽可能地接近)**

**线性判别分析个人觉得蛮巧妙的**

**多分类学习:一对一(OvO)，一对其余(OvR)，多对多(MvM)**

**纠错输出码,简称ECOC，编码和解码，ECOC码越长，纠错能力越强，但计算和储存开销也会越大**

**码长超过一定范围为什么会失去意义**

**类别不平衡:分类任务中不同类别的训练样例数目差别很大的情况**

**基本策略:再缩放，三种做法分别为欠采样、过采样、阈值移动**

**决策树是一种常见的机器学习方法**

**信息熵是度量样本集合纯度最常用的一种指标，Ent(D)的值越小，D纯度越高，信息增益越大，意味着使用属性a来进行划分所获得的纯度提升越大**

**增益率准则对可取值较少的属性有所偏好，信息增益准则相反**

**其原理不太懂，为何有如此偏好**

**CART决策树使用基尼指数来选择划分属性，Gini(D)越小，数据集D的纯度越高**

**剪枝处理:对付过拟合的主要手段，基本策略有预剪枝和后剪枝(验证集精度并决策是否剪枝)**

**连续值处理(采用二分法)、缺失值处理**