**机器学习1-3章温习总结**

**第一章**

**机器学习所研究的内容 说白了是一种训练计算机学习的“学习算法”，即如何教会机器自己学习。**

**机器学习始于训练样本，但是目标是让这个模型很好的适用于“新样本”，即“泛化能力”，故而有时过拟合是难以避免而且**

**学习过程：再所有假设组合的空间中进行探索，求得与训练集最最fit的假设。最后这个假设往往就是学习所得的结果，但是有可能存在多个不同的假设都很fit训练集，这时则考虑归纳偏好。**

**归纳偏好由奥卡姆剃刀进行引导，即选择“更平滑”,“更简单”的一个假设，这是常用的，自然科学研究中最基本的原则。**

**归纳学习：从样例中学习，狭义的归纳学习是从训练数据中学得概念（应该是更为苛刻的要求，因为有限的训练集所得出的概念往往很难适用于所有样本，即很难有很强的泛化能力）**

**没有免费的午餐定理：看起来难以理解，实际上通过书上计算可得出对于不同学习算法，总误差的期望性能与学习算法无关。即所有的算法都和胡猜拥有一样的可能产生误差大小的可能性。所以需要针对具体问题具体分析，找到不同问题偏重的方面。**

**人工智能从“推理期”——“知识期”到面临“知识工程瓶颈”，机器学习是研究发展到一定阶段的必然产物。**

**第二章**

**NP难：P问题是可以在多项式时间内被确定机(通常意义的计算机)解决的问题， NP(Non-Deterministic Polynomial, 非确定多项式)问题,是指可以在多项式时间内被非确定机(他可以猜,他总是能猜到最能满足你需要的那种选择,如果你让他解决n皇后问题,他只要猜n次就能完成----每次都是那么幸运)解决的问题。**

**即不是所有机器学习的任务都是能再多项式时间内解决（是NP问题），但是有效的学习算法是P问题，NP!=P。**

**自助法：用于数据集较小，难以有效划分训练/测试集是有用，对集成学习等方法有很大的好处，但是改变了初始数据集的分布，引入估计偏差。**

**留出法、交叉验证：用于初始数据量充足。**

**泛化能力性能度量：回归：均方误差**

**分类任务：accuracy 精度**

**Error-rate 错误率**

**Precision 查准率**

**Recall 查全率**

**F1 （查准率和查全率的调和平均）**

**比较检验：**

**假设检验：得到一次测试错误率：二项检验**

**由多次留出法和交叉验证多次训练、测试：t检验**

**k折交叉验证即可用“成对t检验”**

**若测试错误率小于某个临界值，则假设不可拒绝（应该是两个比较的学习器的性能没有显著差别）**

**多个算法比较时运用基于算法排序的Friedman检验；列表排序，给出序值。**

**偏差：算法本身拟合能力**

**方差：不同的数据集对学习性能的变化**

**噪声：学习问题本身的难度（所有算法所能达到的期望方差下界）**

**第三章**

**线性回归：输入属性只有一个，学得一个函数使得均方误差最小化**

**多元线性回归则利用矩阵得运算，可能有多租解，这时由学习算法的归纳偏好决定，可以引入正则化项。**

**“极大似然法”令每个样本属于其真实标记得概率越大越好**

**线性判别分析：“类内散度矩阵”同类样例投影点得协方差**

**“类间散度矩阵”一类阳历的投影点尽可能远离**

**应用到多类时，“全局散度矩阵”，“类内散度矩阵”重定义为每个类别散度矩阵之和。**

**多分类学习的基本思路：拆解法**

**拆分策略：一对一 OvO N个类别两两配对 N(N-1)/2个二分类任务**

**一对其余 OvR N个分类任务**

**多对多 MvM 纠错输出码**

**编码：N个类别M次划分，M个训练集训练M个分类器**

**译码：用M个分类器分别对测试样本进行预测，返回其中距离最小的类别作为预 测的类别结果。**

**类别划分：编码矩阵（二元码 分为正类反类，三元码 分为正类反类+停用类）**

**（对于同一个学习任务，ECOC编码越长，纠错能力越强，但是编码越长，所训练 的分类器越多）**

**Class-imbalance 分类任务中不同类别的训练样例数目差别很大。**

**基本策略：对预测值进行调整，进行“再缩放 rescaling”（也是代价敏感学习的基础）这是根据训练样例中正例与反例的比例来调整，而不再是直接用0.5作为阈值判断。**

**做法：1.对训练集的反类样例进行欠采样**

**对训练集里的正类样例进行过采样**

**直接基于原始训练集学习，但再用训练好的分类器进行预测时，将再缩放加入到决策过程中，——“阈值移动 threshold-moving**