<<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 F β 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

第三、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

第五、六章

神经网络:是由具有适应性的简单单元组成的广泛并行互连的网络，它的组织能够模拟生物神经系统对真实世界物体所作出的交互反应，最基本的成分是神经元模型，典型的一种是 M-P 神经元模型

SNN 指标准的神经网络，CNN 为卷积神经网络，RNN 为循环神经网络

激活函数:Sigmoid 函数，即挤压函数

Sigmoid 函数我觉得有一点不好的地方，在 X 轴趋近于正无穷或者负无穷的时候，斜率梯度会趋近于零，这是不是代表着学习进度会慢慢趋于缓慢，是否可以优化激活函数将它的斜率梯度一直保持下去，好像这就是梯度下降法的本质所在吧

感知机:由两层神经元组成，即阈值逻辑单元，可解决一些线性可分的问题比如与、或、非问题，要解决非线性可分问题则需要考虑多层

功能神经元

多层前馈神经网络:更常见的神经网络

误差逆传播算法:训练多层网络的一种学习算法，BP 算法基于梯度下降策略，目标是最小化训练集 D 上的累积误差

累积 BP 算法与标准梯度算法，区别类似于标准梯度下降和随机梯度下降之间的区别

梯度下降策略说白了是为了增加计算速度，这样能让我们训练更大规模的神经网络或者在合理时间内完成计算，也方便我们在优化自己的神经网络的时候能多次试验

通过试错法调整设置隐层神经元的个数问题

缓解 BP 网络的过拟合:早停和正则化

全局最小和局部极小:参数寻优需要尽可能接近或者找到全局最小，有多种途径和技术，但大多在理论上缺乏保障

其他常见的神经网络:RBF 网络(单隐层前馈神经网络)、ART 网络(竞争型学习的代表)、SOM 网络(竞争学习型的无监督神经网络)、级联相关网络(结构自适应网络的代表)、Elman 网络(递归神经网络的代表)、Boltzmann 机(基于能量的模型)

感觉神经网络的规模足够大才能保证深度学习性能，同时也需要很多数据，数据很少的时候各种学习算法性能都差不多吧，可以说规模在推动深度学习的进步

深度学习:训练很深层的神经网络，可通过增加隐层的数目或者隐层神经元的数目来实现，前者更有效果

无监督逐层训练是多隐层网络训练的有效手段，预训练+微调策略或者权共享策咯可以节省开销

用划分超平面将两类训练样本分开，选择泛化能力最强的

距离超平面最近的几个训练样本点使公式等号成立,它们被称为支持向量，两个异类支持向量到超平面的距离之和被称为间隔

支持向量机简称 SVM，为了最大化间隔，取其基本型

要解决对偶问题，SMO 算法为其中著名代表

模型最优解可通过训练样本的核函数展开,这一展式称为支持向量展式

很难确定合适的核函数使得训练样本在特征空间中线性可分,缓解该问题的一个方法是允许支持向量机在一些样本上出错,这便是软间隔向量机

一系列基于核函数的学习方法统称为核方法,最常见的是通过引入核函数来将线性学习器拓展为非线性学习器,通过核化来进行非线性拓展，从而得到核线性判别分析

第七、章

贝叶斯决策论是概率框架下实施决策的基本方法,考虑如何基于已知概率和误判损失来选择最优的类别标记

极大似然估计:根据数据采样来估计概率分布参数的经典方法

朴素贝叶斯分类器采用属性条件独立性假设,避开了联合概率难以从有限的训练样本直接估计而得的障碍，后尝试一定程度的放松，产生了半朴素贝叶斯分类器，独依赖估计是常见的一种策略，AODE 是一

种基于集成学习机制、更为强大的独依赖分类器

贝叶斯网:借助有效无环图来刻画属性间的依赖关系，并使用条件概率表来描述属性的联合概率分布,其中三个变量之间的典型依赖关系为同父结构、V 型结构、顺序结构

定义一个评分函数来评估贝叶斯网与训练数据的契合程度,然后基于这个评分函数来寻找结构最优的贝叶斯网

最小描述长度准则:应该选择综合编码长度最短的贝叶斯网

EM 算法:常用的估计参数隐变量的迭代式方法