<<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

合适的核函数是否一定存在

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、八章

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

拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题

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

贝叶斯网的近似推断常使用吉布斯采样来完成

EM 算法:常用的估计参数隐变量的迭代式方法

集成学习：通过构建并结合多个学习器来完成学习任务，有时也被称为多分类器系统、基于委员会的学习等

同质集成中的个体学习器称为基学习器

不同类型个体学习器结合起来是异质的,这样的个体学习器称为组件学习器或个体学习器

集成个体应该好而不同，个体学习器至少不差于弱学习器

Boosting 是一族可将弱学习器提升为强学习器的算法，著名代表为 AdaBoost 算法，通过重赋权法和重采样法来学习

Bagging：并行式集成学习方法的著名代表，基于自助采样法

随机森林：简称 RF，是 Bagging 的一个扩展变体，基学习多样性不

仅来自样本扰动，还来自属性扰动

结合策略有平均法，投票法，学习法

多样性度量是用于度量集成中个体分类器的多样性，典型做法是考虑个体分类器的两两相似或不相似性

要增强多样性，常见做法是对数据样本、输入属性、输出表示、算法参数进行扰动

第九、十章

聚类任务是无监督学习任务中研究最多的领域

聚类性能度量：我们希望聚类结果的簇内相似度高且簇间相似度低，可通过外部指标或内部指标来实现

距离计算：最常用的是闵可夫斯基距离

原型聚类：k 均值算法（针对聚类所得簇划分最小化平方误差）、学习向量量化（也是通过找到一组原型向量来刻画聚类结构，但假设数据样本带有类别标记，学习过程利用样本的学学监督信息来辅助簇类）、高斯混合聚类（采用概率模型来表达聚类模型）、密度聚类（从样本密度的角度考察样本之间的可连续性，并基于可连接样本不断扩展聚类簇以获得最终的聚类结果）、层次聚类（试图在不同层次对数据集进行划分，可以自底向上也可以自顶向下，从而形成树形的聚类结构）

k 近邻(kNN)学习是一种常见的监督学习方法，根据与其靠近的 k 个训练样本的信息进行预测，通常在分类任务中可使用投票法，即选择 k 个样本中出现最多的类别标记作为预测结果；在回归任务中可使用

平均法，即将这 k 个样本的实值输出标记的平均值作为预测结果；还可基于距离远近而加权

懒惰学习:训练阶段仅保存样本，待收到测试样本再处理

急切学习:在训练阶段就对样本进行学习处理

高维情形下出现的数据样本稀疏、距离计算困难等问题，被称为维数灾难，缓解它的一个重要途径是降维，多维缩放(MDS)是一种经典的降维方法；主成分分析(PCA)是常见的一种降维方法；非线性降维的一种常见方法是基于核技巧对线性降维方法进行核化，核主成分分析(KPCA)是一个例子;流形学习是一类借鉴了拓扑流形概念的降维方法等度量映射和局部线性嵌入都是为了得到训练样本在低维空间的坐标，但前者试图保持近邻样本间的距离不同，后者试图保持样本间的线性关系

度量学习目的是找到一个合适的低维空间，在此空间内学习能比原始空间性能更好

第十一、十二章

对当前学习任务有关的属性称为相关特征，没什么用的属性称为无关特征，特征选择是一个重要的数据预处理过程，冗余特征指的是一种所包含的信息是从其他特征中推演出来的特征

产生候选子集:子集搜索和子集评价,后者以计算出的信息增益作为准则，越大意味着特征子集 A 包含的有助于分类的信息越多

常见的特征选择方法大致可分为三类:过滤式、包裹式、嵌入式

过滤式方法先对数据集进行特征选择，然后再训练学习器，即先用特

征选择过程对初始特征进行过滤，再用过滤后的特征来训练模型，Relief 是一种著名的过滤式特征选择方法，该方法设计了一个相关统计量来度量特征重要性

包裹式特征选择直接把最终将要使用的学习器的性能作为特征子集的评价准则，其目的是为给定学习器选择最有利于其性能的特征子集，它直接针对给定学习器进行优化，因此从最终学习性能来看，包裹式特征选择比过滤式更好，但由于需要多次训练学习器，开销也远远大于过滤式特征选择

LVW 是一个典型的包裹式特征选择方法，它在拉斯维加斯方法框架下使用随机策略来进行子集搜索，并以最终分类器的误差为特征子集评价准则

嵌入式特征选择是将特征选择过程和学习器训练过程融为一体，两者在一个优化过程中完成，记在学习器训练过程中自动进行了特征选择当样本具有稀疏表达形式的时候对学习任务有不小好处，对于普通非稀疏数据，我们希望将之转化为稀疏表示的形式，而且是恰当稀疏为普通稠密表达的样本找到合适的字典，将样本转化为合适的稀疏表达形式，从而使学习任务得以简化，模型复杂度得以降低，称为字典学习，亦称稀疏编码

为解决根据部分信息来恢复全部信息，压缩感知提供了不错的思路，它关注的是如何利用信号本身所具有的稀疏性，从部分观测样本中恢复原信号，分为感知测量和重构恢复两个阶段，前者为了获得稀疏样本表示，后者为了从少量观测中恢复原信号，常见求解方法是转化为

LASSO的基寻踪去噪、基寻踪、匹配寻踪等

计算学习理论研究的是关于通过计算来进行学习的理论，即关于机器学习的理论基础，其目的是分析学习任务的困难本质，为学习算法提供理论保证，根据分析结果指导算法设计

几个常用不等式:Jensen不等式、Hoeffding不等式、McDiarmid不等式

概率近似正确(PAC)学习理论，目标概念集中存在假设能将所有示例按与真实标记一致的方式完全分开，则称该问题对学习算法是可分的或一致的，反之则不可分或不一致的，相关概念有PAC辨识、PAC可学习、PAC学习算法、不可知PAC可学习等

有限假设空间H都是PAC可学习的

为针对无限假设空间的可学习性进行研究，需度量假设空间的复杂度，最常见的方法是考虑假设空间的VC维，相关概念有增长函数、对分、打散

基于VC维的泛化误差界是分布无关、数据独立的，任何VC维有限的假设空间H都是不可知PAC可学习的

经验风险最小化(ERM)原则算法

Rademacher复杂度是另一种刻画假设空间复杂度的途径，与VC维不同的是，它在一定程度上考虑了数据分布，所以泛化误差界也更紧一些，不论是哪种方法，所得到的结果都与具体学习算法无关，对所有学习算法都适用

稳定性分析，考察的是算法在输入发生变化的时候，输出是否随之发

生较大变化

损失函数刻画了假设的预测标记和真实标记之间的差别，几种损失为泛化损失、经验损失、留一损失等

算法均匀稳定性的定义

第十三、十四章

主动学习:将每次新获得的有标记样本加入训练样本集中重新训练新模型，其目标是使用尽量少的查询来获得尽量好的性能，引入了额外的专家知识而对未标记样本进行了利用

半监督学习:让学习器不依赖外界交互、自动地利用未标记样本来提升学习性能

为利用未标记样本，需要做一些将未标记样本所揭示的数据分布信息与类别标记相联系的假设，最常见的是聚类假设，其推广为流形假设，它们的本质都是相似的样本拥有相似的输出这个基本假设

半监督学习可进一步划分为纯半监督学习和直推学习，前者假定训练数据中未标记样本并非待预测数据，希望模型能适用于训练过程中未观察到的数据；后者假定学习过程中所考虑的未标记样本恰是待预测数据，仅希望模型对学习过程中观察到的未标记数据进行预测

生成式方法:直接基于生成式模型的方法，此类方法假设所有数据都是由同一个潜在的模型生成的，通常基于 EM 算法进行极大似然估计求解，模型假设必须足够准确

半监督 SVM(半监督支持向量机):是支持向量机在半监督学习上的推广，S3VM 试图找到能将两类有标记样本分开且穿过数据低密度区域

的划分超平面，TSVM 是半监督支持向量机中最著名的例子，也是针对二分类的学习方法

图半监督学习:对于某数据集，我们可将其映射为一个图，数据集中每个样本对应于图中一个结点，若两个样本之间的相似度很高或相关性很强，则对应的结点之间存在一条边，边的强度正比于样本之间的相似度或相关性，一个图对应一个矩阵，我们能基于矩阵运算来进行半监督学习算法的推导分析

基于分歧的方法:使用多学习器，学习器之间的明显分歧对未标记数据的利用至关重要，协同训练是其中的一个代表，也是多视图学习的代表，能很好地利用多视图的相容互补性

半监督聚类:利用监督信息来获得更好的聚类效果，监督信息大致有两种类型，第一种类型是必连和勿连约束，前者是指样本必属于同一个簇，后者指必不属于同一个簇；第二种类型是少量的有标记样本

约束 k 均值算法是利用第一类监督信息的代表

概率模型的核心是如何基于可观测变量推测出未知变量的条件分布

概率图模型是一类用图来表示变量相关关系的概率模型，分为使用有向无环图表示变量间的依赖关系(称为有向图模型或贝叶斯网)和使用无向图表示变量间的相关关系(称为无向图模型或马尔可夫网)

隐马尔可夫模型是结果最简单的动态贝叶斯网，有向图模型的代表，主要用于时序数据建模，需要三组参数(状态转移概率、输出观测概率、初始状态概率)来进行构建

马尔可夫随机场:典型的马尔可夫网，一种著名的无向图模型

条件随机场:一种判别式无向图模型，对条件分布进行建模，试图对多个变量在给定观测值后的条件概率进行建模

对概率图模型还需确定具体分布的参数,这称为参数估计或参数学习问题，通常使用极大似然估计或者最大后验概率估计求解，概率图模型的推断方法大致可分为两类，第一类是精准推断方法，第二类是近似推断方法,这种方法又可以分为采样即通过使用随机化方法完成近似，和使用确定性近似完成近似推断(典型代表为变分推断）

精确推断的实质是一类动态规划算法,利用图模型所描述的条件独立性来削减计算目标概率值所需的计算量,变量消去法是最直观的精确推断算法

信念传播算法将变量消去法中的求和操作看作一个消息传递过程,较好地解决了求解多个边际分布时的重复计算问题

概率图模型中最常用的采样技术是MCMC方法,MH算法是其重要代表,基于拒绝采样来逼近平稳分布 p

变分推断通过使用已知简单分布来逼近需推断的复杂分布,并通过限制近似分布的类型，从而得到一种局部最优、但具有确定解的近似后验分布

话题模型:一族生成式有向图模型,主要用于处理离散型的数据(比如文本集合),LDA模型是话题模型的典型代表