《机器学习》/周志华 读后总结

第一章 绪论

主要内容总结：

1.1

机器学习的基本概念：一门学科，致力于研究如何通过计算的手段，利用经验来改善系统自身的性能。机器学习所研究的是关于再计算机上从数据产生“模型”的算法，即”学习算法“（learning algorithm）。

1.2

基本术语：“色泽“”根蒂“”敲声“ ——**属性/特征** attribute/feature

“青绿“”乌黑“——**属性值** attribute value

属性张成的空间——**属性空间/样本空间/输入空间** attribute space/sample space（每个西瓜可在空间中找到一个坐标位置，即每个点对应一个坐标向量，故而一个示例称为一个“特征向量” feature vector）

（色泽=青绿；根蒂=蜷缩；敲声=浊响）——**记录/示例/样本** instance/sample

记录的集合——**数据集** data set

所学得的模型对应关于数据的某种潜在规律——**假设** hypothesis

样本的结果信息——**标记** label

拥有标记信息的示例——**样例** example

标记的集合——**标记空间/输出空间** label space

学习任务分类：一、监督学习 （有导师学习）supervised learning

A、分类 classification（所预测的是离散值）

1、 二分类 binary classification 只涉及两个类别，其中一个为正类，另一个为反类。

2、 多分类 multi-class classification 回归 regression （所预测的是连续值）

B、 回归 regression （所预测的是连续值）

二、无监督学习（无导师学习） unsupervised learning

A、聚类 clustering （将训练集中的西瓜分成若干组，每组称为一个簇（cluster）如：浅色瓜，深色瓜，本地瓜……

1.3

假设空间：学习过程就是在所有假设所组成的空间中进行搜索，找到与训练集最“匹配”（fit）的假设。

可能出现学习结果产生的多个假设与训练集一致，则这些假设集合为版本空间 version space

1.4

归纳偏好：在多个与训练集一致的假设中找到算法“偏好”的假设。

引导算法确立正确偏好的原则：“奥卡姆剃刀” (Occam’s razor) 即在多个假设中选择更简单的一个。

第二章 模型评估与选择

2.1 经验误差与过拟合

过拟合：overfitting 学习能力过于强大，关键障碍，无法彻底避免只能缓解，减小其风险。

欠拟合：underfitting 训练样本的一般性质尚未学好，容易克服，在决策树学习中扩展分支等。

2.2 评估方法

一、留出法 直接将数据集 D 划分为两个互斥的集合 S (训练集)T(测试集)

二、交叉验证法 将数据集分为 k 个大小相似的互斥子集，每次把一个当作测试集

三、自助法 每次随机挑选进入数据集 D'且每次采样后下次仍有可能被采样到

2.3 性能度量

——对模型泛化能力的评价标准（进行有效可行的实验评估方法之后进行评价）

一、错误率与精度

二、**查准率 precision、查全率 recall 与 F1** P-R 图直观显示学习器在样本总体上的 PR 值

三、ROC“受试者工作特征“ **AUC** 通过对 ROC 曲线下各部分面积求和可得

四、**代价敏感错误率与代价曲线** 代价矩阵（costij 表示将第 i 类样本预测为第 j 类样本的代价）

2.4 比较检验

对学习器的性能进行评估比较，利用统计假设检验，得出统计意义上的优劣结论和结论的把握多大。

一、假设检验

二、交叉验证 t 检验

三、McNemar 检验

四、Friedman 检验 与 Nemenyi 后续检验

2.5 偏差与方差

泛化误差分解为：偏差、方差与噪声之和。泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度决定。

第三章 线性模型

3.1 基本形式

线性模型（linear model）试图学得一个通过属性的线性组合来进行预测的函数 f(**x**).(x 为由 d 个属性描述的示例）

3.2 线性回归

线性回归（linear regression）试图学得一个线性模型以尽可能准确的预测实值输出标记。

（最小二乘法 least square method 欧氏距离 Euclidean distance）

广义线性模型：除了线性回归还可以进行非线性函数映射（如对数线性回归，输出标记在指数尺度上变化。对数线性回归是广义线性模型的特例。联系函数 link function

3.3 对数几率回归

对数几率函数是一个常用的近似单位阶跃函数的“替代函数 surrogate function

实际上是一种分类学习方法，可以直接对分类可能性进行建模，不需要实现假设数据分布，避免假设分布不准确所带来的问题。

3.4 线性判别分析 LDA

给定训练样例集，设法将样例投影到一条直线上，使得同类样例投影点尽可能接近，异类样例尽可能远离。分类是也可以通过投影的位置确定样本的类别。

（**一般用来降维，是一种经典的监督降维技术**）

3.5 多分类学习

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**）

3.6 类别不平衡问题

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）

做法：1.对训练集的反类样例进行欠采样

2. 对训练集里的正类样例进行过采样

3. 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

第四章 决策树

4.1 基本流程

决策树 decision tree 进行一系列的判断和”子决策“（提出对于某个属性的”测试“的判定问题，如色泽=？ 根蒂=？）

包含：一个根节点，若干个内部节点（属性测试） 若干个叶子节点（决策结果）

目的：产生一棵泛化能力强的决策树

策略：分而治之

决策树的生成是一个递归过程

4.2 划分选择

决策树的学习关键：如何选择最优划分属性，再划分过程的进行中让决策树的分支节点所包含的样本尽可能属于同一类别（节点的纯度越来越高）

信息熵 information entropy 度量样本集合纯度的常用指标。样本集合 D 的信息熵记为 Ent（D），值越小 D 的纯度越高

信息增益 information gain 进行决策树的划分属性选择。值越大，利用某个属性进行划分所获得的纯度提升越大。（对可取数值数目较多的属性有所偏好）

增益率 gain ratio Gain_ratio(D,a）减少属性偏好的不利影响

属性的固有值 intrinsic value IV(a)（属性 a 的可嗯那个取值数目越多，V 越大，IV(a）的值越大）

基尼指数 Gini index CART 决策树划分属性的准则 也可以度量数据集 D 的纯度

4.3 剪枝处理

Pruning 对付"过拟合"(决策树分支过多）的主要手段

基本策略：预剪枝 prepruning 划分节点前进行估计

后剪枝 post-pruning 自底向上对非叶子节点考察

后往往比预保留更多的分支，且后的欠拟和风险小，泛化性能由于预。但训练时间大得多。

4.4 连续与缺失值

连续值策略：采用二分法对连续属性进行处理

（与离散属性不同，当前节点划分属性为连续属性时，该属性还可以作为其后代结点的划分属性）

缺失值策略：(样本的某些属性值缺失）(1）推广信息增益（为每个样本赋予权重）(2）若样本 x 在划分属性 a 上取值未知，则将 x 同时划入所有子节点，同时调整样本在与属性 a 对应的子节点中的权值

4.5 多变量决策树

多个属性描述的样本经过决策树分类即找到空间中的分类边界（轴平行）

策略：斜划分（不是对某个属性，而是对属性的线性组合进行测试）

第五章　神经网络

5.1 神经元模型

神经网络 neural networks 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做出的交互反应。

最基本的成分：神经元模型 neuron

M-P 神经元模型 $y=f\left(\sum_{i=1}^{n} w_i x_i-\theta\right)$ 总输入值（输入乘以权值）与神经元的阈值相比较，通过激活函数处理产生神经元的输出。

激活函数：阶跃函数（不连续不光滑）/优化为：Sigmoid 函数（挤压函数）

5.2 感知机与多层网络

感知机：两层神经元组成：输入层、输出层（M-P 神经元/阈值逻辑单元）

只有一层功能神经元，学习能力有限，可以将阈值和权重统一为权重的学习，学习规则简单。

可以实现与或非等逻辑运算，却不能解决异或这样的非线性可分问题。

简单的感知机学习算法：对不同的样例的输出根据错误的程度进行权重调整。

多层前馈神经网络（multi-layer feedforward neural networks)：

（输入层神经元仅接受输入，不进行函数处理，隐层与输出层包含功能神经元，故而一层输入层，一层隐层一层输出层的神经网络称为“两层网络”/“单隐层网络”）

神经网络学习：连接权+阈值

5.3 误差逆传播算法 BP 算法（迄今最成功的神经网络学习算法）

不仅可用于多层前馈神经网络，还适用于其他类型的神经网络（递归神经网络）

标准 BP 算法：计算均方误差 E_{k}

策略：基于梯度下降 对误差 E_{k} 给定学习率 η，求偏导得 BP 算法中的更新公式。

更新公式：$\Delta w_{hj}=\eta g_j b_h$

学习率控制算法每一轮迭代的更新步长： 太大震荡，太小收敛速度过慢。可令连接权和阈值的学习率不同，作为精细调节。

目标：最小化训练集 D 上的累积误差。

（每次只针对一个训练样例更新，即基于单个的 E_{k} 推导调整参数，参数更新频繁，且每次更新过程可能出现抵消，迭代次数多。）

累积 BP 算法：直接针对累积误差最小化，读取整个训练集 D 一遍后才对参数进行更新，参数更新频率低。一般还是在最后用标准 BP，否则最后的微调也需要执行所有训练集后再调整的话，时间效率过低）

BP 网络表示能力强大，易过拟合，训练误差降低同时容易测试误差上升。解决方

法：1、早停 （当训练集误差降低但验证集误差升高时停止）2、正则化（在误差目标函数中增加一个用于描述网络复杂度的部分，使网络输出更"光滑"，对过拟合进行缓解）

设置隐层神经元个数：试错法调整

5.4 全局最小与局部极小

即一组对于连接权和阈值的最优参数使得神经网络再训练集上的误差 E 最小。

（可以类比函数中的最小值与极小值）

跳出局部极小找到全剧最小的方法：

1、从多组不同参数值初始化开始(即从不同的出发点同时出发开始寻找最优参数，陷入不同的局部极小后再进行比较获得全局最小)

2、"模拟退火" 就是可以接受比局部最小更差的次优解，从而跳出，但是接受次优解的概率要随着时间的推移降低

3、随机梯度下降 陷入了局部极小点所计算的梯度也可能不为零，从而继续探索

5.5 其他常见神经网络

1、**RBF 径向基函数网络** ：单隐层前馈神经网络激活函数为径向基函数

输出层是对隐层神经元输出进行线性组合

具有足够多隐层神经元的 RBF 可以以任意精度逼近任意连续函数

训练过程：确定第 i 个隐层神经元所对应中心 c_i （随机采样，聚类）

BP 算法确定第 i 个隐层神经元所对应的权值和第 i 个输出神经元的输入

2、**ART 自适应谐振理论网络** 竞争型学习代表。

构成：比较层（接收输入样本，并传递给识别神经元）

识别层 （每个神经元对应一个模式类，数目可以动态增加）

识别阈值 （计算输入向量与识别层神经元所对应模式类代表向量之间的距离最小的识别层神经元获胜，相似度大于阈值归为该类，否则新建一个类别） 阈值大：分类精细，阈值小，分类粗略

重置模块

优点：可以进行增量学习或在线学习

3、**SOM 自组织映射网络** 竞争学习型无监督神经网络

降维映射

4、**级联相关网络** 结构自适应网络代表

主要成分：级联（建立层次链接的层级结构）、相关（最大化神经元的输出与网络误差之间的相关性）

无需想和之网络层数，隐层神经元数目，且训练速度较快，训练数据小时容易过拟合

5、**Elman 网络** 常用递归神经网络

6、**Boltzmann 机** 定义能量为网络状态

最小化能量函数。

标准 B..机：全连接图

受限 B..机简化

5.6 深度学习

复杂模型 很深层的神经网络，增加隐层数目（比增加隐层神经元数目有效）

难以直接用 BP 算法进行训练，容易发散

无监督逐层训练 预训练+微调/权共享（手写数字识别任务，卷积神经网络）

第六章　支持向量机

6.1 间隔与支持向量

分类学习：找到基于训练集 D 在样本空间中的一个划分超平面

支持向量：距离超平面最近的，离超平面的距离为+1 或者-1 的几个训练样本点。

间隔：两个异类支持向量到超平面的距离之和

支持向量机：找到最大间隔的划分超平面

6.2 对偶问题

对凸二次规划问题利用拉格朗日乘子法得到”对偶问题“

二次规划问题：SMO 高效算法，固定除了某两个参数的其他参数求解并更新这两个参数。

（SMO 使用启发式：选取一个是目标函数值增长最快的变量和与其间隔最大的另一个变量）

偏移项 b：使用所有支持向量求解的平均值

6.3 核函数

非线性可分问题：映射到更高维的特征空间

核函数 :涉及计算样本映射到特征空间之后的内积，原始样本通过核函数运算。(支持向量展式)

6.4 软间隔与正则化

允许支持向量机再一些样本上出错

引入正则化常数 C 和 0/1 损失函数(可以换成别的替代损失函数以得到其他的学习模型)

优化目标包含划分超平面的”间隔“大小+训练集上的误差

6.5 支持向量回归 SVR

学得回归模型使得函数与分类尽可能接近。

两部分：正则化常数 C+e-不敏感损失（可由松弛变量替代）

6.6 核方法

通过”核化“来将现行学习器拓展为非线性拓展器。(是很强大的学习方法，最优解可以表示为核函数的线性组合)

最后可以求得映射到多维特征空间的关于 h 的范数

第七章 贝叶斯分类器

7.1 贝叶斯决策论

贝叶斯分类器是统计学、概率框架下实施决策的基本方法。贝叶斯决策可以在所有相关概率都已知的理想情形下，考虑如何基于这些概率和误判损失来选择最优的类别标记。

后验概率(P(ci|**x**)) 已知样例各个属性取值 **x**，分类为 ci 类的概率。这个概率一般是用来分类的主要依据，即对某个 ci 该值越大，越有可能是该类别的。

先验概率(P(c)) 就是训练集样本中 ci 类别的概率。可根据各类样本出现的频率进行估计。

类条件概率 P(**x**|ci) 在不同的类别中，如 ci 中，样本各特征值的概率分布。涉及了 x 的所有属性联合概率，容易出现“稀疏性”。（样本空间可能的取值远大于训练样本数，即未被观测到的样本取值将有很多，却会被认为是出现概率为零处理。）

联合概率分布 P(**x**,c) 往往很难求得，特别是 x 中的各个属性不是互相独立的时候。

贝叶斯判定准则就是最小化总体风险，（可由后验概率获得的“条件风险”求得）。而最小化总体风险就需要最小化条件风险。

能达到上述要求的是贝叶斯最优分类器，与之对应的总体风险是贝叶斯风险。

1-风险得到的是分类器能达到的最好性能。

获得后验概率往往难以在现实任务中直接获得。

两种策略估计后验概率：1、判别式模型 通过建模 P(c|**x**) 预测类别（应该是指求的这样一个函数模型/预测模型 比如决策树、BP 神经网络、支持向量）

2、生成式模型 先对联合概率分布 P(**x**,c)建模，再由此获得 P(c|**x**)（由 P(c)先验概率利用贝叶斯公式求得）

7.2 极大似然估计

估计类条件概率的策略：先假定 x 所有属性具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。(概率模型的训练问题变为概率论中的参数估计问题，但是如何确定该种概率分布形式？猜测可能导致误导性的结果)

频率主义学派 参数是固定值。

贝叶斯学派 参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。(参数本身是一个值，它依据什么进行分布？难道是根据数据的不同还会变化？贝叶斯学派的较频率主义学派更难以理解)

极大似然估计中的连乘容易造成下溢，通常使用对数似然变成连加。(因为概率一般小于一，如果连乘过多可能会数值太小了)

7.3 朴素贝叶斯分类器

由于估计后验概率的主要困难在于类条件概率是所有属性上的联合概率，难以从有限的训练样本中直接估计而得。

朴素贝叶斯：解决方法：“属性条件独立性假设”（其中对于连续性属性假设了其服从正态分布，这是普遍适用的方法吗？）

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。(有些预测样本中会出现新属性，而他们的概率分子为 0)

朴素贝叶分类器的学习可以利用“懒惰学习”方式，不断再数据增加过程中进行计数

修正实现增量学习。

7.4 半朴素贝叶斯分类器

对“属性条件独立性假设”的放松。

独依赖估计 ODE：每个属性在类别之外最多仅依赖于一个其他属性。(这里应该是把类别也看作一个属性了)“超父” 假设所有属性都依赖于同一个属性。利用交叉验证来确定超父属性。(SPODE)

TAN 最大权生成树 1、计算两个属性之间的条件互信息
2、以属性为节点构建完全图，折权重
3、构建完全图的最大全省成熟
4、加入类别结点 y，增加 y 到每个属性的有向边

AODE 基于集成学习机制，更为强大的独依赖分类器。尝试将每个属性都作为超父来构建。(计数、无需模型选择、预计算节省时间、易于增量学习)

高阶依赖：多个属性依赖。难以计算高阶联合概率。

7.5 贝叶斯网

信念网

贝叶斯网由结构和参数两部分构成。(网：有向无环图)

结构分为：同父结构、V 型结构、顺序结构。(其中对于某些取值已知或未知会对另两个变量的独立性产生影响，”边际独立性“这种独立性比较难以理解。让人想到薛定谔的猫)

通过有向分离可以分析有向图中变量间的条件独立性。

学习

找出结构最恰当的贝叶斯网。评分函数来评估贝叶斯网与训练数据的契合程度。基于信息论准则。(求得综合编码长度最短的贝叶斯网，类似信息传输中的压缩编码)

D 上的经验分布，即事件在训练数据上出现的频率。

推断

通过已知的变量观测值”证据“推断待查询的变量值。

吉布斯采样，随机采样方法。保证了所求厚颜概率式子收敛于所求得的值。

7.6 EM 算法

对未观测变量”隐变量“估计参数。利用迭代式的方法。(是非梯度下降方法，本来可以通过梯度下降等优化算法求解，但是求和的项数上升过快，难以计算)、

第八章 集成学习

8.1 个体与集成

构建并集合多个学习器来完成学习任务。
同质：包含同种类型的个体学习器。"基学习算法"
异质：不同学习算法构成。"组件学习器"

、一般作用于"弱学习器"更为效果显著。
要求：各个学习器好而不同。

两类集成学习方法：1、个体学习器间存在强依赖关系，必须串行生成序列化方法。

8.2 Boosting （降低偏差）

先让一个基学习器在初始训练集上训练。样本分布调整后（可以是给分类错误的样本更多关注，使得后续的基学习器能纠正之前的分类器的一些错误 利用重赋值法给新南联样本重新赋予权重/重采样法），再让下一个基学习器训练...重复进行。最后将基学习器加权结合。

2、个体学习器间不存在强依赖关系，可以同时生成的并行化。

8.3 Bagging&随机森林（降低方差）

对训练样本采样产生若干个不同的自己进行同时学习。（利用互有交叠的采样子集可以防止每个学习器学习的数据过少了）

自助采样法。根据计算剩下的 36.8%样本可以当作验证集。（采样时从未取到）/辅助剪枝 零训练样本结点的处理 （？如何做到辅助）

标准 AdaBoost 只适用于二分类任务 Bagging 可以不经修改的用于多分类、回归等任务。

"随机森林" **RF 是** Bagging 的扩展变体。再决策树的训练过程中加入了随机属性选择。（也是为了集成的好而不同）

代表集成学习技术水平的方法。

8.4 结合策略

结合的好处：防止泛化性能不佳（应该是假设相似，过拟合了）、防止局部极小点（陷入无法出来，这个点假设性能很差，往往不是最好的（最小点））、扩大假设空间（防止有些正确的假设不存在于现有的假设空间中）

1、平均法 简单平均 加权平均

2、投票法 绝对多数（拒绝预测） 相对多数 加权投票。类标记：硬投票。类概率：软投票。

3、学习法 通过另一个学习器结合。初级学习器 产生新数据集训练次级学习器（新数据集是如何获得的呢？初级学习器的输出作为输入的特征，而初始标记还是样例标记，多响应回归）

8.5 多样性

分歧：表征个体学习器在样本上的多样性。

误差-分歧分解

难以做到优化（个体学习器泛化误差-个体学习器的加权分歧值）目标。

只适用于回归学习，难以推广到分类学习任务。（为何下此定论？）

多样性度量：（成对型多样性度量）不和度量 相关系数 Q-统计量 k-统计量，可以通过二维图画出来。

多样性增强的方法：数据样本扰动。基于采样法进行干扰。

输入属性扰动：不同的属性空间中选出若干子集进行训练。节省时间开销。

输出表示扰动：对输出进行操纵，随机改变一下训练样本的标记。将原任务拆解为多个可同时求解的子任务。

算法参数扰动：随机设置不同的基算法参数。可产生差别较大的个体学习器。