《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）。

·1w1.2

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“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。

第九章 聚类

9.1 聚类任务 clustering

训练样本的标记信息未知。
将数据集中的样本划分为若干个通常是不相交的子集。每个子集称为“簇”。

9.1 性能度量

亦称“有效性指标”。聚类的结果需要“簇内相似度”高，“簇间相似度低”（类似分割）
两类聚类性能度量方法：1、外部指标 将聚类结果与某个“参考模型”比较

Jaccard 系数
FM 指数
Rand 指数
三个数都是越大越好（分子含 a 或 d，代表与参考模型分类结果相同）

3、内部指标 直接考察聚类结果 没有参考模型

Avg(C) 簇 C 内样本间的平均距离
Diam(C) 簇 C 内样本间的最远距离
Dmin(C) 簇 Ci 和簇 Cj 最近样本间的距离
Dcen(C) 簇 Ci 和簇 Cj 中心点之间的距离

DB 指数 越小越好
Dunn 指数 越大越好

9.3 距离计算

闵科夫斯基距离 p=2 时为欧氏距离
P=1 时为曼哈顿距离
属性分类：连续属性 离散属性

其中离散属性可能是有序的可能是无序的(决定了是否可以直接用属性值计算距离)

无序属性：VDM 距离 属性 u 上两个离散值 a，b 的 VDM 距离利用样本数个数计算。不同属性重要性不同可以使用“加权距离”。

非度量距离： 满足直递性 （可以利用三角形两边之和大于第三边不满足来记忆）

距离度量学习来基于数据样本确定合适的距离计算式

9.4 原型聚类

聚类结构能通过一组原型刻画，在现实聚类任务中常用。先对原型初始化，再对原型进行迭代更新求解。(会产生不同的算法)

k 均值算法 k-Means

最小化平方误差 (蔟内样本围绕粗均值向量的紧密程度，公式理解是簇内哥哥向量的平均值) E 越小簇内样本相似度越高。

NP 难问题，采用贪心策略，迭代优化。(算法步骤中如何进行初始化？初始化的值如何确定？利用随机选择的几个样本作为初始向量)

学习向量量化 LVQ

假设数据样本带有类别标记，学习过程利用样本的监督信息来辅助聚类。

随机选取一组原型向量，每次选取与之距离最小的向量进行更新原型向量。(利用学习率)

（Voronoi 剖分）

高斯混合聚类

采用概率模型来表达聚类原型

高斯分布与 n 维均值向量和协方差矩阵两个参数确定。簇划分由原型对应后验概率确定。

采用极大似然估计求解模型参数。且常采用 EM 算法进行迭代优化求解。

（这个为什么会想到利用高斯分布的混合分布来作为模型？比较巧妙）

9.5 密度聚类

假设聚类结构能通过样本分布的紧密程度确定。从样本密度的角度来考察样本之间的可连接性，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。

DBSCAN 一种密度聚类算法，基于一组”邻域“参数（e，MinPts）来刻画样本分布的紧密程度。

e 邻域：样本及中与某一个样本 xj 距离不大于 e 的样本

核心对象 若 xj 的 e 邻域中样本数大于等于 MinPts，则 xj 是一个核心对象

密度直达 xj 位于 xi 的 e 邻域中

密度可达 存在样本序列依次密度直达

密度相连 对 xi，xj，若存在 xk 使得 xi 与 xj 均有 xk 密度可达，则 xi 与 xj 密度相连

簇：由密度可达关系导出的最大的密度项链样本集合。(简单理解就是可以通过一个核心对象进行密度可达放射)

9.6 层次聚类

试图在不同层次对数据集进行划分，从而形成树形的聚类结构。"自底向上"聚合策略/"自顶向下"分拆策略。

ABNES 自底向上。Dmin 两个簇的最近样本 "单链接"

Dmax 两个簇的最远样本 "全链接"

Davg 两个簇的所有样本"均链接"

(由图可知该算法可以通过自底向上的虚线划分得到不同个数的聚类簇结果，其中虚线越靠上，所分出的簇个数越少)

第十章 降维与度量学习

10.1 k 邻近学习 kNN

基于某种距离度量找出训练集中与其最靠近的 k 个训练样本。再根据这 k 个样本出现最多的类别标记进行预测（少数服从多数）

懒惰学习的代表：没有显示的训练过程，收到测试样本才进行学习处理。

(简单但是功能不逊色——只比贝叶斯最优分类器的错误率高不了一倍)

10.2 低维嵌入

"维数灾难"——高维情形下出现的数据样本稀疏，距离计算困难等问题。

缓解途径：降维 （与学习有关的只是一个低维子空间的嵌入

多维缩放：MDS 通过矩阵变化得到一个样本的低维坐标形成的矩阵

线性降维法：基于线性变换。

10.3 主成分分析 PCA

分析：超平面的性质：最近重构性（理解为样本点到这个超平面的距离都足够近）最大可分性

降维后的迪维空间维数 d'可由用户指定或者通过迪维空间中 kNN 交叉验证选取更好的 d'值。

(降维也有去噪的效果)

、

10.4 核化线性降维

线性降维可能会丢失原本的低维结构。

核化 kernelized 是非线性降维

KPCA 核主成分分析 （引用了核函数，具体推导大致是用核矩阵表达映射至高维空间的ϕ的内积等值）

10.5 流形学习

借鉴拓扑流形概念。可以利用欧式距离，可以可视化展示

等度量映射 Isomap（保持近邻样本之间的距离）：测地线距离：无法在多维空间计算直线距离（不合理误导性），利用近邻连接图计算两点之间的最短距离（Dijkstra 算法或 floyd （数据结构中用到））

近邻图的构建：1、指定 k 近邻点个数，得到 k 近邻图

2、指定距离阈值 e

局部线性嵌入 （保持领域内样本之间的线性关系）

10.6 度量学习

降维的目的是找到一个合适的低维空间即合适的距离度量。度量学习就是学习出一个合适的距离度量。

马氏距离：考虑到不同的属性之间相关性，对应坐标轴不再正交。

W 为属性权重所得的对角矩阵。M 为半正定对称矩阵（度量矩阵）。度量学习即是对 M 进行学习。

简单的多数投票改为概率投票。

第十一章 特征选择与稀疏学习

11.1 子集搜索与评价

特征选择（feature selection)的原因：1、缓解维数灾难问题 2、降低学习任务的难度

无关特征(irrelevant feature) 与当前学习任务无关

冗余特征(redundant feature) （但是有时存在可以降低学习任务的难度 “中间概念”）

子集搜索问题：1、评价候选子集的好坏 2、如何根据评价结果获取下一个候选子集？

从子集内特征个数为 1 开始不断增加，直至 k+1 的所有后选择及都不如上一轮的选定集，则停止。（前向搜索）

还有后向搜索和双向搜索（双向搜索是从子集元素为 1 开始还是从完整特征集开始？如何做到选定就不再删除？）

对属性子集，可以根据其取值将 D 划分为 V 个子集，通过比较这个划分与真实划分的差距，可以得到对树形子集的评价（差异越小对分类越有帮助）。

信息熵、决策树（决策树其实就是在不断的利用树结点进行直观的属性划分）

三大类特征选择方法：过滤式、包裹式、嵌入式。

11.2 过滤式选择

先对数据集进行特征选择，再训练学习器，特征选择与后续学习无关。

Relief 过滤式特征选择方法：求得一个统计量的向量，每个分量对应着一个初始特征。特征所对应的相关统计量分量之和确定特征的重要性。

相关统计量：猜中近邻 猜错近邻 ，若对一个属性（与猜中近邻的距离小于与猜错近邻的距离），该属性有益，增大该属性对应的统计分量。时间开销随采样次数以及原始特征数线性增长。（运行效率很高的过滤式特征选择算法）

Relief 二分类 Relief-F 多分类

11.3 包裹式选择

把最终将要使用的学习器的性能作为特征子集的评价标准。（需要学习之后得到学习性能，量身定做，目的在于最终的学习性能，目的性更强，所以包裹式特征选择比过滤式特征选择好，但由于训练次数多，计算开销大得多）

LVW 典型包裹式特征选择方法，拉斯维加斯方法（有趣：赌城命名的随机化算法，赌博）使用随机策略进行子集搜索。评价准则：最终分类器的误差。使用的是在数据集上交叉验证的方法估计学习器的误差。

11.4 嵌入式选择与 L1 正则化

将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成。

引入 L2 范数正则化，“岭回归”式子，能显著降低过拟合的风险。

L1 范数 “LASSO”式 额外好处：更容易获得稀疏解，求得的 w 会有更多的零分量。（这里没有说明 w 指什么，按照线性回归问题应该是指各个特征分量的权重值）

最终正则化求解的结果是 w 对应非零分量的特征作为最终模型中选择的特征。（即完成了特征选择）

求解方法：近端梯度下降法：PGD

11.5 稀疏表示与字典学习

样本具有稀疏表达形式的好处：先行支持向量机等线性可分问题更易于解决。（但是稀疏性的好处应该是相对的，稀疏性也会导致一些障碍（过度稀疏））

字典学习（dictionary learning）亦称 稀疏编码（sparse coding）往往是在同一个优化求解过程中完成，故统称字典学习）

利用变量交替优化的策略反复迭代进行更新字典矩阵 B 的学习。

11.6 压缩感知

令采样频率达到模拟信号最高频率的两倍，数字信号完全保留模拟信号的全部信息（这个理论非常神奇。达到频率要求时。数字离散信号可以表达连续模拟信号。现实的应用也很多）

压缩感知：对接收方基于收到的信号精确的重构源信号。

利用变换处理使得数字信号成为在频域上的稀疏信号。字典和稀疏基可以使未知因

素的影响大为减少。

限定等距性：RIP（这个缩写有歧义..）对任意向量 S 和 A 的所有子矩阵，A 满足 k 限定等距性。问题变为 L1 范数最小化问题。（范数问题是否和 NP 难问题有所联系..）

第十二章　计算学习理论

12.1 基础知识

泛化误差、经验误差、误差参数（泛化误差的上限）、不合

由于训练集使某个隐含未知分布的独立同分布采样，所以 h 的经验误差的期望等于其泛化误差。（这个结论是否有推导？概念上理解是由于独立同分布抽样，所以训练样本和测试样本不会由很大的出入）

Jensen 不等式

Hoeffding 不等式

McDiarmid 不等式

12.2 PAC 学习

概率近似正确学习理论 Probably Approximately Correct

概念（c）（理解为之前章节中的映射，所有可能概念的集合为假设空间）

学习算法可分/一致　不可分/不一致　（是否可以将所有示例按与真实标记一致的方式完全分开，线性可分）

三个定义：

PAC 辨识　泛化误差的不等式

PAC 可学习

PAC 学习算法　算法的运行时间是多项式函数 概念类 C 是高效 PAC 可学习。

样本复杂度 满足 PAC 可学习的所需的最小的复杂度

若概念空间与概念类完全相同，恰 PAC 可学习

（但是现实中不太合理，很难满足）

12.3 有限假设空间

可分情形：

目标概念存在于假设空间中。

根据训练集不断从假设空间中排除，直至满足误差参数要求。

有限假设空间都是 PAC 可学习的。

不可分情形：

目标概念不存在于假设空间中。

无法学的目标概念的 e 近似，只能找泛化误差最小的假设，找出此假设的 c 近似。

不可知 PAC 可学习：

对所有分布都存在学习算法和多项式函数满足最小的样本复杂度要求。

12.4 VC 维

无限假设空间的可学习性研究，需要度量假设空间复杂度：VC 维

增长函数 训练样例的个数 m 增大，假设空间 H 中所有假设对训练集示例所能赋予标记的可能结果数增大。 越大，H 的表示能力越大。

对分 H 中假设对示例赋予标记的没中可能结果称为对 D 的一种对分

散打

若假设空间能实现示例集 D 上的所有对分，则称示例集 D 能被假设空间 H 打散。

VC 维：能被 H 打散的最大示例集的大小。

VC 维与增长函数有密切关系。

基于 VC 维的泛化误差界是分布无关、数据独立的

定理：任何 VC 维有限的假设空间 H 都是（不可知）PAC 可学习的

12.5 Rademacher 复杂度

是区别于 VC 维的另一种刻画假设空间复杂度的途径，在一定程度上考虑了数据分布。

经验 Rademancher 复杂度衡量函数空间 F 与随机噪声在集合 Z 中的相关性。(集合 Z 指实值函数空间的定义域。

12.6 稳定性

VC 维和 Rademancher 复杂度都是脱离具体算法，而是对问题本身的性质描述。

“稳定性”考察的是，对于一个算法，输入发生变化时，输出是否会随之发生较大的变化。

损失函数刻画了假设下的预测标记和真实标记之间的差别。

泛化损失

经验损失

留一损失

均匀稳定性 只根据算法自身的特性来讨论输出。

经验风险最小化：ERM 学习算法所输出的假设满足经验损失最小化，则称算法是满足 ERM 的。

结论：若学习算法是 ERM 且稳定的，则假设空间 H 可学习。(稳定性和假设空间的可学习性由损失函数联系起来。)

第十三章　半监督学习

13.1 未标记样本

主动学习：使用尽量少的“查询”来获得尽量好的性能。但也引用了额外的专家知识。

利用未标记样本：聚类假设，

半监督学习：流形假设输出没有限制，所以比聚类假设的适用范围更广。（但本质相同，　都是相似的样本具有相似的输出）

版监督学习分类：纯半监督学习　（假定训练数据中的未标记样本并非待预测数据，开放世界）

直推学习（假定学习过程中所考虑的未标记样本恰是待预测数据，封闭世界）

（所以纯半监督学习才是真正意义上的具有泛化能力）

13.2 生成式方法

直接基于生成式模型的方法。假设所有数据都是有一个潜在的模型生成的。则学习目标是求得潜在模型的参数。（EM 算法和极大似然估计，但是学习的区别取决于不同生成式模型的假设）。

（方法简单易于实现，标记数据极少情况下往往性能更好。但是需要模型假设必须准确。现实任务难以做到）

13.3 半监督 SVM

TSVM 是针对二分类问题的学习方法。考虑对所有未标记样本进行分别正反假设，求得在所有情况下最大间隔划分超平面。（搜索标记指派）

（但是穷举过程非常耗时，低效，仅仅适用于未标记样本很少时）

13.4 图半监督学习

给定数据集，映射成一个图。每个样本对应于图中一个结点，两个结点之间的边对应于两个样本之间的相似度。

半监督学习过程对应于“颜色”在图上扩散的过程。（类比）

边集：亲和矩阵（关系的强度越亲和）　求关于 f 的最小能量函数

多分类问题的标记传播方法

求得传播矩阵，迭代至收敛得到 F*可以获得数据集中样本的标记。

通过矩阵运算的分析来探索算法性质。存储开销大很难处理大规模数据。而且难以判断新样本在图中的位置，需要引入额外的预测机制。

13.5 基于分歧的方法

使用多学习器，学习器之间分歧对未标记的数据利用至关重要。

“协同训练”是此类方法的重要代表，也是“多视图学习”的代表。类似于“互相学习、共同进步”，要求两个视图充分（每个试图都包含足以产生最优学习器的信息“条件独立（给定类别标记条件下两个试图独立）

多视图数据：一个数据对象拥有多个“属性集”，每个属性及构成一个“视图”(view)

（不同视图具有“相容性”，即包含的关于输出空间 y 的信息是一致的。不同信息的互补性给学习器的构建带来很多便利）

对于单视图可以利用不同的弱学习器之间的差异（分歧）进行互相提供伪标记样本来提升泛化能力。

采用合适的基学习器，而且找到具有显著分歧、性能尚可的做到这一点不是很容易。

13.6 半监督聚类

聚类问题中获得的监督信息：一、必连 样本必属于同一个簇，和勿连约束 样本不需要属于同一个簇。二、少量的有标记样本。

约束 k 均值算法利用第一类监督信息。(k 均值算法的扩展)

利用一致的标记样本作为种子（聚类中心）不断迭代中改变种子样本的隶属关系。得到约束种子 k 均值。

第十四章　概率图模型

14.1 隐马尔可夫模型

概率模型：学习任务归结于计算变量的概率分布，利用已知变量推测未知变量的分布称为”推断“。核心：如何基于可观测变量推测出未知变量的条件分布。

生成式——联合分布

判别式——条件分布

得到其中之一。

概率图模型：用图来表达变量相关关系的概率模型。用一个节点表示一个或一组随机变量。节点之间的边表示变量间的概率相关关系。

概率图模型分类：有向无环图 有向图模型/贝叶斯网

无向图表示 马尔可夫网

隐马尔可夫网：结构最简单的动态贝叶斯网
状态变量（隐变量，不可观测的） 状态空进
观测变量 观测空间

马尔可夫链 系统下一时刻的状态仅有当前状态决定，不依赖以往的状态

隐马尔可夫模型三组参数：状态转移概率 输出观测概率，初始状态概率
条件独立性可以解决三个基本问题： 如何评估模型与观测序列之间的匹配程度 如何根据观测序列推断出隐藏的模型状态 如何训练模型使其能最好的描述观测数据。

14.2 马尔可夫随机场 MRF

是典型的马尔可父网，无向图模型。
势函数/因子 定义在变量子集上的非负实函数。
团 对于图中节点的一个子集，其中任意两个节点间都有边连接。(连通集？)
极大团

多变量联合概率分布基于团分解为多个因子的乘积。

全局马尔科夫性：给定两个变量子集的分离集，这两个变量子集条件独立。(C为 AB 的分离集，给定 C，AB 独立)
推论：
独立马尔可夫性
成对马尔可夫性

势函数的作用：定量刻画变量集之间的相关关系。非负函数（常用指数函数）偏好的变量上取值更大。

14.3 条件随机场 CRF

判别式无向图模型 （判别式模型，而之前两个时生成式模型）
对多个变量在给定观测值后的条件概率进行建模。

标记变量也可以是结构性变量（如输出序列或者语法树等树形结构）

势函数、团、特征函数（刻画观测序列对标记变量的影响关于经验特性）

（条件随机场和马尔可夫随机场在形式上很大差别。区别在于联合概率和条件概率。）

14.4 学习与推断

边际分布：对无关变量求和或者积分后的结果。边际化（积分过程）

推断问题的目标：计算边际概率/条件概率 （将变量集分为两个不相交的变量集合）

精确的推断方法：
变量消去
把多个变量的求和问题转化为对部分变量交替进行求积/求和问题。简化计算。
明显缺点：冗余计算，会重复（对于计算多个边际分布而言）

信念传播
把求和操作看作传递信息，解决了重复计算问题。
结点的边际分布正比于它所接受的消息的乘积

14.5 近似推断

更常用，计算开销较小。
分类：采样 随机化完成近似 使用确定性近似完成近似推断 （变分推断）
1、MCMC 采样 马尔可夫链蒙特卡罗
通过抽取样本计算均值直接近似目标期望。(理论是大数定律)
采样需要基于图模型所描述的概率分布实施。

MH 算法 拒绝采样来逼近平稳分布 需要重复足够多次
吉布斯采样是 MH 算法的特例。

2、变分推断

通过使用已知的简单分布来逼近所需要推断的复杂分布，通过限制近似分布的类型得到一种局部最优但具有确定解的近似后验分布。
概率模型的表示方法：盘式记法（更简洁）
变分法关键：如何对隐变量进行拆解，假设各变量子集符合什么分布。

14.6 话题模型

生成式有向图模型处理离散型数据
隐狄利克雷分配模型 LDA
词 word 文档 document 话题 topic
文档：由一组词组成，不计顺序的表示方式。

话题：一系列相关的词以及他们在该概念下出现的概率。
词袋：
生成文档的过程。（话题根据话题比例产生）
词频是唯一已观测变量。求得参数可以根据词频来推断文档集所对于的话题结构。

第十五章　规则学习

15.1 基本概念

规则 rule：予以明确能描述数据分布所隐含的客观规律或领域概念。
规则体（蕴含符号右边）规则的前提
规则头　（蕴含符号左边）　规则的结果

逻辑文字 literal fk L 规则体中逻辑文字的个数 规则的长度
（规则学习比神经网络、支持向量机（黑箱模型）更具有可解释性，数理逻辑具有极强的表达能力）
逻辑规则的抽象描述能力可以处理高度复杂的 AI 任务，如问答系统。
符合规则=被规则覆盖
（未被覆盖不一定不满足要求，但被要求的否覆盖一定满足不要求）

冲突：一个示例被判别结果不同的多条规则覆盖。
冲突消解：投票法、排序法、元规则法
默认规则：规则学习算法常用的方法，用来处理规则集合未覆盖的样本。

规则分类：（形式语言表达能力）
“命题规则”由原子命题和逻辑连接词构成的简单陈述句
“一阶规则”能描述事物的属性或关系（具有谓词、量词）
（一阶规则能表达复杂的关系“关系型规则”，命题规则是一阶规则的特例）

15.2 序贯覆盖 sequential covering

逐条归纳 不断学习新的规则 不断去除可以覆盖的训练样例。也称“分治”策略。
（是穷尽搜索的做法，在属性和候选值较多时会由于组合爆炸而不可行）

两个策略产生规则：
1、自顶向下 由一般规则开始（空规则） 添加新规则缩小覆盖范围 生成-测试 特化过程 容易产生泛化能力好的规则 对噪声的鲁棒性更强 命题规则学习中通常使用
2、自底向上 从特殊的规则开始 删除文字 数据驱动 泛化过程 适合训练样本较少的情况 在一阶规则学习这类假设空间非常复杂的任务使用较多

评估规则优劣的标准：先考虑规则准确率，再考虑覆盖样例数、再考虑属性次序

局部最优（过于贪心）解决方法：集束搜索 一次保留最优的 b 个逻辑文字。

15.3 剪枝优化

规则生成——贪心搜索 需要有机制缓解过拟合风险。
预剪枝、后剪枝
剪枝可以借助统计显著性检验进行 CN2 算法 似然率统计量（LSR）
后剪枝——减错剪枝 REP 进行多轮剪枝而且穷举最好的规则集
IREP 单挑规则剪枝，更高效

RIPPER 剪枝和后处理优化相结合

15.4 一阶规则学习

关系数据：样例间的关系
引入领域知识：构造出新属性 或 基于领域知识设计某种函数机制（正则化）对假设空间加以约束。FOIL 著名已解决系算法。序贯覆盖+自顶向下 考虑不同的变量组合。

命题规则学习与归纳逻辑程序设计的过渡 比一般的归纳逻辑程序设计技术更高效。

15.5 归纳逻辑程序设计 ILP

在一阶学习中引入函数和逻辑表达式嵌套 （但是会带来计算上的巨大困难）
使得机器学习具备了更强大的表达能力
看作机器学习技术解决基于背景知识的逻辑程序归纳。

15.6 最小一般泛化 LGG

将归纳逻辑程序设计中的自底向上的泛化能力技术。
初始规则选择方法 常用 RLGG
LGG 是特化为一阶公式的最特殊的一个。

15.7 逆归结

机器学习：归纳的过程
归结原理：一阶谓词演算中的演绎推理能用一条十分简洁的规则描述。

逆归结：反过来发明新的概念和关系。
归结逆归结扩展为一阶逻辑形式的操作有两种：
置换　用某些项来替代逻辑表达式中的变量
合一　用一种变量置换令两个或者多个逻辑表达式相等

逆归结可以自动发明新谓词（知识的发现与精化）