6. 其它线性方法1：近邻法

1 一近邻，KNN，无限样本的错误率：，P为无穷样本的错误率

问题：存储量和计算量、票数接近或者有噪声的时候风险大、样本无穷的时候性能优越，但是有限样本

2 快速算法：样本分成多个子集，每个子集可用几个代表，先比代表，最后一个代表再逐个比

2.1 分支定界算法

新样本x，节点，则x的近邻不可能在中；，不是x的最近邻

2.2 剪辑近邻法

分为考试集和参考集，用参考集对考试集分类，去掉考试集分错的，再用剩下的进行分类

2.3 压缩近邻法

开始只有一个样本，每次一个样本用分类，如果分错就放进，最后用分类

2.4 Prototype近邻法

先对每类KMeans，中心点为代表，用中心点一近邻

2.5 可拒绝近邻法

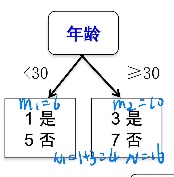
K近邻票数才同意，否则拒绝

3 距离计算

归一化；不同距离度量：闵可夫斯基距离 余弦距离：

其它线性方法2：决策树

3.1 ID3算法

信息增益, ，

3.2 不纯度度量

基尼不纯度：

误差不纯度：

3.3 C4.5

，连续样本离散化为k个取值，再用决策树

3.4 CART算法

对连续取值每次只能二分，构建回归树，选x的阈值让最小

3.5 避免过学习：剪枝 预剪枝：边生长边验证，划分前后验证机提升显著，可以划分；后剪枝：先生长，再剪枝，剪枝后比剪枝前验证集显著提升，可以剪枝

其它线性方法3：集成学习

1 bagging 训练集上有放回抽样产生m个数据集，每个数据集训练分类器，再投票（并行）

2 随机森林 样本同上，每次对特征进行采样，最后决策树投票（并行）

3 boosting 训练新学习器拟合残差

3.1 AdaBoost 首先每个样本有同样权重，训练分类器计算错误率，分类器权重，令，并归一化权重，作为下一次训练的样本权重

3.2 Gradient Boosting

每次用数据集训练一个新学习器，计算更新权重，更新模型。

3.3 XGBoost

加入正则化项，把L在处泰勒展开，变成一个关于的二次函数

4 集成学习投票方式：多数投票；加权投票（对分类器加权）；排序输出（Borda）;概率输出（分类器输出概率）

其它线性方法4：非线性回归

1 线性回归的扩展：引入交叉项或者非线性变换

2 定性变量回归：引入 dummy variable

3 最近邻回归 k个近邻取值加权求和，权重：1/k，1/d，

回归树、支持向量回归

7. 贝叶斯决策理论

1 概念 类条件概率密度; 后验概率, 错误概率如果x属于；平均错误率 。

2 最小错误率贝叶斯决策

条件：类别数一定；和已知

目标：最小化平均错误率，相当于最小化每个x的错误率

决策方法：，或用贝叶斯公式，。

或采用似然比方法：，，则。

错误率计算：

3 最小风险贝叶斯决策

损失函数：状态为，做出决策的损失

条件期望损失：

最小风险的贝叶斯决策：

4 正态分布下的贝叶斯决策：是正态分布，推导：直接代公式

4.1 协方差矩阵相等，且为对角阵，方差相等

先验分布相等：欧氏距离+分类面线性；先验分布不等：欧氏距离+分类面线性

4.2 协方差矩阵相等

马氏距离；分类面线性

4.3 一般情况

分类面为超二次曲面

8 概率密度函数的估计

能用样本估计分布的前提：1. 训练样本能代表真实分布，iid；2. 有充分的样本

1 概念

参数估计：已知概率密度函数的形式，只是估计参数

统计量：样本的函数，用来作为对参数的估计

点估计：估计某个参数的取值

区间估计：估计某个参数的范围

2 最大似然估计

假设：样本都是从某个分布中独立抽取出来的（等）

似然函数：

3 贝叶斯估计

目的：估计某个概率分布的参数

损失函数：，期望风险：，最小化R即最小化

结论：如果损失函数为平方损失时，

计算步骤：确定的先验分布；求样本的联合分布；求的后验概率分布；求的贝叶斯估计量

4 贝叶斯学习

目的：做样本的在线学习迭代

方法：

5 直方图方法（下面开始非参数估计）

，N是样本总数，V是某个仓的体积，k是落入这个仓中的样本数

5.1 kN近邻估计

找k近邻的大小，

5.2 Parzen窗法

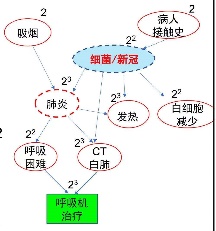
，k表示窗函数，对x的贡献程度

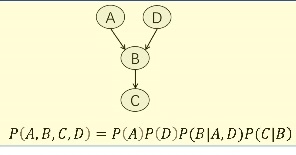
窗宽应该随着样本数目的增大减小，窗宽越宽，密度函数变化越迟缓

9 贝叶斯网络和隐马尔可夫模型

1 贝叶斯网络

希望能够生成模型，因此求联合分布很重要

条件独立：x,y条件独立于z，则；链式法则：

* 1. 概率图

概率图参数量计算

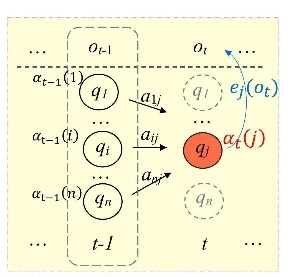
贝叶斯网络写成概率形式

1.2 马尔可夫

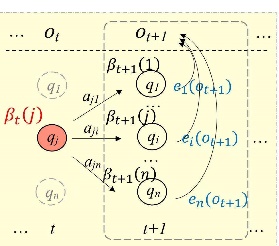
马尔科夫链：

1.3 贝叶斯网络

利用贝叶斯网络作分类：，x是一个时间序列，通过S的取值和阈值分类

1. HMM
   1. HMM建模

整体概率，e为x1状态观察到o1的概率，为初始概率，为隐状态转移概率

* 1. 动态规划

前向算法：，t时刻隐状态为j的概率为前一时刻所有隐状态的概率乘各自的转移概率求和，加上本时刻对应隐状态的发射概率，复杂度

后向算法：

* 1. Viterbi算法：用于HMM解码隐变量

，变前面的隐状态序列，让当步状态为j的概率最大

终止结果，复杂度

* 1. HMM训练问题：已知观测序列，模型状态集和观测值的取值范围，求模型参数：转移矩阵，发射概率，初始概率

利用动态规划求

估计概率：E：，M：

估计E：E：，M：

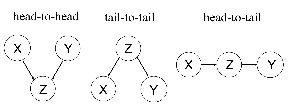
1. 贝叶斯网络

定义：有向无环图

* 1. 朴素贝叶斯分类器

假设：各个分量条件于独立。判别函数

拉普拉斯平滑：, C是类别数，S是x的所有取值可能数

* 1. DAG上的贝叶斯推理

第一个Z未知时XY独立，剩下Z已知时XY独立

分析：

* 1. 贝叶斯网络推断与决策问题：习题（P28）
  2. 贝叶斯网络的学习
     1. 参数学习：已知模型结构，但是不知道模型的参数值

最大化（或对数似然）

无缺失数据的参数学习：按照上式分解为每个参数值的条件概率比较求解

先验分布的选择：选择共轭先验，先验乘似然函数之后的后验概率仍和先验有相同形式

例题：P34

有缺失值的参数学习：通过EM算法求解

E：通过当前参数估计缺失值

M：通过求出的缺失值估计新的参数

例题：P35

* + 1. 结构学习

最大似然，由于D相等，假设p(G)相等，最大化

d为模型参数数量，N为边的数量

10 特征选择

学习算法的方差和偏差分解

是数据真实值，y是观测值，是数据有限时的估计函数，当数据趋于无穷时，趋于f，

分析高维数据的方法：SVM；特征选择；正则化方法

相关名词：特征形成（原始特征）直接观测到的或者经过初步运算的特征；特征选择：从原始特征中直接选出来的；特征提取：把原来的特征变成新的特征

特征选择的目的是降低Variance

特征选择的好处：有利于可视化和理解；降低数据采样和数据存储的开销；降低模型使用时的计算开销；避免维数灾难

10.1 特征选择

最优子集：遍历所有特征组合，找最好的

过滤法：计算输入和输出之间的相关性，选相关性最好的特征；不能在测试集上选特征

Fisher：；相关系数：；互信息：；基于熵的判据：计算依照某一维特征分类的熵；KL散度：；统计检验：t检验

包裹法：需要有验证集，不能在测试集上选特征

前向选择：每次选一个最好的特征加入，复杂度，p为模型个数，维度小于p时也可以适用这个方法

后向选择：每次减少一个特征，选择减少特征后表现降低最多的

混合算法：先加特征，再减，再加。。。

特征选择的遗传算法：

1. 初始化染色体：从p个特征中选k个，有种
2. 计算每条染色体的适应度
3. 按照某种概率模型进行采样，适应度越高的越可能被采到

10.2 模型选择

AIC=2d-2ln(L)；BIC=ln(n)d-2ln(L)，d：模型中拟合参数的数量；L：似然函数的极大值；n：样本数

随机置换法：随机打乱样本的标记，通过同样的学习过程，统计出没有分类信息情况下识别正确率分布，得到真实值的p值，如果p<0.05，则显著

10.3 嵌入法：正则化方法

11 特征提取和数据可视化

11.1 PCA

，a是的第n大特征值对应的规范化特征向量，第二大推导：对a2求导后左右同乘

数据压缩：???

奇异值分解计算PCA：对X作UV分解：，选U的前k列作为特征向量组成矩阵

KPCA：把改成

11.2 KL变换

KL展开：用无穷个的正交归一化向量系作基，展开x，用有限项来逼近x

KL变换：最小化MSE预测误差，变换矩阵也是的前d个特征向量；当求取特征值的矩阵为协方差矩阵时，KL变换等价于PCA（求特征值分解的矩阵可以不同）

监督KL变换：计算类内离散度矩阵；对作KL变换，计算新特征，方差为，计算???，其中为类间离散度矩阵，并用J(yi)排序，选前d个分量作为新的特征向量，uj组成变换阵

PCA例题：P16

监督KL变换例题：P16

11.3 流形学习

认为高维冗余数据是d维数据在D维空间中扭曲的结果。

LLE：认为局部关系在低维空间中也满足

先在高维空间中找X的一个近邻X’，并学习利用X’重建X的权重W；在低维空间中学习固定W，学习重建误差最小的Y

Isomap：认为距离较近时可以利用欧氏距离度量，距离较远时应该用一条最短路径上的欧氏距离之和来度量，重新形成距离矩阵，MDS降维

11.4 高维数据的低维显示

1. PCA

2. MDS：多维尺度法，把样本之间的距离关系在低维中表示出来

3. t-SNE：利用概率分布度量样本之间的距离，把高位空间中的欧氏距离转化为条件概率密度函数

，认为ij距离越近，在i分布中抽到j的概率越大，完整对称定义：；低维概率建模：，优化目标是让高维和低维的KL散度最小，梯度下降求解

困惑度：数据密集的地方上述方差应该小，分布的信息熵随方差的增大而增大，故可以通过困惑度反推方差是否应该增大或减小。

12 聚类分析

12.1 不同距离度量

欧式距离；曼哈顿距离；闵可夫斯基距离；余弦距离；编辑距离

12.2 动态聚类（典型K-Means）

要点：选某种距离作为相似性度量；定义某个准则函数；初始化分类方法和迭代算法

K-Means，，类的初始划分：经验；随机分成k类，选各类中心作为代表点；密度法：选密度最大的作为第一个中心点，离该点一定距离外密度最大的作为第二个；前k个点作为中心点；先k-1聚类，最后一个中心点找离剩下k-1个最远的

缺点：要求类别数已知；最小方差划分，不一定符合内部规律；对类别的硬划分；和初始点有关；和距离度量方式有关

12.2.1 基于核的相似性度量：定义x到之间的距离，准则函数，相当于中心点，相当于距离

比如正态核函数

12.2.2 模糊K均值方法

隶属度函数在0-1取值，表示属于A类的程度

模糊K均值算法：损失函数；求解：分别对m和求偏导，固定一个求另一个迭代？？求解

迭代步骤：初始化中心；用m算；用算m；去模糊化：

12.3 高斯混合模型GMM

假设数据为多个高斯分布的和

求解：E-M算法

E：

M：后验概率最大化：

解得：

样本类别划分：

12.4 分级聚类（画树）

常用的类间距离度量：最大/最小/均值

分级方法：找相似度最大；按类间距离计算新的距离矩阵

12.5 一致聚类

进行S次采样，对每个集合计算不同类别数时的聚类结果，整体一致性矩阵，新的距离度量。

衡量聚类一致性：，用CDF曲线下的面积衡量一致性，画出，找到肘点。

12.6 SOM

竞争：，winner。

权值学习：

聚类：画出相邻神经元的距离矩阵，距离较大的地方分成两类

12.7 聚类结果评估

基于互信息的分数：，U为真实标签，V为聚类结果，MI越高，聚类效果越好

，值越大越好

轮廓系数：，为样本i和同类其它样本的平均距离，b(i)为样本i和距离它最近的另一个聚类中样本的平均距离

在-1到1，取值越大越好

13 深度学习

13.1 计算图梯度下降

13.2 卷积

卷积层参数个数计算：卷积核大小乘数目+1（偏置项）

输出图的大小计算：，n是图像尺寸，k是卷积核尺寸，s是扫描步幅，p是边衬大小

13.3 网络设计

通常用多个小的卷积核达到相同的感受野，替代一个大的卷积核

Alexnet：使用了最大池化层；用了ReLU；用了dropout；用了数据扩增

VGG：增加了网络的深度，用小的核替代大的核

Googlenet：同一个卷积层用多个尺寸的卷积核；用1\*1卷积融合多通道的信息

残差网络：解决梯度消失的问题，直接搭建前向通路；用了skip connection；用了BN

Densenet：多个特征拼在一起，skip connection

EfficientNet：模型深度（更复杂的特征）、宽度（更精细的特征）、图片分辨率应该成比例

13.4 反馈神经网络

RNN：所有隐层共享参数

LSTM：cell state：前向通道；遗忘门：决定前向通道记得多少；输入门：处理输入数据：输出门：综合前面几个输出

GRU：LSTM的简化

13.5 常用训练技巧

训练顺序：整体梯度下降；随机；批量

优化器：SGD；momentum：；Adam：

归一化：BN，归一化+线性变换，加快网络收敛，但是不要抹掉之前的学习结果

Dropout

14 表示学习

AE训练：把每两层的节点作为一个RBM网络，用CD算法，编码器权值对称的到解码器，最后再整体进行finetune

14.1 自然语言处理的表示学习

不用onehot表示：没有包含语义信息；词向量维度过高；信息过于稀疏

Word2vec-skip-gram：输入onehot，通过线性矩阵映射到vec，再通过线性矩阵恢复该词的语境向量（按softmax，概率表示）

Cbow：输入语境，预测词语出现概率

GPT-3：网络结构更好；基于上下文的表示方法（不同上下文词语含义可能不同）；模型参数增加

14.2 网络图的表示学习

节点表示：两个顶点处于同一个社区或具有相同的功能，应该有相同的表示

Deepwalk：某个顶点随机游走，生成一堆序列，再word2vec

图表示：两个结构相似的图应该有相似的表示

社团检测：利用节点表示的距离矩阵进行聚类

15 深度生成模型

15.1 生成式模型的基本思路

从未知分布中采样，产生新的样本

15.2 DBN

RBM：只有一个隐层，隐层单元之间没有连接；给定可见单元，隐层单元条件独立

RBM训练的CD算法：用h和x互相生成无穷多次，就可以认为是平衡时的采样

DBN：类似MLP结构的RBM，sigmoid激活

问题：最后一层由于倒数第二层确定变得不独立；把最后一个隐层变成玻尔兹曼机，从而解除依赖关系

训练：CD，从x层开始往后推

15.3 VAE

目标函数：

再参数化技巧：直接输出分布不能BP，所以在正态分布中采样并通过估计的参数进行平移，再BP

15.4 GAN

目标函数：，生成器：最小化数据和生成数据分布的差异

mtGAN：用于病例样本生成；DCGAN：把卷积引入生成器；cycleGAN：图像来回迁移，求相似程度最小

16 迁移学习、半监督学习、小样本学习

好处：训练起点高、训练快、更好的训练效果

16.1 基于样本的迁移学习

找到其它数据集中和训练数据类似的样本，加入训练数据。

TrAdaBoost：target数据集中数据被错误分类，加大权重；source被错误分类，减小权重

16.2 基于网络的迁移学习

预训练

小样本的时候效果好

16.3 半监督学习

假设：样本距离相近，类标相同；分界面不会穿过高样本密度区域；同一个低维流形中的样本有相同的标签

自训练算法：先有监督，再通过网络标注一部分无类标样本，再重新训练，直到所有样本都被标注

基于图的方法：画出一张图，K近邻有权重，有标签可以通过边影响其它无标签样本

半监督SVM：分界面要让没有类标的样本产生的分类面间隔最大化

16.4 小样本学习

估计误差：由于样本量不足造成的参数误差；近似误差：原空间本身的问题导致误差，无法通过数据改变

16.4.1 基于模型的方法

基于先验知识定义模型，减少模型的搜索范围：比如对数据降维再训练

16.4.2 基于数据的方法

基于先验扩充数据集：数据增强

16.4.3 基于算法的方法

找到一条更好的搜索路径：在大数据集上预训练，在小数据集上做微调