

小抄：重要公式，算式，算法，记不住的概念

习题课题目，过一遍20，21作业，网上找的题查缺补漏

根据19押题

不确定的题：

时序差分更新表达式

Q-learning

图半监督： $E(f)$ ，半正定二次型证明，假设怎么用在最小化里面的，有标记和未标记使用sign求VC维

永远要考的PCA专题

PCA弱点

- 利用不了有标记的样本 (TPCA)
- 只适用于线性 (KPCA)

PCA&LDA相同与区别

- 均使用了矩阵特征分解的思想
- 都假设数据符合高斯分布

不同点：

- 1) LDA是有监督的降维方法，而PCA是无监督的降维方法
- 2) LDA降维最多降到类别数 $k-1$ 的维数，而PCA没有这个限制。
- 3) LDA除了可以用于降维，还可以用于分类。
- 4) LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向

PCA降维第一步必须标准化

- 防止投影后数值大小对特征重要性的影响
- 有利于梯度下降法的收敛。

PCA (降维) 和relief (特征选择) 区别

- 特征选择是去除无关特征和冗余特征，没改变特征
- 降维是将特征映射到新的低维空间

十一、

为什么特征选择？

- 去除不重要的，避免维数灾难
- 去除不相关的，降低任务难度

怎么进行特征评价？

通过估算特征子集和样本标记的差异

比如计算属性子集A的信息增益，越大，说明A中包含有助于分类的信息越多

三种特征选择方法有什么区别

过滤式：预训练时先进行特征选择，和训练过程没关系，然后再训练学习器

包裹式：用训练出的效果作为特征选择的方式，直接用学习器结果评价属性子集，量身定做

嵌入式：加入L1正则化项，得到了稀疏的权重矩阵，完成了特征选择，学习器训练时自动特征选择

嵌入式L1和L2区别

可画（等值线）图说明。L1的解更容易出现在坐标轴上，L2在象限内

L1可以得到更稀疏的解，适合特征选择

λ 对稀疏的影响程度

- λ 很小，惩罚项校，过拟合
- λ 越大，解越稀疏，过拟合程度越低
- λ 超过一个阈值时，开始欠拟合，丢失特征

字典学习的目的

为普通稠密表达的样本找到合适的字典，才能转化成合适的稀疏表示，而不是过度稀疏或稠密，猜简化学习任务，降低模型复杂度

为什么要学稀疏表示

- 使大多数问题变得线性可分
- 已有高效存储的办法

压缩感知的矩阵补全

十二、

➤ PAC学习 **PAC是什么? 要搞清概念.**

- PAC学习理论、PAC辨识、PAC可学习、PAC学习算法、样本复杂度
- 可分情形、不可分情形

➤ VC维 **要评价什么? 为什么引入它们?**

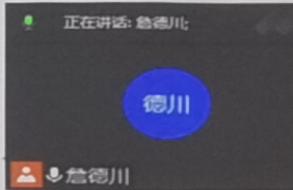
- 为什么需要VC维, 增长函数, 对分, 打散
- VC维计算 **考概念, 区分它们.**

➤ Rademacher复杂度

- Rademacher复杂度作用, 与VC维相比有什么区别
- 相关的定理

➤ 稳定性

- 稳定性评价什么, 与VC维、Rademacher复杂度相比有什么区别
- 相关的定理



十三、

三个半监督区分

主动学习: 选取尽量少的样本去找专家标记, 获得最大的模型提升

纯半监督: 未标记样本不是待测样本

直推学习: 未标记样本是待测样本, 无泛化能力

五个方法总结 (假设, 利用半监督样本, 优缺点, 场景)

1、半监督高斯混合模型:

- 假设样本由高斯混模型产生
- EM算法迭代更新参数
- 优点: 方法简单易于实现, 标记极少时比其他方法效果好
- 缺点: 模型假设必须准确, 否则反噬, 效果更差, 但很难假设准确
- 场景: 有可靠领域知识, 能做出准确假设

2、半监督SVM

- 低密度假设
- 伪标记, 再逐步更新
- 缺点: 开销巨大。可能有多个低密度分割线。目标函数非凸

3、图半监督

- 聚类假设：假设图标签平滑，强边相连的类别大概率相同
- 标记扩散，学习目标看作图的最小割
- 优点：可以用矩阵运算
- 缺点：存储开销大，难以处理大规模数据。泛化性差，难以加入新样本

4、协同学习

- 假设数据有不同充分且条件独立的视图（充分：每个视图都包含足以产生最优学习器的信息；条件独立：在给定类别标记线下视图互相独立）
- 弱假设：仅需弱学习器之间有显著分歧
- 预测自己最有把握的未标记样本，相互提供伪标记，共同进步

5、半监督聚类

- 场景：有少量标记或必连勿连信息
- 直接作为初始化的聚类中心
- x_i 的划分是否违反了已知信息，违反则划进最近的

十四、

什么是/哪些是生成，判别

- 生成式：计算联合分布 $P(Y, R, O)$
- 判别式：计算条件分布 $P(Y, R|O)$

□ 符号约定

- Y 为关心的变量的集合， O 为可观测变量集合， R 为其他变量集合

- 判别式：对条件分布建模，不考虑联合分布，**直接对后验概率建模**

判别式模型：**线性回归模型、线性判别分析、支持向量机SVM、神经网络等**，条件随机场面对预测往往学习准确度更高。

对条件概率建模，学习不同类别之间的最优边界。

捕捉不同类别特征的差异信息，不学习本身分布信息，无法反应数据本身特性。

学习成本较低，需要的计算资源较少。

需要的样本数可以较少，少样本也能很好学习。

无法转换成生成式。

- 生成式：**通过贝叶斯定理使问题转化为求联合分布**，再转化过去（但两次计算可能有两次误差）。更全面，拎出一部分都有用

反映同类数据本身的相似度，它不关心到底划分不同类的边界在哪里。

学习收敛速度更快，当样本容量增加时，学习到的模型可以更快的收敛到真实模型

当存在隐变量时，依旧可以用生成式模型，此时判别式方法就不行了。

对联合概率建模，学习所有分类数据的分布。

学习到的数据本身信息更多，能反应数据本身特性。

学习成本较高，需要更多的计算资源。

需要的样本数更多，样本较少时学习效果较差。

推断时性能较差。

一定条件下能转换成判别式。

生成式构建出模型后，可输入噪声产生新样本

生成式模型：朴素贝叶斯，隐马尔可夫模型，马尔可夫随机场

概率图模型是什么

是一类用图来表达变量相关关系的概率模型

为什么用概率图模型

联合概率分布用链式法则表达复杂，**概率图模型**引入条件独立性进行了转化

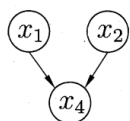
直接利用概率求和规则消去变量 R 的时间和空间复杂度为指数级别 $O(2^{|Y|+|R|})$ ，需要一种能够简洁紧凑**表达变量间关系**的工具

怎么把有向图道德化成无向图，会画盘式记法

- 找所有V型结构，添加一条横边

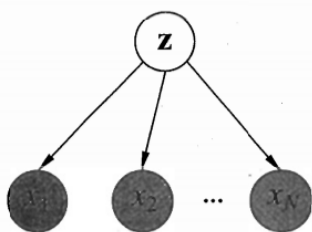
基于道德图能直观、迅速地找到变量间的条件独立性：

变量集合 z 去除后， x 和 y 分属两个连通分支，则称变量 x 和 y 被 z 有向分离

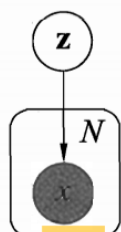


V型结构

- 盘内表示的是相互独立，由相同机制生成的多个变量



(a) 普通变量关系图

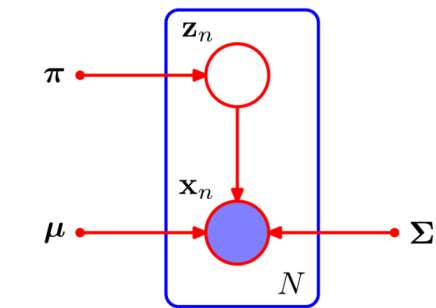


(b) 盘式记法

图 14.10 盘式记法的例示

$$p(\boldsymbol{x}) = \sum_{k=1}^K \pi_k p(\boldsymbol{x}|k)$$

答案：



N个数据点的高斯混合模型图表示

1. (5 points) 请使用盘式记法表示联合分布 $p(\mathbf{t}, \mathbf{w}, \alpha)$ 。

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t} \mid \mathbf{w})p(\mathbf{w} \mid \alpha)p(\alpha)$$

$$p(\mathbf{t} \mid \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^T \boldsymbol{\phi}_n, \beta^{-1})$$

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha^{-1} \mathbf{I})$$

$$p(\alpha) = \text{Gam}(\alpha \mid a_0, b_0) = \frac{b_0^{a_0} \alpha^{a_0-1} e^{-b_0 \alpha}}{\Gamma(a_0)}$$

什么是马尔科夫链

图 14.1 中的箭头表示了变量间的依赖关系。在任一时刻，观测变量的取值仅依赖于状态变量，即 x_t 由 y_t 确定，与其他状态变量及观测变量的取值无关。同时， t 时刻的状态 y_t 仅依赖于 $t - 1$ 时刻的状态 y_{t-1} ，与其余 $n - 2$ 个状态无关。这就是所谓的“马尔可夫链” (Markov chain)，即：系统下一时刻的状态仅由当前状态决定，不依赖于以往的任何状态。基于这种依赖关系，所有变量的联合概率分布为

$$P(x_1, y_1, \dots, x_n, y_n) = P(y_1)P(x_1 \mid y_1) \prod_{i=2}^n P(y_i \mid y_{i-1})P(x_i \mid y_i) . \tag{14.1}$$

马尔可夫随机场这些概念干什么的，有什么用

势函数，定义在变量子集上，用于定义概率分布函数

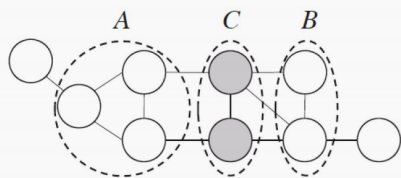
极大团

- 团：全连接子图
 - 极大图：无法再添加点以保持全连接性
- 不是极大团就被包含，（无法被其他团包含，可以分解成全部都是极大团）

全局、局部、部分马尔可夫性

• 分离

□ 借助“**分离**”的概念，若从结点集A中的结点到B中的结点都必须经过结点集C中的结点，则称结点集A, B被结点集C分离，称C为分离集 (separating set)



□ **全局马尔可夫性 (global Markov property)** : 在给定**分离集**的条件下，两个变量子集条件独立

- 若令A,B,C对应的变量集分别为 x_A, x_B, x_C ，则 x_A 和 x_B 在 x_C 给定的条件下独立，记为 $x_A \perp x_B \mid x_C$

□ 由全局马尔可夫性可以导出：

- **局部马尔可夫性 (local Markov property)** : 在给定**邻接变量**的情况下，一个变量条件独立于其它所有变量
 - 令 V 为图的结点集， $n(v)$ 为结点 v 在图上的邻接节点， $n^*(v) = n(v) \cup \{v\}$ ，有 $x_v \perp x_{V \setminus n^*(v)} \mid x_{n(v)}$
- **成对马尔可夫性 (pairwise Markov property)** : 在给定**所有其它变量**的情况下，两个非邻接变量条件独立
 - 令 V 为图的结点集，边集为 E ，对图中的两个结点 u, v ，若 $\langle u, v \rangle \notin E$ ，有 $x_u \perp x_v \mid x_{V \setminus \{u, v\}}$

HMM&MRF&CRF

MRF&CRF均使用团上的势函数定义概率

HMM在任何时刻观察值仅仅与状态（即要标注的标签）有关，HMM有了条件分布后，就变成了线性条件随机场

CRF：判别式，处理的是条件概率

概率图模型学习最常用的是极大似然，有隐变量时要用EM，EM过程

进一步，若我们不是取 \mathbf{Z} 的期望，而是基于 Θ^t 计算隐变量 \mathbf{Z} 的概率分布 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$ ，则EM算法的两个步骤是：

- **E步 (Expectation)**: 以当前参数 Θ^t 推断隐变量分布 $P(\mathbf{Z} \mid \mathbf{X}, \Theta^t)$ ，并计算对数似然 $LL(\Theta \mid \mathbf{X}, \mathbf{Z})$ 关于 \mathbf{Z} 的期望

$$Q(\Theta \mid \Theta^t) = \mathbb{E}_{\mathbf{Z} \mid \mathbf{X}, \Theta^t} LL(\Theta \mid \mathbf{X}, \mathbf{Z}) . \quad (7.36)$$

- **M步 (Maximization)**: 寻找参数最大化期望似然，即

$$\Theta^{t+1} = \arg \max_{\Theta} Q(\Theta \mid \Theta^t) . \quad (7.37)$$

简要说，EM算法使用两个步骤交替计算：第一步是期望(E)步，利用当前估计的参数值来计算对数似然的期望值；第二步是最大化(M)步，寻找能使E步产生的似然期望最大化的参数值。然后，新得到的参数值重新被用于E步，……直至收敛到局部最优解。

缺点，引出了近似推断

计算出目标变量的边际分布或条件分布的精确值，计算复杂度随极大团规模增长呈指数增长

吉布斯采样算法的收敛速度较慢.此外, 若贝叶斯网中存在极端概率"0"或"1", 则不能保证马尔可夫链存在平稳分布, 此时吉布斯采样会给出错误的估计结果.

直接计算或逼近, 比推断概率分布更容易

MCMC随机采样: 通过构造一条马尔可夫链, 使其收敛至平稳分布恰为待估计参数的后验分布, 然后通过该马尔可夫链产生样本, 用这些样本进行估计

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

变分推断用在EM的哪一部分, 做了什么, 了解一下怎么做, 和EM有什么关系

□ 可使用EM算法最大化对数似然

- E步: 根据 t 时刻的参数 θ^t 对 $p(z | x, \theta^t)$ 进行推断, 并计算联合似然函数 $p(x, z | \theta^t)$
- M步: 基于E步结果进行最大化寻优, 对关于变量 θ 的函数 $Q(\theta; \theta^t)$ 进行最大化从而求取:

$$\begin{aligned}\theta^{t+1} &= \arg\max_{\theta} Q(\theta; \theta^t) \\ &= \arg\max_{\theta} \sum_z p(z | x, \theta^t) \ln p(x, z | \theta)\end{aligned}$$

对数联合似然函数 $\ln p(x, z | \theta)$ 在分布 $p(z | x, \theta^t)$ 下的期望, 分布 $p(z | x, \theta^t)$ 与 z 的真实后验分布相等, $Q(\theta; \theta^t)$ 近似于对数似然函数

□ $p(z | x, \theta^t)$ 是隐变量 z 的近似分布, 将这个近似分布用 $q(z)$ 表示:

对数似然

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q \| p)$$

构成下界

分布 q 和 p 的差异度量

$$\mathcal{L}(q) = \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} \right\} d\mathbf{z}$$

$$\text{KL}(q \| p) = - \int q(\mathbf{z}) \ln \left\{ \frac{p(\mathbf{z})|p(\mathbf{x})}{q(\mathbf{z})} \right\} d\mathbf{z}$$

假设 z 的分布:

$$q(\mathbf{z}) = \prod_{i=1}^M q_i(\mathbf{z}_i)$$

- 变量子集 z_j 所服从的最优分布 q_j^* 应满足：

$$q_j^*(\mathbf{z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]) d\mathbf{z}_j}$$

- 因此，通过恰当分割变量子集 z_j 并选择 q_i 服从的分布， $\mathbb{E}_{i \neq j}[\ln p(\mathbf{x}, \mathbf{z})]$ 往往有闭式解，使得上式能对隐变量高效推断
- 由于在对 z_j 所服从的分布 q_j^* 估计时融合了 z_j 之外的其它 $z_{i \neq j}$ 的信息，这是通过联合似然函数 $\ln p(\mathbf{x}, \mathbf{z})$ 在 z_j 之外的隐变量分布上求期望得到的，因此亦称为“平均场”（mean field）方法
- 在实际应用中，最重要的是考虑如何对隐变量进行拆解，以及假设各变量子集服从何种分布，在此基础上结合EM算法对概率图模型进行推断和参数估计

话题模型稍微了解一下

- 话题模型（topic model）是一类生成式有向图模型，主要用来处理离散型的数据集合（如文本集合）。作为一种非监督产生式模型，话题模型能够有效利用海量数据发现文档集合中隐含的语义。隐狄里克雷分配模型（Latent Dirichlet Allocation, **LDA**）是话题模型的典型代表。

- 强化学习
 - 基本概念、四元组。
- 多摇臂赌博机
 - 探索和利用、解决方法
- 有模型强化学习
 - 状态值函数Q和状态-动作值函数V
 - Bellman等式，最优Bellman等式（关于Q和V的）
 - 策略迭代和值迭代
- 免模型强化学习
 - 蒙特卡洛学习，时序差分学习
 - on-policy, off-policy
- 值函数近似
- 模仿学习
- 逆强化学习

类别不平衡问题

EM&Kmeans

PCA推导，降维理解

马氏距离

特征选择方法比较，relief，LVM计算

稀疏最优解，L1正则化

图半监督学习的模型约束，公式，求最优解，步骤，实际含义

PAC定义，12章

概率图模型

生成式&判别式

HMM，表达式。与CRF区别

MCMC

Sarsa, Q-learning

rl-Bellman

规则学习，自顶向下

VC维

最近邻分类器

TSVM算法

高斯混合模型，盘式记法，EM算法

半监督，伪标签，与主动学习区别