東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

人工智能概论

第七章：贝叶斯分类器

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn

目录

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 贝叶斯网络
- EM 算法

贝叶斯决策论

基本概念

- 贝叶斯决策论（Bayesian Decision Theory）
- 一个概率框架下实施决策的基本方法
- 对分类任务来说，在所有相关概率都已知的理想情形下
- 贝叶斯决策论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

贝叶斯决策论

基本概念

- 假设有 N 种可能的类别标记，即 $y = \{c_1, c_2, \ldots, c_N\}$
- λ_{ij} 是将一个真实标记为 c_j 的样本误分类为 c_i 所产生的损失
- 基于后验概率 $P(c_i|x)$
- 可获得将样本 x 分类为 c_i 所产生的期望损失（Expected Loss）
- 即在样本 x 上的“条件风险（Conditional Risk）”

$$R(c_i \mid \boldsymbol{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \boldsymbol{x}) \ .$$

贝叶斯决策论

贝叶斯判定准则

- 我们的任务是寻找一个判定准则 h 以最小化总体风险：

$$R(h) = \mathbb{E}_{\boldsymbol{x}}\left[R\left(h\left(\boldsymbol{x}\right) \mid \boldsymbol{x}\right)\right] .$$

- 对每个样本 x，若 h 能最小化条件风险 $R(h(x)|x)$
- 则总体风险 $R(h)$ 也将被最小化

贝叶斯决策论

贝叶斯判定准则

- 贝叶斯判定准则（Bayes Decision Rule）
- 为最小化总体风险，只需在每个样本上选择那个能使条件风险 $R(h(x)|x)$ 最小的类别标记，即：

$$h^*(\boldsymbol{x}) = \arg\min_{c \in \mathcal{Y}} R(c \mid \boldsymbol{x}) \ ,$$

- h^* 被称为贝叶斯最优分类器（Bayes Optimal Classifier）
- 与之对应的总体风险 $R(h^*)$ 被称为贝叶斯风险（Bayes Risk）

贝叶斯决策论

贝叶斯判定准则

- 若目标是最小化分类错误率，则误判损失 λ_{ij} 可写为：

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j \; ; \\ 1, & \text{otherwise}, \end{cases}$$

- 此时的条件风险：

$$R(c \mid \boldsymbol{x}) = 1 - P(c \mid \boldsymbol{x}) \; ,$$

贝叶斯决策论

贝叶斯判定准则

- 于是，最小化分类错误率的贝叶斯最优分类器为：

$$h^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c \mid \boldsymbol{x})$$

- 即对每个样本 x
- 选择能使后验概率 $P(c|x)$ 最大的类别标记

贝叶斯决策论

贝叶斯判定准则

- 使用贝叶斯判定准则来最小化决策风险
- 首先要获得后验概率 $P(c|x)$
- 然而，在现实任务中这通常难以直接获得
- 因此，机器学习所要实现的是：
- 基于有限的训练样本集尽可能准确地估计出后验概率 $P(c|x)$

贝叶斯决策论

估计方法

- 对生成式模型来说，必然考虑：

$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

- 基于贝叶斯定理，$P(c|x)$ 可写为：

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

贝叶斯决策论

估计方法

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

- 其中，$P(c)$ 是类的“先验（Prior）概率”
- $P(x|c)$ 是样本 x 相对于类 c 的“条件（Conditional）概率”
- 条件概率也称为“似然（Likelihood）”
- $P(x)$ 是用于归一化的“证据（Evidence）因子”

贝叶斯决策论

估计方法

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

- 对给定样本 x，证据因子 $P(x)$ 与类的标签无关
- 因此估计 $P(c|x)$ 的问题就转化为：
- 如何基于训练数据 D 来估计先验概率 $P(c)$ 和似然 $P(x|c)$

贝叶斯决策论

估计方法

- 先验概率 $P(c)$ 表达了样本空间中各类样本所占的比例
- 根据大数定律，当训练集包含充足的独立同分布样本时
- $P(c)$ 可通过各类样本出现的频率来进行估计

贝叶斯决策论

估计方法

- 似然 $P(x|c)$ 涉及关于 x 所有特征的联合概率
- 直接根据样本出现的频率来估计将会遇到严重的困难
- 例如，假设样本的 d 个特征性都是二值的
- 则样本空间将有 2^d 种可能的取值

贝叶斯决策论

估计方法

- 在现实应用中，这个值往往远大于训练样本数量
- 也就是说，很多样本取值在训练集中根本没有出现
- 直接使用频率来估计 $P(x|c)$ 显然不可行
- 因为“未被观测到”与“出现概率为零”通常是不同的

贝叶斯决策论

极大似然估计

- 估计类条件概率的一种常用策略是：
- 先假定其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
- 具体来说，假设 $P(x|c)$ 具有确定的形式并且被参数向量 θ_c 唯一确定
- 则我们的任务就是利用训练集 D 估计参数 θ_c
- 为明确起见，我们将 $P(x|c)$ 记为 $P(x|\theta_c)$

贝叶斯决策论

极大似然估计

- 极大似然估计（Maximum Likelihood Estimation，MLE）
- 根据数据采样来估计概率分布参数的经典方法
- 令 D_c 表示训练集 D 中第 c 类样本组成的集合
- 假设这些样本是独立同分布的
- 则参数 θ_c 对于数据集 D_c 的似然是：

$$P(D_c \mid \boldsymbol{\theta_c}) = \prod_{\mathbf{x} \in D_c} P(\mathbf{x} \mid \boldsymbol{\theta_c})$$

贝叶斯决策论

极大似然估计

- 对 θ_c 进行极大似然估计
- 就是去寻找能最大化似然 $P(D_c|\theta_c)$ 的参数值 $\hat{\theta}_c$
- 直观上看，极大似然估计是试图在 θ_c 所有可能的取值中
- 找到一个能使数据出现的“可能性”最大的值

贝叶斯决策论

极大似然估计

- 这种参数化的方法虽能使类条件概率估计变得相对简单
- 但估计结果的准确性严重依赖于：
- 所假设的概率分布形式是否符合潜在的真实数据分布
- 在现实应用中，想要做出能较好地接近潜在真实分布的假设
- 往往需在一定程度上利用关于应用任务本身的经验知识
- 否则若仅凭“猜测”来假设概率分布形式
- 很可能产生误导性的结果

目录

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 贝叶斯网络
- EM 算法

朴素贝叶斯分类器

基本概念

- 似然 $P(x|c)$ 涉及关于 x 所有特征的联合概率
- 在现实应用中，很多样本取值在训练集中根本没有出现
- 直接使用频率来估计 $P(x|c)$ 显然不可行
- 因为“未被观测到”与“出现概率为零”通常是不同的
- 朴素贝叶斯分类器是解决这个问题的一个方法

朴素贝叶斯分类器

基本概念

- 朴素贝叶斯分类器（Naïve Bayes Classifier）
- 采用了特征条件独立性假设：
- 对已知类别假设所有特征相互独立
- 换言之，假设每个特征独立地对分类结果发生影响

朴素贝叶斯分类器

基本概念

- 基于属性条件独立性假设：

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})}$$

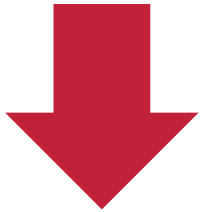
$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^{d} P(x_i \mid c)$$

朴素贝叶斯分类器

基本概念

- 由于对所有类别来说 $P(x)$ 相同，朴素贝叶斯分类器的表达式为：

$$h^*(x) = \underset{c \in y}{\operatorname{argmin}} R(c \mid x)$$



$$h_{nb}(\mathbf{x}) = \underset{c \in y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i \mid c)$$

朴素贝叶斯分类器

训练过程

- 朴素贝叶斯分类器的训练过程就是：

1. 基于训练集 D 来估计类先验概率 $P(c)$
2. 为每个特征估计条件概率 $P(x_i|c)$

朴素贝叶斯分类器

训练过程

- 基于训练集 D 来估计类先验概率 $P(c)$
- 令 D_c 表示训练集 D 中第 c 类样本组成的集合
- 若有充足的独立同分布样本，则可容易地估计出类先验概率：

$$P(c) = \frac{|D_c|}{D}$$

朴素贝叶斯分类器

训练过程

- 为每个特征估计条件概率 $P(x_i|c)$：离散特征
- 令 D_{c,x_i} 表示 D_c 中在第 i 个属性上取值为 x_i 的样本组成的集合
- 则条件概率 $P(x_i|c)$ 可估计为：

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|}$$

朴素贝叶斯分类器

训练过程

- 为每个特征估计条件概率 $P(x_i|c)$：连续特征
- 可考虑概率密度函数，假定 $P(x_i|c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)$
- 其中 $\mu_{c,i}$ 和 $\sigma_{c,i}^2$ 分别是第 c 类样本在第 i 个属性上取值的均值和方差
- 则有：

$$P(x_i \mid c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2{\sigma_{c,i}}^2}\right)$$

朴素贝叶斯分类器

案例分析

表 4.3　西瓜数据集 3.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

- 色泽：青绿
- 根蒂：蜷缩
- 敲声：浊响
- 纹理：清晰
- 脐部：凹陷
- 触感：硬滑
- 密度：0.697
- 含糖率：0.460

朴素贝叶斯分类器

案例分析

- 首先估计类先验概率 $P(c)$，显然有：

$$P(\text{好瓜}=\text{是}) = \frac{8}{17} \approx 0.471 ,$$

$$P(\text{好瓜}=\text{否}) = \frac{9}{17} \approx 0.529 .$$

朴素贝叶斯分类器

案例分析

❖ 然后，为每个特征估计条件概率 $P(x_i|c)$：

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375 ,$$

$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 ,$$

$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{是}) = \frac{5}{8} = 0.375 ,$$

$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂} = \text{蜷缩} \mid \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333 ,$$

$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750 ,$$

$$P_{\text{浊响}|\text{否}} = P(\text{敲声} = \text{浊响} \mid \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444 ,$$

$$P_{\text{清晰}|\text{是}} = P(\text{纹理} = \text{清晰} \mid \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875 ,$$

$$p_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是})$$
$$= \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959 ,$$

$$p_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否})$$
$$= \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203 ,$$

$$p_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是})$$
$$= \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788 ,$$

$$p_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否})$$
$$= \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066 .$$

朴素贝叶斯分类器

案例分析

$$P(\text{好瓜}=\text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}}$$
$$\times P_{\text{硬滑}|\text{是}} \times p_{\text{密度: 0.697}|\text{是}} \times p_{\text{含糖: 0.460}|\text{是}} \approx 0.038 ,$$
$$P(\text{好瓜}=\text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}}$$
$$\times P_{\text{硬滑}|\text{否}} \times p_{\text{密度: 0.697}|\text{否}} \times p_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5} .$$

- 由于 0.038 > 6.80 x 10^{-5}
- 朴素贝叶斯分类器将测试样本判别为“好瓜”

朴素贝叶斯分类器

拉普拉斯修正

- 需注意，若某个特征值在训练集中没有与某个类同时出现过
- 则其条件概率可能为 0，并影响后续概率判别
- 例如在上述数据里：

$$P_{\text{清脆}|\text{是}} = P(\text{敲声} = \text{清脆} \mid \text{好瓜} = \text{是}) = \frac{0}{8} = 0$$

- 无论该样本的其他特征是什么，哪怕在其他特征上明显像好瓜
- 分类的结果都将是“好瓜：否”，这显然不太合理

朴素贝叶斯分类器

拉普拉斯修正

- 为了避免其他特征携带的信息被训练集中未出现的特征值“抹去”
- 在估计概率值时通常要进行“平滑（Smoothing）”
- 常用“拉普拉斯修正（Laplacian Correction）”

朴素贝叶斯分类器

拉普拉斯修正

- 拉普拉斯修正（Laplacian Correction）
- 令 N 表示训练集 D 中可能的类别数，N_i 表示第 i 个属性可能的取值数
- 先验概率和条件概率分别修正为：

$$P(c) = \frac{|D_c|}{D} \quad \Rightarrow \quad \hat{P}(c) = \frac{|D_c| + 1}{|D| + N}$$

$$P(x_i \mid c) = \frac{|D_{c,x_i}|}{D} \quad \Rightarrow \quad \hat{P}(x_i \mid c) = \frac{|D_{c,x_i}| + 1}{|D| + N_i}$$

朴素贝叶斯分类器

进阶使用方法

- 在现实任务中：
- 任务对预测速度要求高：“查表”
- 任务数据更替频繁：“懒惰学习（Lazy Learning）”
- 数据不断增加：“增量学习（Incremental Learning）”

目录

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 贝叶斯网络
- EM 算法

贝叶斯网络

结构

- 贝叶斯网络（Bayesian Network）
- 亦称“信念网络（Belief Network）”
- 它借助有向无环图（Directed Acyclic Graph，DAG）
- 来刻画特征之间的依赖关系
- 并使用条件概率表（Conditional Probability Table，CPT）
- 来描述特征的联合概率分布

贝叶斯网络

结构

- 一个贝叶斯网络 B 由结构 G 和参数 Θ 两部分构成
- 网络结构 G 是一个有向无环图，其每个结点对应于一个特征
- 若两个特征有直接依赖关系，则它们由一条边连接起来
- 参数 Θ 定量描述这种依赖关系
- 假设特征 x_i 在 G 中的父结点集为 π_i
- 则 Θ 包含了每个特征的条件概率表：

$$\theta_{x_i|\pi_i} = P_B(x_i|\pi_i)$$

贝叶斯网络

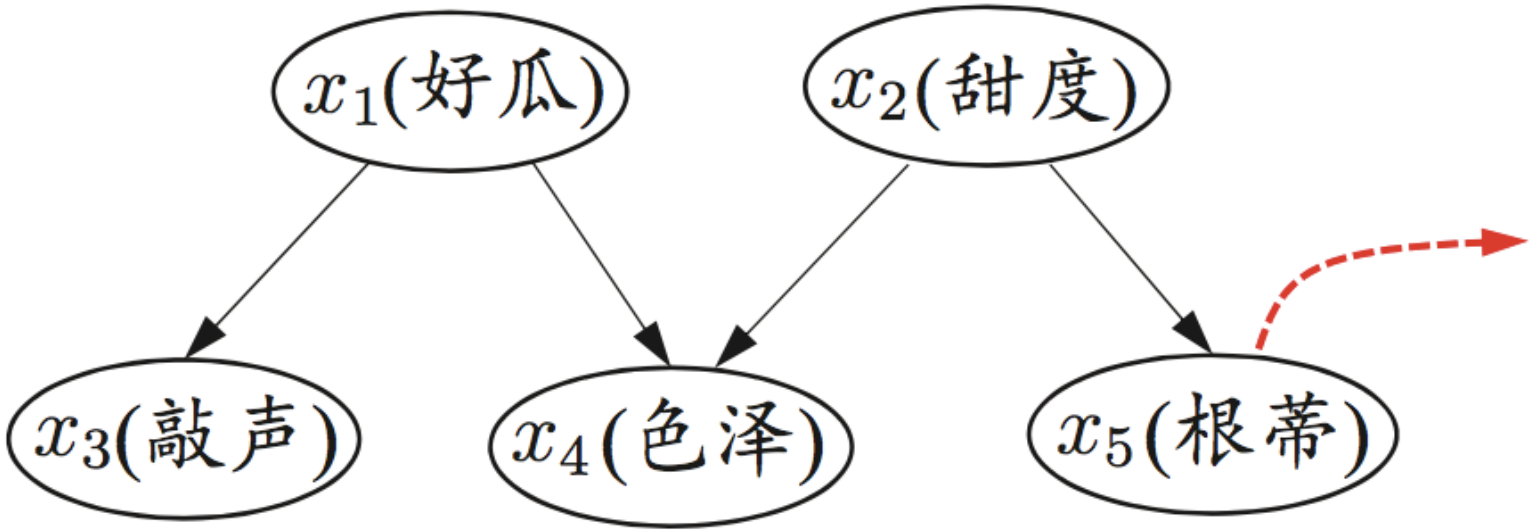
结构

- 贝叶斯网络结构有效地表达了特征间的条件独立性
- 给定父结点集，贝叶斯网络假设每个特征与它的非后裔特征独立
- 于是 $B = \langle G, \Theta \rangle$ 将特征的联合概率分布定义为：

$$P_B(x_1, x_2, \cdots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i}$$

贝叶斯网络

结构



根蒂

甜度	硬挺	蜷缩
高	0.1	0.9
低	0.7	0.3

贝叶斯网络

训练

- 若网络结构已知，即特征间的依赖关系已知
- 则贝叶斯网络的学习过程相对简单
- 只需通过对训练样本“计数”，估计出每个结点的条件概率表即可
- 但在现实应用中我们往往并不知晓网络结构
- 于是，贝叶斯网络学习的首要任务就是：
- 根据训练数据集来找出结构最“恰当”的贝叶斯网络

贝叶斯网络

训练

- 评分搜索
- 先定义一个评分函数（Score Function）
- 以此来评估贝叶斯网络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络
- 显然，评分函数引入了关于我们希望获得什么样的贝叶斯网的归纳偏好

贝叶斯网络

训练

- 最小描述长度（Minimal Description Length，MDL）
- 常用评分函数通常基于信息论准则
- 学习的目标是找到一个能以最短编码长度描述训练数据的模型
- 给定训练集 $D = \{x_1, x_2, \ldots, x_m\}$
- 贝叶斯网络 $B = \langle G, \Theta \rangle$ 在 D 上的评分函数可写为：

$$S(B \mid D) = f(\theta)|B| - LL(B \mid D)$$

贝叶斯网络

训练

- 搜索最优贝叶斯网络结构是一个 NP-Hard 问题
- 有两种常用的策略能在有限时间内求得近似解：

1. 贪心法：例如从某个网络结构出发，每次调整一条边，直到评分函数值不再降低为止
2. 约束法：给网络结构施加约束来削减搜索空间，例如将网络结构限定为树形结构等

贝叶斯网络

推断

- 贝叶斯网络训练好之后就能用来回答“查询（Query）”
- 即通过一些特征的观测值来推测其他特征的取值
- 例如在西瓜问题中：
- 若我们观测到西瓜色泽青绿、敲声浊响、根蒂蜷缩
- 就可以推断它是否成熟、甜度如何
- 通过已知变量观测值来推测待查询变量的过程称为“推断（Infer）”
- 已知变量观测值称为“证据（Evidence）”

贝叶斯网络

推断

- 最理想的推断方式是：
- 直接根据贝叶斯网络定义的联合概率分布来精确计算后验概率
- 但这样的“精确推断”已被证明是 NP-Hard 的
- 此时需借助“近似推断”，在有限时间内求得近似解
- 吉布斯采样（Gibbs Sampling）就是一种常见的随机采样方法

贝叶斯网络

推断

- 吉布斯采样是在贝叶斯网络所有变量的联合状态空间与证据一致的子空间中进行“随机漫步（Random Walk）”
- 每一步仅依赖于前一步的状态
- 这是一个“马尔可夫链（Markov Chain）”

目录

- 贝叶斯决策论
- 朴素贝叶斯分类器
- 贝叶斯网络
- EM 算法

EM 算法

基本概念

- 在前面的讨论中
- 我们一直假设训练样本所有属性变量的值都已被观测到
- 即训练样本是“完整”的
- 但在现实应用中往往会遇到“不完整”的训练样本
- 例如由于西瓜的根蒂已脱落，无法看出是“蜷缩”还是“硬挺”
- 则训练样本的“根蒂”特征变量值未知
- 在这种存在“未观测”变量的情形下
- 是否仍能对模型参数进行估计呢？

EM 算法

基本概念

- 未观测变量的学名是“隐变量（Latent Variable）”
- 令 X 表示已观测变量集，Z 表示隐变量集，Θ 表示模型参数
- 若欲对 Θ 做极大似然估计，则应最大化对数似然：

$$LL(\mathbf{\Theta} \mid \mathbf{X}, \mathbf{Z}) = \ln P(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta})$$

EM 算法

基本概念

- 然而由于 Z 是隐变量，上式无法直接求解
- 此时我们可通过对 Z 计算期望
- 来最大化已观测数据的对数“边际似然（Marginal Likelihood）”

$$LL(\boldsymbol{\Theta} \mid \mathbf{X}) = \ln P(\mathbf{X} \mid \boldsymbol{\Theta}) = \ln \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})$$

EM 算法

基本概念

- EM（Expectation-Maximization）算法
- 常用的估计参数隐变量的算法
- 它是一种迭代式的方法，其基本想法是：
- 若参数 Θ 已知，则可根据训练数据推断出最优隐变量 Z 的值（E 步）
- 若 Z 的值已知，则可方便地对参数 Θ 做极大似然估计（M 步）

EM 算法

基本概念

- 于是，以初始值 Θ^0 为起点
- 可迭代执行以下步骤直至收敛：

1. 基于 Θ^t 推断隐变量 Z 的期望，记为 Z^t
2. 基于已观测变量 X 和 Z^t 对参数 Θ 做极大似然估计，记为 Θ^{t+1}

- 这就是 EM 算法的原型

EM 算法

基本概念

- 简要来说， EM 算法使用两个步骤变替计算：
- 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

EM 算法

基本概念

- 事实上，隐变量估计问题也可通过梯度下降等优化算法求解
- 但由于求和的项数将随着隐变量的数目以指数级上升
- 会给梯度计算带来麻烦
- 而 EM 算法则可看作一种非梯度优化方法

总结

- 「西瓜书」周志华《机器学习》，清华大学出版社
- https://item.jd.com/12762673.html

- 「南瓜书」谢文睿、秦州《机器学习公式详解》，人民邮电出版社
- https://github.com/datawhalechina/pumpkin-book/

总结

- Scikit-Learn
- https://scikit-learn.org

- TensorFlow
- https://www.tensorflow.org

TensorFlow

東莞理工學院
DONGGUAN UNIVERSITY OF TECHNOLOGY

Thank You!

丁烨，计算机科学与技术学院
dingye@dgut.edu.cn