7.1 贝叶斯决策论

7.1.1 贝叶斯决策论的基本概念

贝叶斯决策论 (Bayesian decision theory) 是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**..

7.1.2 贝叶斯决策论的总体数学描述

假设有 N 种可能的类别标记, 即 $\mathcal{Y}=\{c_1,c_2,\ldots,c_N\}$, 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** $P\left(c_i|\boldsymbol{x}\right)$ 可获得样本 $\boldsymbol{x}$ 分类为 c_i 所产生的期望损失, 即在样本 $\boldsymbol{x}$ 上的 "条件风险".

$$R\left(c_i|\boldsymbol{x}\right)=\sum_{j=1}^{N}\lambda_{ij}P\left(c_j|\boldsymbol{x}\right) \tag{7.1}$$

我们的任务就是寻找一个判定准则 $h:\mathcal{X}\mapsto\mathcal{Y}$ 以最小化总体风险

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

显然, 对每个样本 $\boldsymbol{x}$, 若 h 能最小化条件风险 $R(h(\boldsymbol{x})|\boldsymbol{x})$, 则总体风险 $R(h)$ 也将被最小化.

于是, **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** $R(c|\boldsymbol{x})$ **最小的类别标记** , 即

$$h^*(\boldsymbol{x})=\arg\min_{c\in\mathcal{Y}}R(c|\boldsymbol{x}) \tag{7.3}$$

此时, h^* 称为**贝叶斯最优分类器**, 相对应的**总体风险** $R\left(h^*\right)$ 称为**贝叶斯风险**. $1-R\left(h^*\right)$ 反映了分类器所能达到的**最好性能**，即通过机器学习所能产生的**模型精度的理论上限**.

> 注1: **先验概率和后验概率**
>
> **先验概率**:
>
> **事件发生前的预判概率。**可以是基于历史数据的统计，可以由背景常识得出，也可以是人的主观观点给出。一般都是单独事件概率，如P(x),P(y)。
>
> **后验概率**:
>
> 事件发生后求的反向条件概率；或者说，基于先验概率求得的反向条件概率。概率形式与条件概率相同。
>
> **条件概率**:
>
> 一个事件发生后另一个事件发生的概率。一般的形式为P(x|y)表示y发生的条件下x发生的概率。
>
> 贝叶斯公式:

> $$P(y|x) = (P(x|y) * P(y))/P(x)$$
>
> 这里:
>
> $P(y|x)$ 是后验概率，一般是我们求解的目标。
>
> $P(x|y)$ 是条件概率，又叫似然概率，一般是通过历史数据统计得到。一般不把它叫做先验概率，但从定义上也符合先验定义。
>
> $P(y)$ 是先验概率，一般都是人主观给出的。贝叶斯中的先验概率一般特指它。
>
> $P(x)$ 其实也是先验概率，只是在贝叶斯的很多应用中不重要（因为只要最大后验不求绝对值），需要时往往用全概率公式计算得到。

7.1.3 贝叶斯决策论的具体数学表示

具体来说, 若目标是最小化分类错误率, 则**误判损失** λ_{ij} 可写成

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{otherwise} \end{cases} \tag{7.4}$$

此时条件风险

$$R(c|\boldsymbol{x}) = 1 - P(c|\boldsymbol{x}) \tag{7.5}$$

> 推导1:
>
> 由式7.1和式7.4可得： $R(c_i|\boldsymbol{x}) = 1 * P(c_1|\boldsymbol{x}) + 1 * P(c_2|\boldsymbol{x}) + \ldots + 0 * P(c_i|\boldsymbol{x}) + \ldots + 1 * P(c_N|\boldsymbol{x})$ 又 $\sum_{j=1}^{N} P(c_j|\boldsymbol{x}) = 1$，则： $R(c_i|\boldsymbol{x}) = 1 - P(c_i|\boldsymbol{x})$ 此即为式7.5

于是, 最小化分类错误类的贝叶斯最优分类器为

$$h^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c|\boldsymbol{x}) \tag{7.6}$$

即最优为: **对每个样本 $\boldsymbol{x}$, 选择能使后验概率 $P(c|\boldsymbol{x})$ 最大的类别标记.**

因此, 首先要获得后验概率 $P(c|\boldsymbol{x})$. 然而, 现实中很难直接获得. 从这个角度来看, 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**.

主要**两种**策略:

- 1 给定 $\boldsymbol{x}$, 可通过直接建模 $P(c|\boldsymbol{x})$ 来预测 c , 这样得到的是 "**判别式模型**".
- 2 也可先对联合概率分布 $P(\boldsymbol{x}, c)$ 建模, 然后再由此获得 $P(c|\boldsymbol{x})$, 这样得到的是 "**生成式模型**"

决策树, BP神经网络, 支持向量机等, 属于判别式模型的范畴.

对生成模型来说, 必须考虑

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})} \tag{7.7}$$

基于贝叶斯定理 (全概率公式), $P(c|\boldsymbol{x})$ 可写为:

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} \tag{7.8}$$

其中, $P(c)$ 是类**"先验"(prior) 概率**; $P(\boldsymbol{x}|c)$ 是样本 $\boldsymbol{x}$ 相对于类标记 c 的**类条件概率**, 或称为"**似然**"; $P(\boldsymbol{x})$ 是用于归一化的"**证据**"(evidence)**因子**.

对给定样本 $\boldsymbol{x}$，证据因子 $P(\boldsymbol{x})$ 与类标记无关, 因此估计 $P(c|\boldsymbol{x})$ 的问题就转化为如何基于训练数据 D 来估计**先验** $P(c)$ **和似然** $P(\boldsymbol{x}|c)$

- **类先验概率** $P(c)$: 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.
- **类条件概率** $P(\boldsymbol{x}|c)$: 由于它涉及关于 m 所有属性的联合概率，直 接根据样本出现的频率来估计将会遇到严重的困难. 因为"未被观测到"与"出现概率为零"通常是不同的.

7.2 极大似然估计

7.2.1 极大似然估计的概念

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**.具体地, 记关于类别 c 的类条件概率为 $P(\boldsymbol{x}|c)$, 假设 $P(\boldsymbol{x}|c)$ 具有**确定的形式**并且**被参数向量 $\boldsymbol{\theta}_c$ 唯一确定**, 则我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$. 为明确起见, 我们**将 $P(\boldsymbol{x}|c)$ 记为 $P(\boldsymbol{x}|\boldsymbol{\theta}_c)$.**

实际上, 概率模型的训练过程就是参数估计过程. 参数估计, 存在两个学派:

- **频率主义学派**: 认为参数虽然未知, 但却是客观存在的固定值, 因此, 可通过优化似然函数等准则来确定参数值
- **贝叶斯学派**: 认为参数是未观察到的随机变量, 其本身也可有分布, 因此, 可假定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.

本节介绍频率主义学派的**极大似然估计** (MLE) (Maximum Likelihood Estimation)

7.2.2 极大似然估计的数学表示

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 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是:

$$P(D_c|\boldsymbol{\theta}_c) = \prod_{\boldsymbol{x}\in D_c} P(\boldsymbol{x}|\boldsymbol{\theta}_c) \tag{7.9}$$

对 $\boldsymbol{\theta}_c$ 进行极大似然估计, 就是去寻找能最大化似然函数 $P(D_c|\boldsymbol{\theta}_c)$ 的参数 $\hat{\boldsymbol{\theta}}_c$.

> **注2**: 关于**似然函数的概念**: <概率论与数理统计教程>(第二版 茆诗松) p314 定义6.3.1 (似然函数)
>
> 似然函数的定义:

> 设总体的概率函数为 $p(x;\theta), \theta \in \Theta$, 其中 θ 是一个未知参数或几个未知参数组成的参数向量, Θ 是参数空间, $x_1, \cdots, x_n$ 是来自该总体的样本, 将样本的联合概率函数看成 θ 的函数, 用 $L\left(\theta; x_1, \cdots, x_n\right)$ 表示, 简记为 $L(\theta)$,
>
> $$L(\theta) = L\left(\theta; x_1, \cdots, x_n\right) = p\left(x_1; \theta\right) p\left(x_2; \theta\right) \cdots p\left(x_n; \theta\right)$$
>
> $L(\theta)$ 称为样本的似然函数. 如果某统计量 $\hat{\theta} = \hat{\theta}\left(x_1, \cdots, x_n\right)$ 满足
>
> $$L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$$
>
> 则称 $\hat{\theta}$ 是 θ 的最大似然估计, 简称 MLE.

为了便于分析和计算, 通常使用对数似然 (log-likelihood):

$$\begin{aligned} LL\left(\boldsymbol{\theta}_c\right) &= \log P\left(D_c | \boldsymbol{\theta}_c\right) \\ &= \sum_{\boldsymbol{x} \in D_c} \log P\left(\boldsymbol{x} | \boldsymbol{\theta}_c\right) \end{aligned} \tag{7.10}$$

> 注3: 这里的 log, 我觉得应该是 ln

此时参数 θ 的极大似然估计 $\hat{\theta}$ 为:

$$\hat{\boldsymbol{\theta}}_c = \arg\max_{\boldsymbol{\theta}_c} LL\left(\boldsymbol{\theta}_c\right) \tag{7.11}$$

对于正态分布, 概率密度函数 $p(\boldsymbol{x}|c) \sim \mathcal{N}\left(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2\right)$, 参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} \boldsymbol{x} \tag{7.12}$$

$$\hat{\boldsymbol{\sigma}}_c^2 = \frac{1}{|D_c|} \sum_{\boldsymbol{x} \in D_c} \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c\right) \left(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_c\right)^{\mathrm{T}} \tag{7.13}$$

需要注意的一点是, 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.

7.3 朴素贝叶斯分类器

7.3.1 朴素贝叶斯分类器的数学表示

基于贝叶斯公式 (7.8) 来估计后验概率 $P(c|\boldsymbol{x})$ 的主要用难在于: 类条件概率 $P(\boldsymbol{x}|c)$ 是所有属性上的联合概率，难以从有限的训练样本直接估计而得.

为避开这个, **朴素贝叶斯分类器**采用了"**属性条件独立性假设**": 对已知类别, **假设所有属性相互独立**. 也就是, 假设**每个属性独立地对分类结果发生影响**.

基于**属性条件独立性**假设，式 (7.8) 可重写为:

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d} P\left(x_i|c\right) \tag{7.14}$$

其中, d 为属性数目, x_i 为 $\boldsymbol{x}$ 在第 i 个属性上的取值.

由于对所有类别来说 $P(\boldsymbol{x})$ 相同, 因此基于式 (7.6) 的贝叶斯判定准则可写为

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P\left(x_i|c\right) \tag{7.15}$$

这就是朴素贝叶斯分类器的表达式.

显然, 朴素贝叶斯分类器的训练过程就是基于训练集 D 来**估计类先验概率** $P(c)$, 并为每个属性**估计条件概率** $P(x_i|c)$

7.3.2 朴素贝叶斯分类器具体计算方法

1 类先验概率 $P(c)$ 的计算

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, 则易计算出类先验概率

$$P(c) = \frac{|D_c|}{|D|} \tag{7.16}$$

2 属性的条件概率 $P\left(x_i|c\right)$

对**离散属性**而言, 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:

$$P\left(x_i|c\right) = \frac{|D_{c,x_i}|}{|D_c|} \tag{7.17}$$

对连续属性而言, 可考虑概率密度函数, 假定 $p\left(x_i|c\right) \sim \mathcal{N}\left(\mu_{c,i}, \sigma_{c,i}^2\right)$, 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

$$p\left(x_i|c\right) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}}\exp\left(-\frac{\left(x_i-\mu_{c,i}\right)^2}{2\sigma_{c,i}^2}\right) \tag{7.18}$$

7.3.3 朴素贝叶斯分类器计算方法的改进

若**某个属性值**在训练集中**没有与某个类同时出现**过，则直接基于式 (7.17) 进行概率估计, 再根据式 (7.15) 进行判别将出现问题.

为了避免其他属性携带的信息被训练集中未出现的属性值"抹去', 在估计概率值时通常要进行"平滑", 常用**"拉普拉斯修正"**. 具体来说, 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, 则式 (7.16) 和 (7.17) 分别修正为:

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \tag{7.19}$$

$$\hat{P}\left(x_i|c\right) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i} \tag{7.20}$$

优点: 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, 并且在训练集变大时, 修正过程所引入的先验 (prior) 的影响也会逐渐变得可忽略, **使得估值渐趋向于实际概率值**.

7.4 半朴素贝叶斯分类器

暂放

7.5 贝叶斯网

暂放

7.6 EM算法

7.6.1 EM算法问题的提出

在前面的讨论中, 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. 这样会造成某些属性变量值未知. 在这种存在"未观测"变量的情形下, 如何对模型参数进行估计.

7.6.1 EM 算法的数学表达

未观测变量的学名叫"隐变量" (latent variable). 令 $\mathbf{X}$ 表示已观测变量集, $\mathbf{Z}$ 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

$$LL(\Theta|\mathbf{X},\mathbf{Z}) = \ln P(\mathbf{X},\mathbf{Z}|\Theta) \tag{7.34}$$

由于 $\mathbf{Z}$ 是隐变量, 上式无法直接求解. 我们可以通过对 $\mathbf{Z}$ 计算期望, 来最大化已观测数据的对数"边际似然"

$$LL(\Theta|\mathbf{X}) = \ln P(\mathbf{X}|\Theta) = \ln \sum_{\mathbf{Z}} P(\mathbf{X},\mathbf{Z}|\Theta) \tag{7.35}$$

EM 算法是常用的估计参数隐变量的方法. 其基本思想为: 若参数 Θ 已知, 则可根据训练数据推断出最优隐变量 $\mathbf{Z}$ 的值 (E 步); 反之, 若 $\mathbf{Z}$ 的值已知, 则可方便地对参数 Θ 做极大似然估计 (M 步).

于是，以初始值 Θ^0 为起点, 对式 (7.35) ，可迭代执行以下步骤直至收敛:

- 基于 Θ^t 推断隐变量 $\mathbf{Z}$ 的期望, 记为 $\mathbf{Z}^t$;
- 基于已观测变量 $\mathbf{X}$ 和 $\mathbf{Z}^t$ 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.

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

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

$$Q\left(\Theta|\Theta^t\right)=\mathbb{E}_{\mathbf{Z}|\mathbf{X},\Theta^t}LL(\Theta|\mathbf{X},\mathbf{Z}) \tag{7.36}$$

- $\mathbf{M}$ 步 (Maximization): 寻找参数最大化期望似然, 即

$$\Theta^{t+1}=\arg\max_{\Theta}Q\left(\Theta|\Theta^t\right) \tag{7.57}$$

总结来说:

$\mathbf{EM}$ 算法使用两个步骤交替计算:

- 第一步是期望 ($\mathbf{E}$)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 ($\mathbf{M}$)步, 寻找能使 $\mathbf{E}$ 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 $\mathbf{E}$ 步, ... 直至收敛到局部最优解.

> 注4: 关于 $\mathbf{EM}$ 算法的整个过程有一些理解不透彻, $\mathbf{EM}$ 算法的详细数学推导和计算过程在<统计学习方法>中会进一步学习.