贝叶斯分类器的训练

翟婷婷

扬州大学
信息工程（人工智能）学院
zhtt@yzu.edu.cn

2023年春

课程目标

- 掌握利用训练数据集训练得到贝叶斯分类器的方法。
- 了解参数化估计方法和非参数估计方法的区别。
- 重点掌握最大似然估计法和贝叶斯估计法。

引言

- 贝叶斯决策论告诉我们，知道先验概率$P(\omega_j)$和类条件概率密度$p(\boldsymbol{x}|\omega_j)$情况下，如何设计一个最优的分类器。
- 在实际的模式分类问题中，完全的概率结构信息很难获得，即$P(\omega_j)$和$p(\boldsymbol{x}|\omega_j)$不能预先知道。
- 实际问题中，我们能够搜集一个待分类模式的特定子集，即训练数据集。
- 可以利用训练数据集对$P(\omega_j)$和$p(\boldsymbol{x}|\omega_j)$进行估计。
- 贝叶斯分类器的训练，就是从训练数据集中估计出先验概率$P(\omega_j)$和类条件概率密度$p(\boldsymbol{x}|\omega_j)$。

先验概率$P(\omega_j)$的估计

- 估计先验概率经常使用的方法：

① 当训练集中的样本数量足够多，且每个样本都是从样本空间中随机抽取的。

用训练集中ω_j类样本所占的比例来估计$P(\omega_j)$的值：

$$\hat{P}(\omega_j) = \frac{n_j}{N}$$

其中，n_j为训练集中ω_j类样本的总数，N为训练集中样本的总数。

② 如果训练样本集不是随机抽样得到的：

可以假设各类样本出现的概率是相等的(均匀先验)，即取$P(\omega_j) = 1/c$，其中c是类别的总数。

类条件概率密度$p(\boldsymbol{x}|\omega_j)$的估计

- 类条件概率密度 $p(\boldsymbol{x}|\omega_j)$ 表示ω_j类特征向量取值的分布情况，估计$p(\boldsymbol{x}|\omega_j)$的方法：

①参数化的方法：

先假定$p(\boldsymbol{x}|\omega_j)$具有某种确定的分布形式，例如正态分布、二项分布等，只是分布的参数未知，再用训练集对分布的未知参数进行估计。

②非参数化的方法：

直接对概率密度函数$p(\boldsymbol{x}|\omega_j)$本身进行估计，而不必假设$p(\boldsymbol{x}|\omega_j)$具有某个确定的分布形式。

理论上，能够估计任意形式的概率分布。

参数化的估计方法

- 最大似然估计 (Maximum Likelihood Estimation)
  - ① 把待估计参数$\boldsymbol{\theta}$看作是固定的量，只是其取值未知。
  - ② 找到一个/组参数值，使得训练样本集所有样本出现的联合概率密度$p(\mathcal{D}^j|\boldsymbol{\theta})$最大化。
- 贝叶斯估计(Bayesian Estimation)
  - ① 把待估计的参数$\boldsymbol{\theta}$看作是随机量，具有某个已知的先验概率密度函数$p(\boldsymbol{\theta})$。
  - ② 当观察到ω_j类的样本集$\mathcal{D}^j$后，能够把参数$\boldsymbol{\theta}$的先验概率密度$p(\boldsymbol{\theta})$转化为后验概率密度函数$p(\boldsymbol{\theta}|\mathcal{D}^j)$，求该分布的数学期望作为参数的估计值。

最大似然估计法

- 将训练数据集$\mathcal{D}$按照类别划分为：$\mathcal{D}^1,\mathcal{D}^2\cdots\mathcal{D}^c$，其中$\mathcal{D}=\mathcal{D}^1\cup\mathcal{D}^2\cdots\cup\mathcal{D}^c$，$\mathcal{D}^j$为$\omega_j$类样本的集合。

$$\mathcal{D}\left\{\begin{array}{ccc} \boxed{\mathcal{D}^1} & p(\boldsymbol{x}|\omega_1) & \boldsymbol{\theta}^1 \\ \boxed{\mathcal{D}^2} & p(\boldsymbol{x}|\omega_2) & \boldsymbol{\theta}^2 \\ \vdots & \vdots & \vdots \\ \boxed{\mathcal{D}^j} & p(\boldsymbol{x}|\omega_j) & \boldsymbol{\theta}^j \\ \vdots & \vdots & \vdots \\ \boxed{\mathcal{D}^c} & p(\boldsymbol{x}|\omega_c) & \boldsymbol{\theta}^c \end{array}\right.$$

最大似然估计法

- 最大似然估计法假设：

①每类的$p(\boldsymbol{x}|\omega_j)$的形式已知，但是其参数$\boldsymbol{\theta}^j$是未知的。

②每类的样本$\mathcal{D}^j$都是独立地根据该类的类条件件密度函数$p(\boldsymbol{x}|\omega_j)$所定义的分布抽取得到的，即每类的样本是独立同分布的(i.i.d.)。

③每类的样本与其余类的类条件概率密度函数无关。因此要估计$p(\boldsymbol{x}|\omega_j)$的参数$\boldsymbol{\theta}^j$，只需用$\mathcal{D}^j$。

原问题：利用$\mathcal{D}$对$p(\boldsymbol{x}|\omega_1), \cdots p(\boldsymbol{x}|\omega_c)$进行估计。

简化为：利用$\mathcal{D}^j$对$p(\boldsymbol{x}|\omega_j)$的参数$\boldsymbol{\theta}^j$进行估计，$j = 1,2\cdots c$。

最大似然估计法

- 假设$\mathcal{D}^j$中包含n个实例/样本，$\mathcal{D}^j = \{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_n\}$。
- 将ω_j类的类条件概率密度$p(\boldsymbol{x}|\omega_j)$表示为$p(\boldsymbol{x})$，同时，为了强调$p(\boldsymbol{x})$依赖于参数$\boldsymbol{\theta}^j$，把它重写成$p(\boldsymbol{x}|\boldsymbol{\theta}^j)$；

- 根据MLE的假设：
  - ✓ $\mathcal{D}^j$中每个$\boldsymbol{x}_i$都是根据密度函数为$p(\boldsymbol{x}|\boldsymbol{\theta}^j)$的分布独立采样得到的。
  - ✓ 密度函数$p(\boldsymbol{x}|\boldsymbol{\theta}^j)$的分布形式已知，其分布参数$\boldsymbol{\theta}^j$是未知的。

最大似然估计法

- 在参数$\boldsymbol{\theta}^j$给定条件下，样本集$\mathcal{D}^j$中所有样本的联合概率密度可以表示为:

$$p(\mathcal{D}^j|\boldsymbol{\theta}^j) = p(\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_n|\boldsymbol{\theta}^j) = \boxed{\prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\theta}^j)}$$

$p(\mathcal{D}^j|\boldsymbol{\theta}^j)$是关于参数$\boldsymbol{\theta}^j$的函数，称为$\boldsymbol{\theta}^j$的似然函数:

$$\boxed{L(\boldsymbol{\theta}^j) = p(\mathcal{D}^j|\boldsymbol{\theta}^j)}$$

- 最大似然估计：就是找到最优的$\boldsymbol{\theta}^j$取值，使得似然函数$L(\boldsymbol{\theta}^j)$取得最大值。
- 假设似然函数满足连续可微的条件，按照一般的求极值点方法，则在极值点处函数的梯度为零向量:

$$\boxed{\boldsymbol{\nabla}_{\boldsymbol{\theta}^j} L(\boldsymbol{\theta}^j) = \boldsymbol{0}}$$

最大似然估计

- 由于似然函数是乘积形式，不容易求导。因此依据对数函数的单调递增性，用对数似然函数进行参数估计：

$$\ln L(\boldsymbol{\theta}^j) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}^j)$$

- 令对数似然函数关于$\boldsymbol{\theta}^j$的梯度为零向量，求得对数似然函数的极值点，从中找到最值点就是$\boldsymbol{\theta}^j$的估计值。

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}^j} \ln L(\boldsymbol{\theta}^j) = \sum_{i=1}^{n} \boldsymbol{\nabla}_{\boldsymbol{\theta}^j} \ln p(\boldsymbol{x}_i|\boldsymbol{\theta}^j) = \mathbf{0}$$

- 注意: 满足梯度为零向量的解可能有多个，要对每个解进行检查，找到全局最优解，还要检查边界条件。

例题1:

- 假设$\mathcal{D}^j$中样本是根据正态分布$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$采样得到的，其中参数$\boldsymbol{\mu}, \boldsymbol{\Sigma}$是未知的。要求用MLE对$\boldsymbol{\mu}, \boldsymbol{\Sigma}$进行估计。
- 求解过程：

①写出似然函数$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma})$

②写出对数似然函数$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

③对$\ln L(\boldsymbol{\mu}, \boldsymbol{\Sigma})$分别关于$\boldsymbol{\mu}, \boldsymbol{\Sigma}$求梯度，令其为零。

按照上述过程，最终估计得到：$\widehat{\boldsymbol{\mu}}$ 和$\widehat{\boldsymbol{\Sigma}}$

例题1:

➢ 求解过程:

$$L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \left(\frac{1}{(2\pi)^{\frac{d}{2}}\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\right)^{n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right]$$

$$\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{dn}{2}\ln 2\pi - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

$$\boldsymbol{\nabla}_{\boldsymbol{\mu}}\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2}\sum_{i=1}^{n}2\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) = \mathbf{0}$$

$$\boldsymbol{\nabla}_{\boldsymbol{\Sigma}}\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{n}{2}(\boldsymbol{\Sigma}^{-1})^{\top} + \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\Sigma}^{-\top}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-\top} = \mathbf{0}$$

求解上述两个方程得到:

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\,, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})^{\top}(\boldsymbol{x}_i-\widehat{\boldsymbol{\mu}})$$

观察发现，$\widehat{\boldsymbol{\mu}}$是$\boldsymbol{\mu}$的无偏估计，而$\widehat{\boldsymbol{\Sigma}}$是$\boldsymbol{\Sigma}$的有偏估计。

常用求导公式:

- 对矩阵求导常用公式(大写字母表示矩阵，小写字母表示向量):

$$\frac{\partial\, \boldsymbol{a}^\top \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^\top$$

$$\frac{\partial\, \boldsymbol{a}^\top \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = - \boldsymbol{X}^{-\top}\boldsymbol{a}\boldsymbol{b}^\top \boldsymbol{X}^{-\top}$$

$$\frac{\partial\, ln\, |\boldsymbol{X}|}{\partial \boldsymbol{X}} = (\boldsymbol{X}^{-1})^\top$$

例题2：

- 假设$\mathcal{D}^j$的样本是根据$Bernoulli(\theta)$分布采样得到的，即$p(x|\theta) = \theta^x(1-\theta)^{1-x}$，其中$x = 0$或$1$，$0 \le \theta \le 1$。用MLE对$\theta$进行估计。
- 求解过程：

①写出似然函数$L(\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)}$

②写出对数似然函数$\ln L(\theta) = (\sum_{i=1}^n x_i)\ln\theta + (\sum_{i=1}^n (1-x_i))\ln(1-\theta)$

③对$\ln L(\theta)$分别关于θ求导数，令其为零。

最终得到：

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$$

贝叶斯估计法

- 贝叶斯估计(Bayesian Estimation)
  ① 把待估计的参数$\boldsymbol{\theta}$看作是随机量，具有某个已知的先验概率密度函数$p(\boldsymbol{\theta})$。
  ② 观察到某类的样本集$\mathcal{D}^j$后，能够把参数$\boldsymbol{\theta}$的先验概率密度$p(\boldsymbol{\theta})$转化为后验概率密度$p(\boldsymbol{\theta}|\mathcal{D}^j)$，求该分布的数学期望作为参数的估计值。

- 设$\mathcal{D}^j=\{\boldsymbol{x}_1,\boldsymbol{x}_2\cdots\boldsymbol{x}_n\}$，贝叶斯估计法基本假设:
  - ✓ $p(\boldsymbol{x}|\boldsymbol{\theta}^j)$的形式已知，未知参数$\boldsymbol{\theta}^j$是一个随机量，具有已知的先验概率密度函数$p(\boldsymbol{\theta}^j)$;
  - ✓ $\mathcal{D}^j$中的样本都是独立地根据密度为$p(\boldsymbol{x}|\boldsymbol{\theta}^j)$的分布采样得到的，$\mathcal{D}^j$中样本与$p(\boldsymbol{x}|\boldsymbol{\theta}^i)$无关，$i\neq j$。

贝叶斯估计法

➢ 基本估计步骤：

①计算观察到$\boldsymbol{\theta}^j$后，$\mathcal{D}^j$中所有样本的联合概率密度：

$$p(\mathcal{D}^j|\boldsymbol{\theta}^j) = p(\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_n|\boldsymbol{\theta}^j) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\theta}^j)$$

②利用贝叶斯公式，计算观察到$\mathcal{D}^j$后$\boldsymbol{\theta}^j$的后验概率密度$p(\boldsymbol{\theta}^j|\mathcal{D}^j)$：

$$p(\boldsymbol{\theta}^j|\mathcal{D}^j) = \frac{p(\boldsymbol{\theta}^j)p(\mathcal{D}^j|\boldsymbol{\theta}^j)}{p(\mathcal{D}^j)} = \frac{p(\boldsymbol{\theta}^j)p(\mathcal{D}^j|\boldsymbol{\theta}^j)}{\int p(\boldsymbol{\theta}^j)p(\mathcal{D}^j|\boldsymbol{\theta}^j)d\boldsymbol{\theta}^j}$$

③参数$\boldsymbol{\theta}^j$的估计为：

$$\widehat{\boldsymbol{\theta}^j} = \int \boldsymbol{\theta}^j \, p(\boldsymbol{\theta}^j|\mathcal{D}^j) \, d\boldsymbol{\theta}^j$$

例题1:

- 给定一个样本集$\mathcal{D}=\{x_1,x_2\cdots x_n\}$，设$\mathcal{D}$中的每个样本都是根据一维的正态分布$\mathcal{N}(\mu,\sigma^2)$相互独立地采样得到的，其中$\mu$未知，$\sigma^2$已知。已知未知参数$\mu$服从先验概率分布$\mathcal{N}(\mu_0,\sigma_0^2)$。
  要求：用贝叶斯估计法对参数μ进行估计。
- 根据题目得知：$p(x|\mu)=\frac{1}{\delta\sqrt{2\pi}}exp^{\left[-\frac{1}{2\delta^2}(x-\mu)^2\right]}$

$$p(\mu)=\frac{1}{\delta_0\sqrt{2\pi}}exp^{\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]}$$

①计算$p(\mathcal{D}|\mu)$：

$$p(\mathcal{D}|\mu)=\prod_{i=1}^{n}p(x_i|\mu)$$

$$=2\pi^{-\frac{n}{2}}\delta^{-n}\ exp^{\left[-\frac{1}{2\delta^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]}$$

例题1:

②计算参数μ的后验概率密度$p(\mu|\mathcal{D})$:

$$p(\mu|\mathcal{D}) \propto p(\mu)p(\mathcal{D}|\mu)$$

$$\propto exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}$$

$$\propto exp\left\{-\frac{1}{2\sigma_n^2}(\mu-\mu_n)^2\right\}$$

$$\mu_n=\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0+\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\mu_{MLE},\quad \mu_{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

$$\frac{1}{\sigma_n^2}=\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}$$

先验信息和观测数据信息的加权平均!

③ 密度为$p(\mu|\mathcal{D})$的分布的数学期望是μ_n，所以对μ的贝叶斯估计是：$\hat{\mu}=\mu_n$

例题2:

- 给定一个训练样本集 $\mathcal{D} = \{x_1, x_2 \cdots x_n\}$，设$\mathcal{D}$中样本是根据$Bernoulli(\theta)$采样得到，但是参数$\theta$未知：

$$p(x|\theta) = \theta^x(1-\theta)^{1-x}$$

其中$x = 0$或1，$0 \leq \theta \leq 1$。

已知参数θ服从$Beta$分布$\theta \sim Beta(\alpha, \beta)$：

$$p(\theta) = constant \cdot \ \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

且$Beta$分布的期望为$\alpha/(\alpha+\beta)$。

要求：用贝叶斯估计法对参数θ进行估计。

例题2:

①计算$p(\mathcal{D}|\theta)$:

$$p(\mathcal{D}|\theta)=\prod_{i=1}^{n}p(x_i|\theta)=\theta^{\sum_{i=1}^{n}x_i}(1-\theta)^{n-\sum_{i=1}^{n}x_i}$$

②计算$p(\theta|\mathcal{D})$:

$$p(\theta|\mathcal{D})=\frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})}\propto\ \theta^{\alpha+\sum_{i=1}^{n}x_i-1}(1-\theta)^{\beta+n-\sum_{i=1}^{n}x_i-1}$$

$p(\theta|\mathcal{D})$是$Beta(\alpha+\sum_{i=1}^{n}x_i,\beta+n-\sum_{i=1}^{n}x_i)$的密度函数!

③对θ的贝叶斯估计是密度为$p(\theta|\mathcal{D})$的分布的期望: $(\theta_{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i)$

$$\hat{\theta}=\frac{\alpha+\sum_{i=1}^{n}x_i}{\alpha+\beta+n}=\frac{\alpha+\beta}{\alpha+\beta+n}\cdot\frac{\alpha}{\alpha+\beta}+\frac{n}{\alpha+\beta+n}\theta_{MLE}$$

共轭分布和共轭先验(conjugate distribution)

- 如果参数的后验分布$p(\theta|\mathcal{D})$与其先验分布$p(\theta)$具有相同的概率分布形式，那么称$p(\theta)$与$p(\theta|\mathcal{D})$为“共轭分布”。$p(\theta)$被称为似然函数$p(\mathcal{D}|\theta)$的“共轭先验”。
- 这种分布间的关系能够简化计算。

生成样本的概率分布 $p(x\|\theta)$	相应的共轭先验分布 $p(\theta)$
高斯分布	高斯分布
指数分布	Gamma分布
泊松分布	Gamma分布
二项分布	Beta分布
多项式分布	Dirichlet分布

小结

- 贝叶斯分类器的训练，就是从训练数据集中估计出先验概率$P(\omega_j)$和类条件概率密度函数$p(\boldsymbol{x}|\omega_j)$。
- 估计$p(\boldsymbol{x}|\omega_j)$的方法包括参数化的方法和非参数化的方法：

  参数化方法假定$p(\boldsymbol{x}|\omega_j)$具有某种确定的分布形式，但是分布的参数未知，然后利用训练集对分布的参数进行估计。

  非参数化的方法直接对$p(\boldsymbol{x}|\omega_j)$本身进行估计，而不必假设$p(\boldsymbol{x}|\omega_j)$是某个类型确定的分布的概率密度函数，能够估计任意分布的概率密度函数。

小结

- 参数化估计方法包括：最大似然估计和贝叶斯估计。

  最大似然估计基于训练样本集$\mathcal{D}$找到参数$\boldsymbol{\theta}$的一个估计值，使得$\mathcal{D}$中所有的样本被抽取的概率最大化。不同的样本集会导致不同的参数估计值。

  贝叶斯估计假设参数$\boldsymbol{\theta}$符合一个已知的先验分布，具有概率密度函数$p(\boldsymbol{\theta})$，然后利用观测到的训练样本集$\mathcal{D}$得到参数$\boldsymbol{\theta}$的后验概率密度函数$p(\boldsymbol{\theta}|\mathcal{D})$，求取该分布的数学期望作为$\boldsymbol{\theta}$的估计值。贝叶斯估计法得到的参数估计值是参数的先验信息和样本集$\mathcal{D}$中的信息的加权平均!

重难点