机器学习

第七讲 贝叶斯分类器

魏凤凤 助理教授

fengfeng_scut@163.com

计算机科学与工程学院

课程内容

- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网
- EM算法

背景知识

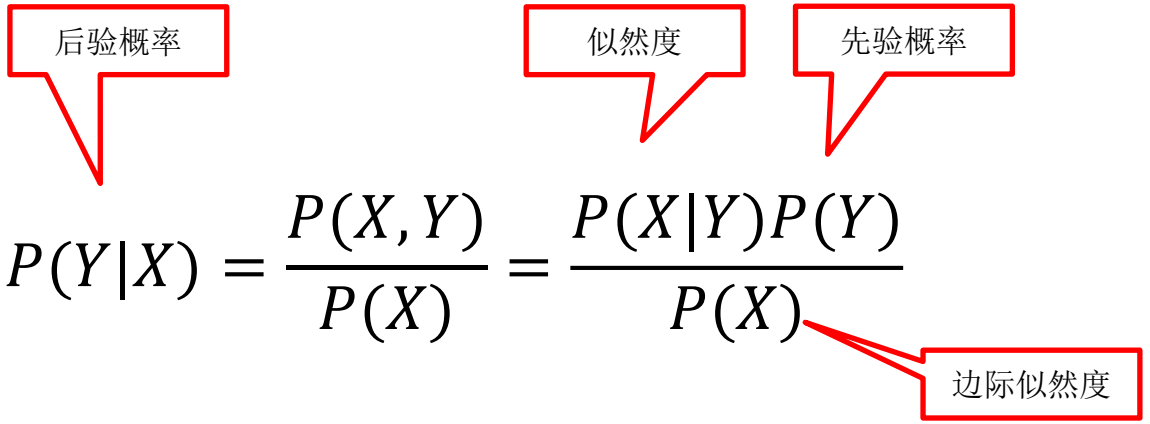
- **贝叶斯分类**
  - 贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。
- **先验概率**
  - 根据以往经验和分析得到的概率。我们用$P(Y)$来代表在没有训练数据前假设Y拥有的初始概率。
- **后验概率**
  - 根据已经发生的事件来分析得到的概率。以$P(Y|X)$代表假设X成立的情下观察到Y数据的概率，因为它反映了在看到训练数据X后Y成立的置信度。

背景知识

- **联合概率**
  - 联合概率是指在多元的概率分布中多个随机变量分别满足各自条件的概率。X与Y的联合概率表示为$P(X,Y)$、 $P(XY)$ 或 $P(X\cap Y)$ 。
  - 假设X和Y都服从正态分布，那么$P(X<5,Y<0)$就是一个联合概率，表示 $X<5,Y<0$两个条件同时成立的概率。表示两个事件共同发生的概率。

背景知识

- **贝叶斯公式**



  - $P(Y|X)$称为后验概率(posterior),这是我们需要结合先验概率和证据计算之后才能知道的。
  - $P(X|Y)$称为似然(likelihood),在事件Y发生的情况下，事件X(或evidence)的概率有多大
  - $P(Y)$称为先验概率(prior), 事件Y发生的概率有多大
  - $P(X)$称为证据(evidence)，即无论事件如何，事件X(或evidence)的可能性有多大。

在某些证据下，某个事件发生的可能性有多大。比如我们想知道一个人的血液检查结果为阳性，那么他得病的概率有多大？但是，我们只能知道在得病的条件下，血液检查结果呈阳性的概率为95%，即在给定事件下，知道证据发生的概率。

贝叶斯决策论

- **贝叶斯决策论（Bayesian decision theory）是在概率框架下实施决策的基本方法**
  - 在分类问题情况下，在所有相关概率都已知的理想情形下，贝叶斯决策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

- 假设有N种可能的类别标记，即$y = \{c_1, c_2, \ldots, c_N\}$，$\lambda_{ij}$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)，即在样本上的“条件风险”(conditional risk)：

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

贝叶斯决策论

- 寻找一个判定准则$h: X \to Y$，以最小化总体风险：

$$R(h) = \mathrm{E}_x[R(h(x)|x)]$$

- 显然，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$R(c|x)$最小的类别标记，即：

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

- 此时，被称为**贝叶斯最优分类器(Bayes optimal classifier)**，与之对应的总体风险$R(h^*)$称为**贝叶斯风险 (Bayes risk)**。

贝叶斯决策论

- $1-R(h^*)$称为反映了分类起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。**
- 具体来说，若目标是最小化分类错误率，则误判损失λ_{ij}可表达为：

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

- 此时，条件风险$R(c|x)=1-P(c|x)$
- 于是，最小化分类错误率的贝叶斯最优分类器为

$$h^*(x)=\underset{c\in y}{\operatorname{argmax}}\,P(c|x)$$

- 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

- 不难看出，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，在现实中通常难以直接获得。机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率$P(c|x)$。
- 主要有两种策略：
  - 判别式模型（Discriminative Models）
    - 给定x，通过直接建模$P(c|x)$，来预测c → 决策树、BP神经网络，SVM
  - 生成式模型（Generative Models）
    - 先对联合概率分布 $P(x,c)$建模，再由此获得$P(c|x)$ → 朴素贝叶斯

贝叶斯决策论

- 生成式模型：

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

- 基于贝叶斯定理，$P(c|x)$可写成：

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

类标记c相对于样本x的“类条件概率” (class-conditional probability), 或称“似然”

先验概率：样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定理）

“**证据**”(Evidence)因子，与类别标记无关

极大似然估计

- 估计类条件概率的常用策略：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参数估计。
- 记关于类别c的类条件概率为$P(x|c)$
  - 假设$P(x|c)$具有确定的形式被参数θ_c唯一确定，我们的任务就是利用训练集D估计参数θ_c

极大似然估计

- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的方案。
  - **频率主义学派 (Frequentist)**认为参数虽然未知，但却存在客观值，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 (Bayesian)**认为参数是未观察到的随机变量、其本身也可由分布，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计算参数的后验分布

极大似然估计

- 令D_c表示训练集中第c类样本的组合的集合，假设这些样本是独立的，则参数θ_c对于数据集D_c 的似然是：

$$P(D_c|\theta_c) = \prod_{x \in D_c} P(x|\theta_c)$$

- 对θ_c 进行极大似然估计，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，找到一个使数据出现的“可能性”最大值

极大似然估计

$$P(D_c|\theta_c) = \prod_{x \in D_c} P(x|\theta_c)$$

- 连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$LL(\theta_c) = logP(D_c|\theta_c) = \sum_{x \in D_c} logP(x|\theta_c)$$

- 此时，参数θ_c的极大似然估计$\hat{\theta}_c = \underset{\theta_c}{\operatorname{argmax}} LL(\theta_c)$

极大似然估计

- 例如，在连续属性情形下，假设概率密度函数$p(x|c)\sim N(\mu_c,\sigma_c^2)$，则参数$\mu_c,\sigma_c^2$的极大似然估计为：

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

$$\hat{\sigma}_c=\frac{1}{|D_c|}\sum_{x\in D_c}(x-\hat{\mu}_c)(x-\hat{\mu}_c)^{\mathrm{T}}$$

- 也就是说，通过极大似然法得到的正态分布均值就是样本均值，方差就是$(x-\hat{\mu}_c)(x-\hat{\mu}_c)^{\mathrm{T}}$的均值，这显然是一个符合直觉的结果。
- 需注意的是，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。

朴素贝叶斯分类器

- 估计后验概率$P(c|x)$主要困难：类条件概率$P(x|c)$是所有属性上的联合概率难以从有限的训练样本估计获得。
- 朴素贝叶斯分类器(Naïve Bayes Classifier)采用了“属性条件独立性假设”(attribute conditional independence assumption)：每个属性独立地对分类结果发生影响。
- 基于属性条件独立习惯假设，朴素贝叶斯模型可改写为

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

- 其中，d为属性数目，x_i为x在第i个属性上的取值。

朴素贝叶斯分类器

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

- 由于对所有类别来说，$P(x)$相同，因此基于贝叶斯判定准则，有

$$h^*(x) = \underset{c \in y}{\operatorname{argmax}}\, P(c)\prod_{i=1}^{d} P(x_i|c)$$

- **朴素贝叶斯分类器**

朴素贝叶斯分类器

- 朴素贝叶斯分类器训练过程：
  - 基于训练集D估计类先验概率$P(c)$，并为每个属性估计条件概率$P(x_i|c)$
- 令D_c表示训练集中第c类样本的组合的集合，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

- 朴素贝叶斯分类器训练过程：
  - 基于训练集D估计类先验概率$P(c)$，并为每个属性估计条件概率$P(x_i|c)$
- 令D_c表示训练集中第c类样本的组合的集合，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：

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

- 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，则概率条件$P(x_i|c\ )$可估计为：

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

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

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

拉普拉斯修正

- 例如：用西瓜数据集3.0训练一个朴素贝叶斯分类器，对测试例“测1”进行分类
- (p151, 西瓜数据集 p84 表4.3)

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

拉普拉斯修正

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题。
- 比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理。

拉普拉斯修正

- 若某个属性值在训练集中没有与某个类同时出现过，则直接计算会出现问题。
- 比如“敲声=清脆”测试例，训练集中没有该样例，因此连乘式计算的概率值为0，无论其他属性上明显像好瓜，分类结果都是“好瓜=否”，这显然不合理。
- 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“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，则可修正为：

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

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

朴素贝叶斯分类器

- 假设我们正在构建一个分类器，该分类器说明文本是否与运动(Sports)有关。我们的训练数据有5句话：

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

我们想要计算句子“A very close game”是 Sports 的概率以及它不是 Sports 的概率。

即$P(\text{Sports} \mid \text{a very close game})$这个句子的类别是Sports的概率

朴素贝叶斯分类器

- 特征：单词的频率

已知贝叶斯定理$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$，则：

$$P(\text{ Sports } | \text{ a very close game }) = \frac{P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })}{P(\text{ a very close game })}$$

由于我们只是试图找出哪个类别有更大的概率，可以舍弃除数，只是比较

$P(\text{ a very close game } | \text{ Sports }) \times P(\text{ Sports })$ 和$P(\text{ a very close game } | \text{ Not Sports }) \times P(\text{ Not Sports })$

朴素贝叶斯分类器

- 我们假设一个句子中的每个单词都与其他单词无关。

$$P(\text{ a very close game }) = P(a) \times P(\text{ very }) \times P(\text{ close }) \times P(\text{ game })$$

$$P(\ a\ very\ close\ game|Sports\ )$$
$$= P(a|\ Sports\ ) \times P(\ very\ |\ Sports) \times P(\ close\ |\ Sports\ ) \times P(\ game\ |\ Sports\ )$$

朴素贝叶斯分类器

- 计算每个类别的先验概率：
- 对于训练集中的给定句子，$P(\text{Sports})$的概率为⅗。$P(\text{Not Sports})$是⅖。
- 然后，计算$P(game|Sports)$就是“game”有多少次出现在Sports的样本，然后除以sports为标签的文本的单词总数（3+3+5=11）。
- 因此，$P(game|Sports)=\frac{2}{11}$。
- “close”不会出现在任何sports样本中！那就是说$P(close|Sports)=0$。

文本	标签
A great game	Sports
The election was over	Not Sports
Very clean match	Sports
A clean but forgettable game	Sports
It was a close election	Not Sports

朴素贝叶斯分类器

- 通过使用拉普拉斯平滑的方法：我们为每个计数加1，因此它永远不会为零。为了平衡这一点，我们将可能单词的数量添加到除数中，因此计算结果永远不会大于1。
- 在这里的情况下，可能单词是['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match']。
- 由于可能的单词数是14，因此应用平滑处理可以得到
- $P(\text{game} \mid \text{sports}) = \frac{2+1}{11+14}$

14个单词

朴素贝叶斯分类器

Word	P (word | Sports)	P (word | Not Sports)
a	$(2+1) \div (11+14)$	$(1+1) \div (9+14)$
very	$(1+1) \div (11+14)$	$(0+1) \div (9+14)$
close	$(0+1) \div (11+14)$	$(1+1) \div (9+14)$
game	$(2+1) \div (11+14)$	$(0+1) \div (9+14)$

$$P(a|\text{ Sports }) \times P(\text{ very }|\text{ Sports }) \times P(\text{ close }|\text{ Sports }) \times P(\text{ game }|\text{ Sports }) \times P(\text{ Sports })$$
$$= 2.76 \times 10^{-5} = 0.0000276$$

$$P(a|\text{ Not Sports }) \times P(\text{ very }|\text{ Not Sports }) \times P(\text{ close }|\text{ Not Sports })$$
$$\times P(\text{ game }|\text{ Not Sports }) \times P(\text{ Not Sports})$$
$$= 0.572 \times 10^{-5} = 0.00000572$$

由于0.0000276大于0.00000572，我们的分类器预测“A very close game”是Sport类。

半朴素贝叶斯分类器

- 为了降低贝叶斯公式中估计后验概率的困难，朴素贝叶斯分类器采用的属性条件独立性假设；对属性条件独立假设记性一定程度的放松，由此产生了一类称为“半朴素贝叶斯分类器” (semi-naïve Bayes classifiers)

谢谢

魏凤凤 助理教授

fengfeng_scut@163.com

计算机科学与工程学院