机器学习

06 贝叶斯分类器

李祎

liyi@dlut.edu.cn

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

目录

贝叶斯决策论

先验概率

■ 鲈鱼与鲑鱼分类问题

- 用w标记类别

$$w = w_1 \quad : \text{鲈鱼}$$
$$w = w_2 \quad : \text{鲑鱼}$$

- 抓到鲈鱼与鲑鱼的事件是随机的，可以用概率（可能性）来表示：

$$P(w_1) \quad : \text{下一条鱼是鲈鱼的先验概率}$$
$$P(w_2) \quad : \text{下一条鱼是鲑鱼的先验概率}$$
$$P(w_1)+P(w_2)=1$$

- 先验概率：反映了我们的经验知识，例如随季节和海域的不同鱼的分布情况不同。

先验概率

一种简单的判决准则

●例如：假设春季在某海域是：鲈鱼的概率是：99.9%，鲑鱼的概率是：0.01%，则可认为打捞上来的鱼是鲈鱼。在冬季他们出现的概率相反，则可认为打捞上来的鱼是鲑鱼。

如果 $P(w_1) > P(w_2) \rightarrow w = w_1$ 即分类为鲈鱼

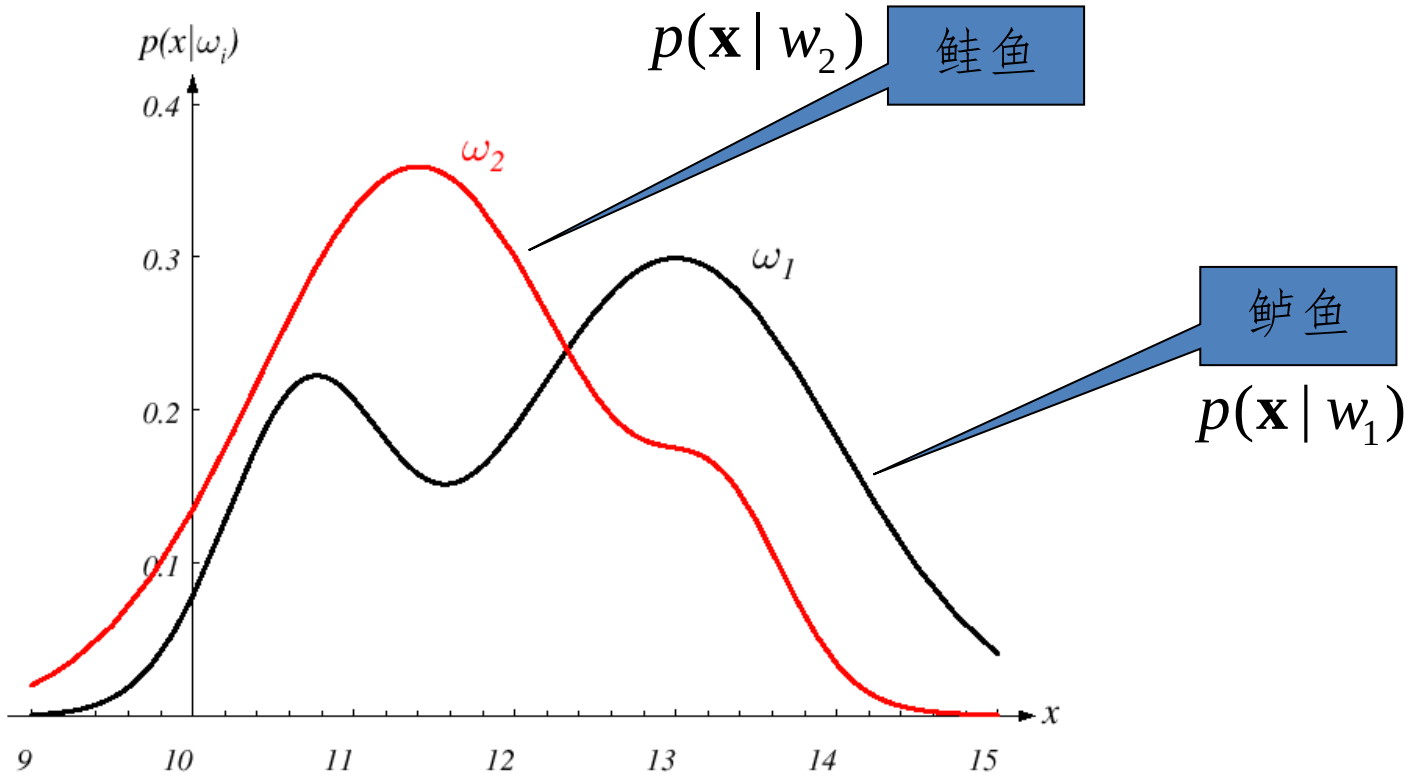
如果 $P(w_2) > P(w_1) \rightarrow w = w_2$ 即分类为鲑鱼

● 当两类鱼出现的概率相差不大时？

● 需要更多的特征信息：如鱼的长度，鱼的外表色泽的深浅、嘴的位置和大小等等。

似然概率

- **似然概率：** 特征的类条件概率密度分布
  - 特征量：鱼外表色泽亮度x
  - 根据大量的统计可以获得鲈鱼和鲑鱼关于亮度x的类条件概率分布情况：



最小错误率贝叶斯

二类判决问题

- 假设已知：

1. 鲈鱼和鲑鱼的先验概率：$p(w_1)$ 和 $p(w_2)$
2. 亮度特征$\mathbf{x}$的类条件概率密度：$p(\mathbf{x}|w_1)$ 和 $p(\mathbf{x}|w_2)$
3. 当前待分类的这条鱼的亮度观测值$\mathbf{x}$

- 判断观测值$\mathbf{x}$属于鲈鱼和鲑鱼的概率情况？

$$p(w_1|\mathbf{x}) \text{ 和 } p(w_2|\mathbf{x})$$

一个合理的规则：

$$p(w_1|\mathbf{x}) > p(w_2|\mathbf{x}) \rightarrow w = w_1$$
$$p(w_2|\mathbf{x}) > p(w_1|\mathbf{x}) \rightarrow w = w_2$$

最小错误率贝叶斯

■ 贝叶斯公式

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

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

- Posterior: $P(w_i \mid \mathbf{x})$ 观测到具有x属性的示例或样本，该样本属于 w_i 的概率；
- Likelihood: $P(\mathbf{x} \mid w_i)$ 似然，即第 w_i 类样本， x属性或特征的分布情况；
- Prior: $P(w_i)$ 先验概率，样本空间中各类样本所占的比例，根据大数定理，可通过各类样本出现的频率估计（大数定理）
- Evidence: “证据” 因子，与类标记无关，保证类别后验概率之和为1。

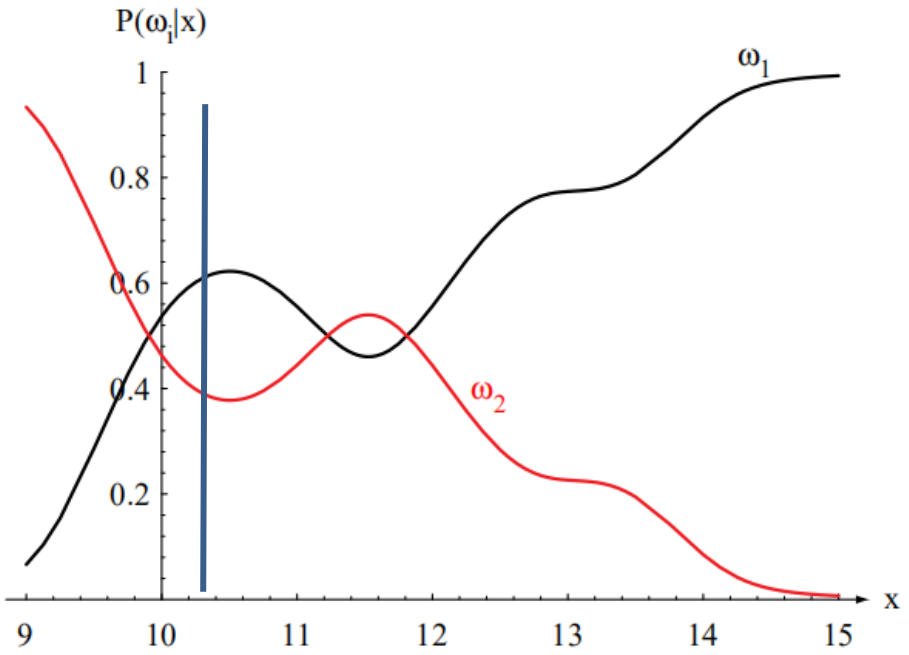
■ 最大后验分类规则：

$$w^* = \arg\max\{P(w_i \mid \mathbf{x})\}$$

最小错误率贝叶斯

$$P(error \mid x) = \begin{cases} P(w_1 \mid x), & w = w_2 \\ P(w_2 \mid x), & w = w_1 \end{cases}$$

$$P(w_1 \mid x) + P(w_2 \mid x) = 1$$



贝叶斯判定准则（Bayes decision rule）：为最小化总体错误率，只需在每个样本上选择那个能使错误最小的类别标记。

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)p(x)dx$$

$$\int_{-\infty}^{\infty} \min\{P(error \mid x)\}p(x)dx \Rightarrow \min P(error)$$

$$p(w_1 \mid \mathbf{x}) > p(w_2 \mid \mathbf{x}) \rightarrow w = w_1$$
$$p(w_2 \mid \mathbf{x}) > p(w_1 \mid \mathbf{x}) \rightarrow w = w_2$$

最小错误率贝叶斯

■ 举例：

●为了对癌症进行诊断，对一批人进行一次普查，给每个人打试验针，观察反应，然后进行统计，规律如下：

(1) 这一批人中，每1000个人中有5个癌症病人；

(2) 这一批人中，每100个正常人中有一个试验呈阳性反应；

(3) 这一批人中，每100个癌症病人中有95人试验呈阳性反应。

问：若某人（甲）呈阳性反应，甲是否正常？

- $w_1 \rightarrow$ 正常； $w_2 \rightarrow$ 癌症
- $P(w_1) = 0.995, P(w_2) = 0.005$
- $P(\mathbf{x} \mid w_1) = 0.01, P(\mathbf{x} \mid w_2) = 0.95$

最小错误率贝叶斯

$$P(\mathbf{x}\,|\,w_1)\cdot P(w_1)=0.00995$$

$$P(\mathbf{x}\,|\,w_2)\cdot P(w_2)=0.00475$$

- w_1 → 正常; w_2 → 癌症
- $P(w_1)=0.995, P(w_2)=0.005$
- $P(\mathbf{x}|w_1)=0.01, P(\mathbf{x}|w_2)=0.95$

$$P(w_2\,|\,x)=\frac{P(\mathbf{x}\,|\,w_2)\cdot P(w_2)}{P(\mathbf{x}\,|\,w_1)\cdot P(w_1)+P(\mathbf{x}\,|\,w_2)\cdot P(w_2)}=0.323$$

$$P(w_1\,|\,\mathbf{x})=1-P(w_2\,|\,\mathbf{x})=1-0.323=0.677$$

$$P(w_1\,|\,\mathbf{x})>P(w_2\,|\,\mathbf{x})\Leftrightarrow P(\mathbf{x}\,|\,w_1)\cdot P(w_1)>P(\mathbf{x}\,|\,w_2)\cdot P(w_2)$$

若某人（甲）呈阳性反应，甲是否正常？

$\mathbf{x}\in w_1$

最小风险贝叶斯

- 最小错误率贝叶斯=>最大后验概率规则

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

$$Posterior = \frac{Likelihood \times Prior}{Evidence}$$

$$w^* = \arg\max \{P(w_i \mid \mathbf{x})\}$$

$$= \arg\max \{P(\mathbf{x} \mid w_i)P(w_i)\}$$

- 问题扩展

1. 允许有其它行为而不仅是判定类别。
2. 引入损失函数，比错误率更具一般性。

最小风险(Minimum Risk)贝叶斯!

最小风险贝叶斯

■问题定义

- 令 $\{w_1, w_2, ..., w_c\}$ 表示一系列类别状态。
- 令 $\{\alpha_1, \alpha_2, ..., \alpha_a\}$ 表示一系列可能采取的行动（或决策）。
- 令 $\lambda(\alpha_i \mid w_j)$ 表示当实际类别状态为 w_j 时，采取 α_i 的行为会带来的风险。那么，与行动 α_i 相关联的损失

$$R(\alpha_i \mid \mathbf{x}) = \sum \lambda(\alpha_i \mid w_j) P(w_j \mid \mathbf{x})$$

■最小风险判决步骤：

- 在给定样本$\mathbf{x}$条件下，计算各类后验概率 $P(w_j \mid \mathbf{x})$
- 求各种判决的条件平均风险 $R(\alpha_i \mid \mathbf{x}) = \sum \lambda(\alpha_i \mid w_j) P(w_j \mid \mathbf{x})$
- 比较各种判决的条件平均风险，把样本$\mathbf{x}$归属于条件平均风险最小的那一种判决 $\alpha^* = \arg\min_i \{R(\alpha_i \mid \mathbf{x})\}$

最小风险贝叶斯

- 举例：
  - 在癌症诊断问题中，所有的化验结果可分为两类。w_1 正常，w_2 癌症。采取的判决行为也有两种 α_1（正常）和 α_2（癌症）。
  - 风险矩阵

判决 \ 类型	w_1	w_2
α_1	0.5	6
α_2	2	0.5

无病判成无病的风险

无病判成有病的风险

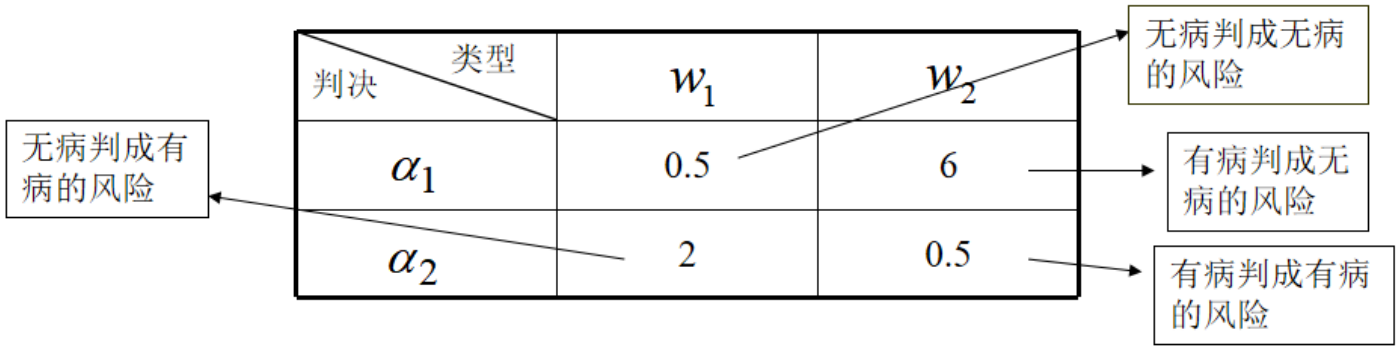
有病判成无病的风险

有病判成有病的风险

  - $P(w_1 \mid \mathbf{x}) = 0.677, P(w_2 \mid \mathbf{x}) = 0.323$

$$R(\alpha_i \mid \mathbf{x}) = \sum \lambda(\alpha_i \mid w_j) P(w_j \mid \mathbf{x})$$

最小风险贝叶斯



w_1：正常，w_2：癌症

α_1：正常，α_2：癌症

- $\lambda(\alpha_1 \mid w_1) = 0.5,\ \ \lambda(\alpha_1 \mid w_2) = 6\ ,\ \lambda(\alpha_2 \mid w_1) = 2\ ,\ \lambda(\alpha_2 \mid w_2) = 0.5$
- $P(w_1 \mid \mathbf{x}) = 0.677,\ P(w_2 \mid \mathbf{x}) = 0.323$

$$R(\alpha_1 \mid \mathbf{x}) = \sum_{j=1}^{2} \lambda(\alpha_1 \mid w_j) \cdot P(w_j \mid \mathbf{x}) = 0.5 \times 0.677 + 6 \times 0.323 = 2.2765$$

$$R(\alpha_2 \mid \mathbf{x}) = \sum_{j=1}^{2} \lambda(\alpha_2 \mid w_j) \cdot P(w_j \mid \mathbf{x}) = 2 \times 0.677 + 0.5 \times 0.323 = 1.5155$$

$$R(\alpha_2 \mid \mathbf{x}) < R(\alpha_1 \mid \mathbf{x})$$

思考

- 最小错误率贝叶斯是最小风险贝叶斯的一个特例

什么情况下，最小风险判决规则退化成最小错误率判决规则？

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。
- 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 $P\{c_i \mid \mathbf{x}\}$ 可获得将样本 $\mathbf{x}$ 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）

$$R(c_i \mid \mathbf{x}) = \sum_{j=1}^{N} \lambda_{ij} P(c_j \mid \mathbf{x}) \tag{7.1}$$

- 我们的任务是寻找一个判定准则 $h : X \mapsto Y$ 以最小化总体风险

$$R(h) = \mathbf{E}_x \left[R(h(\mathbf{x}) \mid \mathbf{x}) \right]$$

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 $P(c \mid \mathbf{x})$ 。
- 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 $P(c \mid \mathbf{x})$。
- 主要有两种策略：
  - 判别式模型（discriminative models）
    - 给定 $\mathbf{x}$，通过直接建模 $P(c \mid \mathbf{x})$，来预测 c
    - 决策树，BP神经网络，支持向量机
  - 生成式模型（generative models）
    - 先对联合概率分布 $P(\mathbf{x}, c)$ 建模，再由此获得 $P(c \mid \mathbf{x})$
    - 生成式模型考虑 $P(c \mid \mathbf{x}) = \dfrac{P(\mathbf{x}, c)}{P(\mathbf{x})}$

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

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）

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

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，与类标记无关

贝叶斯决策论

- 生成式模型

$$P(c \mid \mathbf{x}) = \frac{P(\mathbf{x}, c)}{P(\mathbf{x})} \tag{7.7}$$

- 基于贝叶斯定理，$P(c \mid \mathbf{x})$ 可写成

$$P(c \mid \mathbf{x}) = \frac{P(c)P(\mathbf{x} \mid c)}{P(\mathbf{x})} \tag{7.8}$$

类标记 c 相对于样本 $\mathbf{x}$ 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 $P(\mathbf{x} \mid c)$，
  - 假设 $P(\mathbf{x} \mid c)$ 具有确定的形式被参数 $\boldsymbol{\theta}_c$ 唯一确定，我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$
- 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 $\boldsymbol{\theta}_c$ 对于数据集 D_c 的似然是

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

  - 对 $\boldsymbol{\theta}_c$ 进行极大似然估计，寻找能最大化似然 $P(D_c \mid \theta_c)$ 的参数值 $\hat{\theta}_c$ 。直观上看：极大似然估计是试图在 $\boldsymbol{\theta}_c$ 所有可能的取值中，找到一个使数据出现的“可能性”最大值。

- 连乘操作易造成下溢，通常使用对数似然(log-likelihood)

$$\begin{aligned} LL(\boldsymbol{\theta_c}) &= \log P(D_c \mid \boldsymbol{\theta_c}) \\ &= \sum_{\mathbf{x} \in D_c} \log P(\mathbf{x} \mid \boldsymbol{\theta}_c) \end{aligned}$$

- 此时参数 $\boldsymbol{\theta}_c$ 的极大似然估计 $\hat{\boldsymbol{\theta}}_c$ 为 $\hat{\boldsymbol{\theta}}_{\boldsymbol{c}} = \underset{\boldsymbol{\theta}_c}{\operatorname{argmax}} LL(\boldsymbol{\theta}_c)$

极大似然估计

- 例如，在连续属性情形下，假设概率密度函数 $p(\mathbf{x} \mid c) \sim N(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c^2)$，则参数 $\boldsymbol{\mu}_c$ 和 $\boldsymbol{\sigma}_c^2$ 的极大似然估计为

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

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

- 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 $(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)(\mathbf{x} - \hat{\boldsymbol{\mu}}_c)^{\mathrm{T}}$ 的均值，这显然是一个符合直觉的结果。

注意：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。