统计判别之贝叶斯判别

2019 年 12 月 10 日

目录

1 基本概念 **2**

2 简单贝叶斯判别 **2**

3 贝叶斯决策论——推广的简单贝叶斯判别 **2**

3.1 贝叶斯决策论 2

3.2 回到二分类 3

4 分类器、判别函数及判定面 **4**

4.1 贝叶斯分类器 4

4.2 二类情况下的贝叶斯分类器 5

5 正态分布的贝叶斯分类器 **5**

5.1 正态分布 5

5.2 一些基本概念 6

5.3 正态分布下，贝叶斯决策论的判别函数 7

5.4 两类问题且其类模式都是正态分布的特殊情况 7

6 均值向量和协方差矩阵的参数估计 **8**

6.1 均值和协方差矩阵的非随机参数的估计 8

6.2 均值向量和协方差矩阵的贝叶斯学习：随机参数估计 10

1 基本概念

1. 概率：给定参数 θ 下，样本随机变量 $\mathbf{X} = \mathbf{x}$ 的可能性。
2. 似然：给定样本 $\mathbf{X} = \mathbf{x}$ 参数 θ_1 为真实值的可能性。
3. 先验概率 (Prior)：根据以往的经验得到的概率。已知的概率。$P(w_i)$
4. 后验概率：给定参数下在考虑和给出相关证据或数据后所得到的条件概率 $P(w_j|x)$
5. 证据因子：$p(x)$
6. 为什么叫 Naive Bayes：因为各个特征之间相互独立。

2 简单贝叶斯判别

贝叶斯决策论是基于常识的决策过程的形式化。

在给定特征向量 $\mathbf{x}$ 的条件下，判断为类别 w_i 的概率最大的类别 i。——这是符合常识的，那个类别更有可能就判断给那个类别。计算后验概率的依据为贝叶斯公式:

$$P(w_j|x) = \frac{p(x|w_j)P(w_j)}{p(x)} \tag{1}$$

判断失误的概率

$$P(error) = \begin{cases} P(w_1, |x), & if\ judged\ as\ w_2 \\ P(w_2, |x), & if\ judged\ as\ w_1 \end{cases}$$

结合 (1) 公式，可得到:

$$P(error) = min[P(w_1|x), P(w_2|x)] \tag{2}$$

对于二分类，带入贝叶斯公式可知，只需要比较 $P(x|w_i)P(w_i)$ 即可。

3 贝叶斯决策论——推广的简单贝叶斯判别

3.1 贝叶斯决策论

相比于简单的贝叶斯判别，贝叶斯决策论在如下几个方面做了推广:

- 多特征：特征值 x 变为特征向量 $\mathbf{x}$。可以根据多种特征值进行判断。
- 多类别：不止有两类，而是有 d 类，特征空间为 $\mathbf{R}^d$。
- 多行为：行为可以指判决动作，也可以指代其他。多行为主要是为了允许存在拒绝决策的可能性。比如，在后验概率相近的情况下可以拒绝作出判断，如果因此所付出的代价不太大的话。
- 引入损失函数：损失函数用来精确阐释美中判决的代价。

其他的一些符号定义：

- 决策函数 α：$\alpha(\mathbf{x})$ 指根据输入 $\mathbf{x}$，判别函数会告诉我们该如何分类，输出为一个确定的 α 数值 $\alpha_1, \alpha_2 \dots \alpha_a$
- 损失量 $\lambda(\alpha_i|w_j)$ 指把类别 j 误判为 i 时候的损失。一般可以简写为 λ_{ij}。

显然，对于普遍的贝叶斯决策论，后验概率 $P(w_j|\mathbf{x})$ 可以通过贝叶斯公式得到：

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|\mathbf{w_j})P(w_j 0)}{p(\mathbf{x})} \tag{3}$$

此时的证据因子为：

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j) \tag{4}$$

与行为 α_i（比如把输入判定为 i 类）相关联的损失就是：

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|\mathbf{x}) \\ &= \sum_{j=1}^{c} \lambda_{ij}P(w_j|\mathbf{x}) \end{aligned} \tag{5}$$

值得注意的是，**条件风险** $R(\alpha_i|\mathbf{x})$ 是和后验概率有关的，一般后验概率大的，自然条件风险就会小。

3.2 回到二分类

二分类时，条件风险函数 $R(\alpha_i|\mathbf{x})$ 可以分别写为：

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(w_1|\mathbf{x}) + \lambda_{12}P(w_2|\mathbf{x}) \tag{6}$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(w_1|\mathbf{x}) + \lambda_{22}P(w_2|\mathbf{x}) \tag{7}$$

风险越小的选择越好，因此带入方程，判断为 w_1 时，会有：

$$(\lambda_{21} - \lambda_{11})P(w_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(w_2|\mathbf{x}) \tag{8}$$

带入贝叶斯公式可以得到：

$$\frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(w_2)}{P(w_1)} \tag{9}$$

显然，假如判断错误时 λ 为 1，正确时为 0，则 (9) 式就是简单贝叶斯里的公式。而此时条件风险函数：

$$\begin{aligned} R(\alpha_i) &= \sum_{j=1}^{c} \lambda(\alpha_i|w_j)P(w_j|\mathbf{x}) \\ &= \sum_{j \neq i} P(w_j|\mathbf{x}) \\ &= 1 - P(w_j|\mathbf{x}) \end{aligned} \tag{10}$$

这很符合常理。

4 分类器、判别函数及判定面

分类器是多种多样的，贝叶斯判别只是其中一种。有非常多的方法来表述模式分类器，其中用的最多的是一种判别函数 $g_i(\mathbf{x})$ 的形式，如果对于所有的 $j \neq i$，有

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \tag{11}$$

则分类器就会把这个特征向量 $\mathbf{x}$ 判断为 $w(i)$。

4.1 贝叶斯分类器

贝叶斯分类器就可以简单自然的表示称这种方式。我们的 $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$（这是由于最大判别函数和最小条件风险是相对应的）。在最小误差概率的情况下，可以进一步简化问题，让 $g_i(\mathbf{x}) = P(\alpha_i|\mathbf{x})$，此时的最大判别函数就和最大后验概率相对应了。

显然，判别函数只需要保证各个类别的大小次序不变即可（递增），在此基础上可以随意变化：

$$g_i(\mathbf{x}) = \frac{p(\mathbf{x}|w_i)P(w_i)}{\sum_{j=1}^{c} p(\mathbf{x}|w_j)P(w_j)} \tag{12}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|w_i)P(w_i) \tag{13}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|w_i) + \ln P(w_i) \tag{14}$$

尽管根据参数不同，判别函数可以写成不同的形式，但是其判决规则是相通的。每种判决规则均是将特定空间分成 c 个判决区域（对应 c 个类别），$\mathscr{R}_1, \ldots \mathscr{R}_c$

4.2 二类情况下的贝叶斯分类器

则种分类器有一个专门的名字——二分分类器。此时并非使用两个判别函数 g_1 和 g_2，取而代之的是定义一个简单的判别函数:

$$g(\mathbf{x}) \equiv g_1\mathbf{x} - g_2(\mathbf{x}) \tag{15}$$

若上式大于 0，则判断为 w_1, 否则为 w_2。

二分分类器可以看成**一个简单的判别函数** $g(\mathbf{x})$ 并**根据结果的符号进行分类**的机器。

$$g(\mathbf{x}) = P(w_1|\mathbf{x}) - P(w_2|\mathbf{x}) \tag{16}$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|w_1)}{p(\mathbf{x}|w_2)} + \ln \frac{P(w_1)}{P(w_2)} \tag{17}$$

5 正态分布的贝叶斯分类器

这一节终于要开始真正的计算了，上面全是符号，都快傻了都。。。计算之前还是要了解一些数学概念。

贝叶斯分布的判决函数 $g_i(\mathbf{x})$ 的形式已经如 (12)(13)(14) 所示清楚的表达出来了，那么剩下的就是只要知道先验分布 $P(w_i)$（作为先验概率，这个一般是已知的，或者显而易见的），和似然函数 $p(\mathbf{x}|w_i)$ 即可以进行求解。而正态分布概率密度函数是最为常见的一种函数（中心极限定理）。

5.1 正态分布

一维正态分布密度函数

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{18}$$

μ 为变量 x 的均值，σ 为变量 x 的方差。

多维整体分布密度函数

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{C}^{-1}(\mathbf{x}-\mu)} \tag{19}$$

其中，每一类模式的分布密度都完全被其均值向量 m_i 和协方差矩阵 $\mathbf{C_i}$ 所规定，其定义为：

$$m_i = E_i\{x\} \tag{20}$$

$$C_i = E_i\{(x - m_i)(x - m_i)^T\} \tag{21}$$

$E_i\{x\}$ 表示对类别属于 w_i 的模型的数学期望。（对于不同的种类，其均值向量是不一样的）在上述公式中，d 为模式向量的维数，$|C_i|$ 为矩阵 C_i 的行列式，协方差矩阵 C_i 是对称的正定矩阵，其对角线上的元素 C_{kk} 是模式向量第 k 个元素的方差，非对角线上的元素 C_{jk} 是 x 的第 j 个分量 x_j 和第 k 个分量 x_k 的协方差。当 x_j 和 x_k 统计独立时，$C_{jk} = 0$。当协方差矩阵的全部非对角线上的元素都为零时，多变量正态类密度函数可简化为 d 个单变量正态类密度函数的乘积。

5.2 一些基本概念

均值向量、方差、协方差、正定矩阵、半正定矩阵、行列式、二次型

- 样本的均值是样本数据集的零维表达。
  均值向量：对于不同的类别 w_i，其特征空间是不同的，特征空间中的特征向量也自然不相同。因此每一个类别都有自己的该均值向量可以表示为：

$$\mu = \mathscr{E}[\mathbf{x}] \tag{22}$$

- 方差是协方差的一种，协方差适用于统计两个变量的总体误差。而当两个变量相同时，就变成了方差。协方差矩阵中第 i 行第 j 列的定义是：

$$\sigma_{ij} = \mathscr{E}[(\mathbf{x_i} - \mu_\mathbf{i})(\mathbf{x_j} - \mu_\mathbf{j})] \tag{23}$$

- 二次型、正定、半正定 请参考此处 。信息量很大很有用!!!
- 行列式：行列式（Determinant）是数学中的一个函数，将一个 $n \times n$ 的矩阵 A 映射到一个标量，记作 $\det(\mathbf{A})$ 或 $|\mathbf{A}|$。行列式可以看做是有向面积或体积的概念在一般的欧几里得空间中的推广。或者说，在 n 维欧几里得空间中，行列式描述的是一个线性变换对“体积”所造成的影响。

- 概率密度函数的内积:

5.3 正态分布下，贝叶斯决策论的判别函数

已知类别 w_i 的判别函数可写成如下形式:

$$d_i(\mathbf{x}) = p(\mathbf{x}|w_i)P(w_i), \qquad i = 1, 2, \ldots, M \tag{24}$$

对于正太密度函数，按照公式 (11) 的形式，可以继续写成:

$$\begin{aligned} d_i(\mathbf{x}) &= \ln\left[p(\mathbf{x}|w_i)\right] + \ln P(w_i), \qquad i = 1, 2, \ldots, M \\ &= \ln P(w_i) - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|C_i| - \frac{1}{2}(\mathbf{x} - m_i)^T C_i^{-1}(\mathbf{x} - m_i), \qquad i = 1, 2, \ldots, M \end{aligned} \tag{25}$$

可以去掉和 i 无关的项，继续简化为:

$$\begin{aligned} d_i(\mathbf{x}) &= \ln\left[p(\mathbf{x}|w_i)\right] + \ln P(w_i), \qquad i = 1, 2, \ldots, M \\ &= \ln P(w_i) - \frac{1}{2}\ln|C_i| - \frac{1}{2}(\mathbf{x} - m_i)^T C_i^{-1}(\mathbf{x} - m_i), \qquad i = 1, 2, \ldots, M \end{aligned} \tag{26}$$

公式 (26) 即位正态分布模式的贝叶斯判别函数。其中 $P(w_i)$ 是先验概率，是已经知道的。$|C_i|$ 是行列式，是一个标量，也是已知的。那么显然该方程是一个二次型方程 ($\mathbf{X^T AX}$)，即判别函数是一个超二次曲面。（见二次型、正定、半正定的参考）

协方差矩阵 $\mathbf{C}$ 一般是对称的且半正定的，我们将严格限定 $\mathbf{C}$ 是正定的，是的 $\mathbf{C}$ 的行列式是一个正数。协方差矩阵的对角线元素就是 x_i 的方差，非对角线元素是 x_i 和 x_j 的协方差。假如 x_i 和 x_j 统计独立，则 $\sigma_{ij} = 0$，如果所有的肺对角线元素都为 0，则公式 (19) 中 $p(\mathbf{x})$ 变成了 $\mathbf{x}$ 中各个元素的单变量正态密度函数的内积。

5.4 两类问题且其类模式都是正态分布的特殊情况

注意，这里只是确定了解的形式，但是并没有确定如何求的解。只有在确定了类条件概率密度的参数的时候，才能求得真正的解。

1. $\mathbf{C_i \neq C_2}$

   此时 $p(\mathbf{x}|w_1)$ 可以表示为 $\mathcal{N}(\mathbf{m_1}, \mathbf{C_1})$，$p(\mathbf{x}|w_2)$ 可以表示为 $\mathcal{N}(\mathbf{m_2}, \mathbf{C_2})$，对应的判别函数带入 (26) 式，可以得到 $d_1(x) - d_2(x)$ 是一个关于 $\mathbf{x}$ 的二次型方程，其关于 0 的解是一个超二次曲面。即 w_1 和 w_2 两类模式可用二次判别界面分开。

   当 x 是二维模式时，判别界面为二次曲线，如椭圆，圆，抛物线或双曲线等。

2. $\mathbf{C_i} = \mathbf{C_2}$ 带入公式 (26) 可得：

$$\begin{aligned} d_1(\mathbf{x}) - d_2(\mathbf{x}) &= \ln P(w_1) - \ln P(w_2) + (m_1 - m_2)^T C^{-1}\mathbf{x} - \frac{1}{2}m_1^T C^{-1} m_1 + \frac{1}{2}m_2^T C^{-1} m_2 \\ &= 0 \end{aligned} \tag{27}$$

判别界面为 $\mathbf{x}$ 的线性函数，为一超平面。

当 x 是二维时，判别界面为一直线。

6 均值向量和协方差矩阵的参数估计

在贝叶斯分类器中，构造分类器需要知道类概率密度函数 $p(x|w_i)$。如果按先验知识**已知其分布**，则只需知道**分布的参数**即可。

- 例如：类概率密度是正态分布，它完全由其均值向量和协方差矩阵所确定。（第五章节中已经给出了正态分布情况下判决函数的最终表达形式，只需要知道参数，就真正解完了。）

对均值向量和协方差矩阵的估计即为贝叶斯分类器中的一种参数估计问题。估计的方法一般有两种：

- 一种是将参数作为非随机变量来处理，例如矩估计就是一种非随机参数的估计。
- 另一种是随机参数的估计，即把这些参数看成是随机变量，例如贝叶斯参数估计。

6.1 均值和协方差矩阵的非随机参数的估计

此类算法以样本的平均值作为均值向量的近似值，从而计算得到均值向量和协方差矩阵。

待估计量的定义：均值向量、协方差矩阵

使用样本的平均值作为均值向量的近似值，那么均值估计量 $\hat{m}$ 为：

$$\hat{m} = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{28}$$

其中 N 为样本的数目。

那么此时的协方差矩阵为：

$$\mathbf{C} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1d} \\ c_{21} & c_{22} & \cdots & c_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ c_{d1} & c_{d2} & \cdots & c_{dd} \end{pmatrix} \tag{29}$$

其中的每一个元素 c_{ij}，是第 i 维和第 j 维的协方差。其公式定义为：

$$\begin{aligned} c_{ij} &= conv(x_i, x_j) = E\left\{(x_i - m_i)(x_j - m_j)^T\right\} \\ &= \frac{1}{N-1}\sum_{k=1}^{N}(x_{ik} - m_i)(x_{jk} - m_j)^T \end{aligned} \tag{30}$$

其中 x_i 和 x_j 指的是特征空间 $\mathbf{X}$ 上的第 i 维和第 j 维，是一个随机变量（连续）。N 是样本的数量（每一维上自然是一样的）。除以 $N-1$ 是为了无偏估计。协方差矩阵是一个对称矩阵，且是一个半正定矩阵，主对角线是各个随机变量的方差（各个维度上的方差）。

协方差矩阵原本的数值为：

$$\begin{aligned} \mathbf{C} &= E\left\{(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T\right\} \\ &= E\left\{\mathbf{x}\mathbf{x}^T - \mathbf{x}m^T - m\mathbf{x}^T + mm^T\right\} \\ &= E\{\mathbf{x}\mathbf{x}^T\} - m^T E\{\mathbf{x}\} - m^T E\{\mathbf{x}\} + mm^T \\ &= E[\mathbf{x}\mathbf{x}^T] - mm^T \end{aligned} \tag{31}$$

根据上式，在离散情况下的的协方差矩阵的估计可以写为：

$$\begin{aligned} \widehat{\mathbf{C}} &\approx \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{m})(x_k - \hat{m})^T \\ &= \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T - \hat{m}\hat{m}^T \end{aligned} \tag{32}$$

注意，输入样本模式的总体为 $\{x_1, x_2, \ldots, x_k \ldots, x_N\}$。因为估计量时没有真实的均值向量 $\mathbf{m}$ 可以用，只能用均值向量的估计量 $\hat{m}$ 代替。x_i 是第 i 个样本，它是一个 k 维向量（有 k 个特征）。

待估计量的迭代运算

假设 $\hat{m} = \frac{1}{N}\sum_{j=1}^{N} x_j$，那么对于第 $N+1$ 个样本加入后，均值向量为：

$$\hat{m}(N+1) = \frac{1}{N+1}\sum_{j=1}^{N+1} x_j = \frac{1}{N+1}\left[N\hat{m}(N) + x_{N+1}\right] \tag{33}$$

取初始状态 $\hat{m} = x_1$，即可迭代求出 $\hat{m}(N)$。

协方差矩阵的运算，输入 N 个样本时的协方差矩阵为：

$$\begin{aligned}\widehat{\mathbf{C}}(N) &\approx \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{m})(x_k - \hat{m})^T \\ &= \frac{1}{N}\sum_{k=1}^{N} x_k x_k^T - \hat{m}\hat{m}^T\end{aligned} \tag{34}$$

$N = N + 1$ 时，有：

$$\begin{aligned}\widehat{\mathbf{C}}(N+1) &= \frac{1}{N+1}\sum_{k=1}^{N+1} x_k x_k^T - \hat{m}(N+1)\hat{m}^T(N+1) \\ &= \frac{1}{N+1}\left[\sum_{k=1}^{N} x_k x_k^T + x_{N+1}x_{N+1}^T\right] - \hat{m}(N+1)\hat{m}^T(N+1)\end{aligned} \tag{35}$$

带入 (34) 方程可以得到：

$$\begin{aligned}\widehat{\mathbf{C}}(N+1) =& \frac{1}{N+1}\left[N\hat{C}(N) + N\hat{m}(N)\hat{m}^T(N) + x_{N+1}x_{N+1}^T\right] - \\ & \frac{1}{(N+1)^2}\left[N\hat{m}(N) + x_{N+1}\right]\left[N\hat{m}(N) + x_{N+1}\right]^T\end{aligned} \tag{36}$$

$\hat{C}(1) = x_1 x_1^T - \hat{m}(1)\hat{m}^T(1)$ 且 $\hat{m}(1) = x_1$，因此 $\hat{C}(1) = 0$ 为零矩阵。

(31) 推导使用了协方差公式的性质，遇到这种数学推导，先去网上搜索一下！自己瞎推太浪费时间了！ 。

6.2 均值向量和协方差矩阵的贝叶斯学习：随机参数估计

此方法是把参数 θ 作为一个参数，通过训练模式样本集 $\{x_i\}$，利用贝叶斯公式设计一个迭代运算过程求出参数的后验概率密度 $p(\theta|x_i)$。

单变量正态密度函数的均值学习

略

一般概念

略