第2章 生成式分类器

卿来云
lyqing@ucas.ac.cn

大纲

- 贝叶斯最优分类器
- 生成式分类器
- 概率密度函数的参数估计
- 朴素贝叶斯分类器
- 高斯判别分析

大纲

- 贝叶斯最优分类器
  - 错误率、最小错误率决策
  - 损失函数、条件风险、期望风险、最小期望风险决策
- 生成式分类器
- 概率密度函数的参数估计
- 朴素贝叶斯分类器
- 高斯判别分析

模式分类

- 模式分类：给定某个给定的模式样本，确定其所属类别
- 通过测量被识别对象的某些特征值，并将其作为某一个判决规则的输入，按此规则来对样本进行分类

- 例：鸢尾花分类
  - 给定一朵鸢尾花的4个特征值（花萼长度 、花萼宽度 、花瓣长度、花瓣宽度）
  - 确定该朵花属于哪一类鸢尾花（山鸢尾、变色鸢尾、维吉尼亚鸢尾）

数据示例

- 鸢尾花数据集：$\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
  - 150个样本： $N = 15$
  - 每个样本有$D = 4$维特征$\boldsymbol{x}_i$：花瓣的长度和宽度、花萼的长度和宽度
  - 标签y_i：共3类样本

特征　　目标/标签列

sepal length	sepal width	petal length	petal width	species
6.7	3.0	5.2	2.3	virginica
6.4	2.8	5.6	2.1	virginica
4.6	3.4	1.4	0.3	setosa
6.9	3.1	4.9	1.5	versicolor
4.4	2.9	1.4	0.2	setosa
4.8	3.0	1.4	0.1	setosa
5.9	3.0	5.1	1.8	virginica
5.4	3.9	1.3	0.4	setosa
4.9	3.0	1.4	0.2	setosa
5.4	3.4	1.7	0.2	setosa

特征值　样本　标签

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ \cdots & \cdots & \cdots & \cdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix}$$

确定性分类

- 确定性现象：在获取模式的观测值时，有些事物具有确定的因果关系，即在一定的条件下，它必然会发生或必然不发生。

- 例如：识别一块模板是不是直角三角形，只要凭“三条直线边闭合连线和一个直角”这个规则，测量它是否有三条直线边的闭合连线并有一个直角，就完全可以确定它是不是直角三角形。

非确定性分类

- 但在现实世界中，有许多客观现象的发生。就每一次观测来说，即使在基本条件保持不变的情况下也具有不确定性。
- 只有在大量重复的观察下，其结果才能呈现出某种规律性，即对它们观察到的特征具有统计特性。
- 特征的值不再是一个确定的向量，而是一个随机向量。
- 此时，只能利用模式集合的统计特性来分类，以使分类器发生错误的概率最小。

最小化错误率决策

- 我们希望做出的决策是**平均错误率/平均误差概率**最小的。
- 平均错误率： $P(error) = \int P(error|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$
- 如果对于每个样本$\boldsymbol{x}$，保证 $P(error|\boldsymbol{x})$最小，则平均错误率就最小。
- 给定观测值$\boldsymbol{x}$，判断其类别为c的错误率是

$$P(error|\boldsymbol{x}) = 1 - P(Y = c|\boldsymbol{x})$$

- 要是错误率最小，则$P(Y = c|\boldsymbol{x})$的概率最大。
- 所以最小错误率决策等价于最大后验概率决策。

最大后验概率决策

■最小错误率决等价于最大后验概率决策：$\hat{y} = \underset{c}{\operatorname{argmax}} P(Y = c|\boldsymbol{x})$

■根据贝叶斯规则，后验概率

$$P(Y = c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|Y = c)P(Y = c)}{\sum_{c'} P(\boldsymbol{x}|Y = c')P(Y = c')}$$

$$\propto P(\boldsymbol{x}|Y = c)P(Y = c)$$

条件概率：

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

全概率公式：

$$P(B) = \sum_i P(B|A_i)P(A_i)$$

$P(Y = c)$：类别c的先验概率

$P(\boldsymbol{x}|Y = c)$：类条件概率；当$\boldsymbol{x}$为连续值，用概率密度函数$p(\boldsymbol{x}|Y = c)$替代

■判别规则：$\hat{y} = \underset{c}{\operatorname{argmax}} P(\boldsymbol{x}|Y = c)P(Y = c)$

判别函数

- 类别c的判别函数可写成：$d_c = P(Y = c|\boldsymbol{x}) = P(\boldsymbol{x}|Y = c)P(Y = c)$
- 通常取自然对数的形式以方便计算，则有

$$d_c = \ln\big(P(\boldsymbol{x}|Y = c)\big) + \ln\big(P(Y = c)\big)$$

例：最大后验概率决策

■例：地震预测　$Y=1$：地震

$Y=0$：正常

- 对某一地震高发区进行统计：每周发生地震的概率为20%，即

$$P(Y=1)=0.2$$

$$P(Y=0)=1-P(Y=1)=0.8$$

- 因为$P(Y=0)>P(Y=1)$，说明在没有其他证据的情况下，正常的可能性大。

- 通常地震与生物异常反应之间有一定的联系。
- 若用生物是否有异常反应这一观察现象来对地震进行预测，生物是否异常这一结果以特征X代表，这里X为1维特征，且只有$X=1$（异常）和$X=0$（正常）2种取值

例：最大后验概率决策

■例：地震预测

- $P(Y=1)=0.2$ ，$P(Y=0)=0.8$
- X为二值特征：$X=1$（异常）、$X=0$（正常）

- 假设根据观测记录，有如下统计结果：
- 地震前一周内出现生物异常反应的概率为0.6，即

$$P(X=1|Y=1)=0.6$$
$$P(X=0|Y=1)=0.4$$

- 一周内没有发生地震但也出现了生物异常的概率为

$$P(X=1|Y=0)=0.1$$
$$P(X=0|Y=0)=0.9$$

例：最大后验概率决策

■ 例：地震预测

- 先验概率：$P(Y=1)=0.2$ ，$P(Y=0)=0.8$
- 类条件概率： $P(X=1|Y=1)=0.6$，$P(X=0|Y=1)=0.4$
- $P(X=1|Y=0)=0.1$，$P(X=0|Y=0)=0.9$

- 若某日观察到明显的生物异常反应现象，一周内发生地震的概率为多少，即求$P(Y=1|X=1)$（后验概率）

$$P(Y=1|X=1)=\frac{P(X=1|Y=1)P(Y=1)}{P(X=1|Y=1)P(Y=1)+P(X=1|Y=0)P(Y=0)}$$

$$=\frac{0.6\times0.2}{0.6\times0.2+0.1\times0.8}=0.6>0.5$$

- 或

$P(X=1|Y=1)P(Y=1)=0.6\times0.2=0.12$，
$P(X=1|Y=0)P(Y=0)=0.1\times0.8=0.08$

两种方式都会判断地震的概率更大，即$\hat{y}=1$

最小化风险决策

- 最小化错误率决策没有考虑不同方式错误带来的损失可能不同
  - 如在癌症筛查中，将一个非癌症人员被判断为癌症患者，带来的损失是一些额外的进一步检查，有损失但相对较小；如果将一个癌症患者误判为正常人员，则带来的损失是错失最佳治疗时机，损失较大。

- 决策时应该引入损失函数或代价函数，描述每个决策所付出的代价的大小。

损失函数

- 样本$\boldsymbol{x}$的真实类别记作 $y \in \{1, 2, \ldots, C\}$
- 分类器的输出类别记为$\hat{y}(\boldsymbol{x}) \in \{1, 2, \ldots, C\}$
- 记损失函数（loss function）为$L(y, \hat{y})$，表示将本应属于类别y的模式判别成属于类别$\hat{y}$的代价。

- 例如：0-1损失函数：$L(\hat{y}, y) = \begin{cases} 0 & y = \hat{y} \\ 1 & y \neq \hat{y} \end{cases}$
- 亦写成代价矩阵的形式

$\hat{y}$ \ y	0	1
0	L_{00}	L_{01}
1	L_{10}	L_{10}

期望风险（expected risk）

- 定义期望风险为平均损失：

$$R_{\exp}(\hat{y}(\boldsymbol{x})) = \int L(\hat{y}(\boldsymbol{x}), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x} dy$$

$$= \int \left(\int L(\hat{y}(\boldsymbol{x}), y)\, p(y|\boldsymbol{x}) dy \right) p(\boldsymbol{x}) d\boldsymbol{x} \qquad p(\boldsymbol{x}, y) = p(y|\boldsymbol{x}) p(\boldsymbol{x})$$

$$= \int \left(\underbrace{\sum_{y=1}^{C} L(\hat{y}(\boldsymbol{x}), y)\, P(Y = c|\boldsymbol{x})}_{R(\hat{y}(\boldsymbol{x})|\boldsymbol{x})} \right) p(\boldsymbol{x}) d\boldsymbol{x} \qquad \text{离散积分为求和}$$

$$= \int R(\hat{y}(\boldsymbol{x})|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

- 将样本$\boldsymbol{x}$分类为类别c的条件风险定义为：

$$R(\hat{y}(\boldsymbol{x}) = c|\boldsymbol{x}) = \sum_{y=1}^{C} L(c, y) P(Y = y|\boldsymbol{x}) = \sum_{y=1}^{C} L_{cy} P(Y = y|\boldsymbol{x})$$

损失加权的后验概率之和

最小风险决策

- 期望风险： $R_{\exp}(\hat{y}(\boldsymbol{x})) = \mathbb{E}[L(\hat{y}(\boldsymbol{x}), y)] = \mathbb{E}_{\boldsymbol{x}}[R(\hat{y}(\boldsymbol{x})|\boldsymbol{x})]$

- 对比期望风险和平均错误率，

$$R_{\exp}(\hat{y}(\boldsymbol{x})) = \int R(\hat{y}(\boldsymbol{x})|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

$$P(error) = \int P(error|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$

- 条件风险和错误率的作用相同，条件风险是错误率的推广。
- 选择对每个样本条件风险最小的分类规则$\hat{y}(\boldsymbol{x})$，将使期望风险最小化。
- 最小风险决策的决策规则为： $\hat{y} = \underset{c}{\operatorname{argmin}} R(c|\boldsymbol{x})$

最小风险决策 vs. 最小错误率决策

- 当损失函数为0-1损失时，$L(\hat{y}, y) = \begin{cases} 0 & y = \hat{y} \\ 1 & y \neq \hat{y} \end{cases}$
- 条件风险为：

$$R(c|\boldsymbol{x}) = \sum_{y=1}^{C} L_{cy} P(Y = y|\boldsymbol{x}) = \sum_{y \neq c} P(Y = y|\boldsymbol{x}) = 1 - P(Y = c|\boldsymbol{x})$$

- 等于错误率 $P(error|\boldsymbol{x}) = 1 - P(Y = c|\boldsymbol{x})$
- 此时最小风险决策等价于最小无错率决策。
- 但一般情况下，最小风险决策用途更广。

例：选课

- 课程分为2种：Y

课程质量	好（$Y=1$）	差（$Y=0$）
概率（先验）	0.6	0.4

- 课堂是否有趣：X

- 条件概率

$P(X\|Y)$	课程质量好（$Y=1$）	课程质量差（$Y=0$）
课堂有趣（$X=1$）	0.8	0.1
课堂无聊（$X=0$）	0.2	0.9

例：选课

■损失函数

$L(y,\hat{y})$	课程质量好（$Y=1$）	课程质量差（$Y=0$）
选课（$\hat{Y}=1$）	0	10
退课（$\hat{Y}=0$）	20	0

■听了一次课后，发现课堂有趣，则最小风险决策？

$$R(c|\boldsymbol{x}) = \sum_{y=1}^{C} L_{cy} P(Y=y|\boldsymbol{x})$$

■课堂有趣，课程质量后验

课程质量	好（$Y=1$）	差（$Y=0$）
概率（先验）	0.6	0.4

$$P(Y=1|X=1)$$
$$=\frac{P(X=1|Y=1)P(Y=1)}{P(X=1|Y=1)P(Y=1)+P(X=1|Y=0)P(Y=0)}$$
$$=\frac{0.8\times0.6}{0.8\times0.6+0.1\times0.4}=0.92$$

$$P(Y=0|X=1)=\frac{0.1\times0.4}{0.8\times0.6+0.1\times0.4}=0.08$$

$P(X\|Y)$	课程质量好（$Y=1$）	课程质量差（$Y=0$）
课堂有趣（$X=1$）	0.8	0.1
课堂无聊（$X=0$）	0.2	0.9

■课堂有趣，课程质量后验

$$P(Y=1|X=1)=0.92$$

$$P(Y=0|X=1)=0.08$$

$L(y,\hat{y})$	课程质量好（$Y=1$）	课程质量差（$Y=0$）
选课（$\hat{Y}=1$）	0	10
退课（$\hat{Y}=0$）	20	0

■选课的风险

$$R(c|\boldsymbol{x})=\sum_{y=1}^{C}L_{cy}P(Y=y|\boldsymbol{x})=0.92\times0+0.08\times10=0.8$$

■退课的风险

$$R(c|\boldsymbol{x})=\sum_{y=1}^{C}L_{cy}P(Y=y|\boldsymbol{x})=0.92\times20+0.08\times0=18.4$$

引入拒识的决策

- 在必要情况下，分类器对于某些样本可以拒绝给出一个输出结果（例如转交给人工处理）。
  - 如后验概率接近的情况下，拒绝做判决
- 拒识（reject）：分类器可以拒绝将样本判为C个类别中的任何一类。
- 此时损失的定义为：$$L(y,\hat{y})=\begin{cases}0 & y=\hat{y}\\ L_s & y\neq\hat{y}\\ L_r & reject\end{cases}$$
- 拒识代价L_r必须小于错分代价L_s，否则永远不会对样本拒识。

引入拒识的决策

- 引入拒识（reject）的损失函数：$L(\hat{y}, y) = \begin{cases} 0 & y = \hat{y} \\ L_s & y \neq \hat{y} \\ L_r & reject \end{cases}$

- 此时条件风险为：$R(c|\boldsymbol{x}) = \begin{cases} L_s(1 - P(Y = c|\boldsymbol{x})) & c = 1,2,\dots,C \\ L_r & reject \end{cases}$

- 所以在引入拒识的情况下，最小风险决策为

$$\underset{c}{\operatorname{argmin}} R(c|\boldsymbol{x}) = \underset{c}{\operatorname{argmin}} \begin{cases} \underset{c}{\operatorname{argmax}} P(Y = c|\boldsymbol{x}) & \text{if } \max_c P(Y = c|\boldsymbol{x}) > 1 - \dfrac{L_r}{L_s} \\ reject & otherwise \end{cases}$$

贝叶斯最优分类器

- 贝叶斯最优分类器：最小风险决策所决定出的贝叶斯分类器。
- 相应地，风险被称为贝叶斯风险。
- 给定损失函数时，计算贝叶斯最优分类器的关键是计算后验概率

$$\underset{c}{\operatorname{argmin}} R(c|\boldsymbol{x}) = \underset{c}{\operatorname{argmin}} \begin{cases} \underset{c}{\operatorname{argmax}} P(Y = c|\boldsymbol{x}) & \text{if } \max_c P(Y = c|\boldsymbol{x}) > 1 - \dfrac{L_r}{L_s} \\ reject & otherwise \end{cases}$$

生成式分类器

- 计算贝叶斯最优分类器的关键是计算后验概率

$$\underset{c}{\operatorname{argmin}} R(c|\boldsymbol{x}) = \underset{c}{\operatorname{argmin}} \begin{cases} \underset{c}{\operatorname{argmax}} P(Y=c|\boldsymbol{x}) & \text{if } \max_c P(Y=c|\boldsymbol{x}) > 1 - \dfrac{L_r}{L_s} \\ reject & otherwise \end{cases}$$

- 一种后验概率计算方式为： $P(Y=c|\boldsymbol{x}) = \dfrac{P(\boldsymbol{x}|Y=c)P(Y=c)}{\sum_{c'} P(\boldsymbol{x}|Y=c')P(Y=c')}$
- 需要已知先验概率$p(y)$和类条件概率$p(\boldsymbol{x}|y)$。
- 此时亦被称为生成式分类器。因为已知$p(y), p(\boldsymbol{x}|y)$可得到联合分布$p(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p(y)$，从而可以从联合分布通过采样生成数据$(\boldsymbol{x}_i, y_i)$。
- 另一种方式是判别式分类器：直接计算后验概率或者判别函数。

生成式分类器

- 生成式分类器中，后验概率为： $P(Y=c|\boldsymbol{x})=\frac{P(\boldsymbol{x}|Y=c)P(Y=c)}{\sum_{c'}P(\boldsymbol{x}|Y=c')P(Y=c')}$
- 需要已知概率（密度）函数$p(y)$， $p(\boldsymbol{x}|y)$。
- 实际中，估计概率密度函数很困难。尤其是类条件概率$p(\boldsymbol{x}|y=c)$，因为$\boldsymbol{x}$通常是高维随机向量。

- 类先验概率$P(Y=c)$
- 类条件概率 $p(\boldsymbol{x}|Y=c)$：由于$\boldsymbol{x}$为多维向量，条件分布$p(\boldsymbol{x}|Y=c)$建模困难，可根据数据的实际情况对其做适当假设或简化
- 朴素贝叶斯： 在给定$Y=c$的情况下， $\boldsymbol{x}$的各维独立
- 高斯判别分析：在给定$Y=c$的情况下， $\boldsymbol{x}$为多元高斯分布

大纲

- 贝叶斯最优分类器
- 生成式分类器
- 概率密度函数的参数估计
  - 极大似然估计
  - 贝叶斯估计
  - 常见分布的参数估计
- 朴素贝叶斯分类器
- 高斯判别分析

概率密度参数估计

- 给定随机变量X或随机向量$\boldsymbol{X}$的概率密度函数$p(\boldsymbol{x})$的形式，但其参数未知
  - 例如，在两类分类任务中，类先验分布$Y \sim \text{Berloulli}(\theta)$，但参数$\theta$的值未知

- 有多种方法可用来估计模型的参数
  - 矩方法
  - 极大似然估计：频率学派
  - 贝叶斯方法：贝叶斯学派

极大似然估计

- 令$\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$为来自分布$p(\boldsymbol{x}|\boldsymbol{\theta})$的独立同分布（Identical Independent Distribution, IID）的样本，
- 定义似然函数定义为

$$L(\boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

- 似然函数在数值上是数据$\mathcal{D}$的联合密度，
- 但它是参数$\boldsymbol{\theta}$的函数，不满足密度函数的性质（如对$\boldsymbol{\theta}$的积分不必为1）。

极大似然估计

- 极大似然估计（Maximize Likelihood Estimation, MLE）是使得似然函数$p(\mathcal{D}|\boldsymbol{\theta})$最大的$\boldsymbol{\theta}$，即

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D}|\boldsymbol{\theta})$$

- log似然函数定义为似然函数的自然对数：$l(\boldsymbol{\theta}) = \ln\big(L(\boldsymbol{\theta})\big)$
  - 自然对数函数为单调增函数，所以$l(\boldsymbol{\theta})$和似然函数$L(\boldsymbol{\theta})$在相同的位置取极大值
  - 数值计算更稳定：似然函数$L(\boldsymbol{\theta})$涉及多个小的概率值相乘，容易下溢出
  - 计算更简单：很多概率密度函数是指数函数，取对数运算后更简单

- 在不引起混淆的情况下，有时记log似然函数为似然函数

贝叶斯估计

- MLE认为参数只是一个值（点估计）
- 贝叶斯估计：参数也是随机变量，亦可用概率分布描述其性质
  - 先验分布$p(\boldsymbol{\theta})$：在没有看到数据之前，参数的分布
    - 先验反映我们对参数取值的信念：通常偏好更简单或更光滑的模型
    - 为计算方便，我们一般采用共轭先验（先验分布与后验分布为同族分布）
  - 似然：同MLE相同，为$p(\mathcal{D}|\boldsymbol{\theta})$
  - 后验分布$p(\boldsymbol{\theta}|\mathcal{D})$ ： 在看到数据$\mathcal{D}$后，对参数分布的更新

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

    - 参数估计不再是一个点估计，而是一个分布（信息更多）

贝叶斯估计

- 参数的贝叶斯后验分布

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})} \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

  - 参数估计不再是一个点估计，而是一个分布（信息更多）

- 亦可用后验分布的均值或众数得到参数$\boldsymbol{\theta}$的点估计

贝努利分布的极大似然估计

- 若随机变量X只有0、1两种取值，则$X \sim \text{Bernoulli}(\theta)$，
  - 其中θ表示随机变量取值为1的概率

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

- log似然函数为

$$l(\theta) = \ln p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \ln p(x_i) = \sum_{i=1}^{N} \ln(\theta^{x_i}(1-\theta)^{1-x_i})$$

$$= \sum_{i=1}^{N} (x_i \ln(\theta) + (1-x_i)\ln(1-\theta))$$

- 求参数θ：令 $\dfrac{\partial l(\theta)}{\partial \theta} = \dfrac{1}{\theta}\underbrace{\sum_{i=1}^{N} x_i}_{N_1} - \dfrac{1}{1-\theta}\underbrace{\sum_{i=1}^{N}(1-x_i)}_{N_0} = 0$

$$\frac{N_1}{\theta} - \frac{N_0}{1-\theta} = 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{N_1}{N_1+N_0} = \frac{N_1}{N}$$

X =1的样本占所有样本的比例

贝努利分布的贝叶斯估计

- $X \sim \text{Bernoulli}(\theta)$
- 其似然为：$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{N_1}(1-\theta)^{N_0}$
  - $N_1 = \sum_{i=1}^{N} x_i$为取值为1的样本的数目，$N_0 = \sum_{i=1}^{N}(1-x_i)$为取值为0的样本的数目
  - $N = N_1 + N_0$
- 该似然对应的共轭先验为Beta分布 ：$\text{Beta}(\theta|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$
- 则后验为： $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

$$\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\propto \theta^{N_1+\alpha-1}(1-\theta)^{N_0+\beta-1}$$

$$= \text{Beta}(\theta|\alpha+N_1,\beta+N_0)$$

贝努利分布的贝叶斯估计

- 类先验的贝叶斯后验估计为：$\mathrm{Beta}(\alpha + N_1, \beta + N_0)$
  - 将超参数α, β加到经验计数N_1和N_0上
  - 伪计数的和 $\alpha + \beta$称为先验的强度（先验的有效样本大小），与样本数$N_1 + N_0$的作用类似

- 参数的点估计可取最大后验估计（后验的众数，MAP）：$\hat{\theta} = \dfrac{\alpha + N_1 - 1}{\alpha + \beta + N - 2}$
  - 当$\alpha = \beta = 1$ 为均匀先验， MAP退化为MLE

- 或者后验的均值： $\hat{\theta} = \dfrac{\alpha + N_1}{\alpha + \beta + N}$
  - 当$\alpha = \beta = 1$时， $\hat{\theta} = \dfrac{N_1 + 1}{N + 2}$ ，称为Laplace平滑。

多项分布的极大似然估计

- 若随机变量X只有$1, 2, \ldots, K$共K种取值，则$X \sim \text{Multinoulli}(\theta)$，例如有$K$个面的骰子
- 多项分布Multinoulli可视为Bernoulli的推广
- log似然函数

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(x_i) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(X_i = k) \ln \theta_k$$

$$\text{Multinoulli}(x|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}(X=k)}$$

其中$\mathbb{I}$为指示函数，当 $X = k$ 成立时为1，否则为0

- 求参数θ_k，需约束条件：$\sum_{k=1}^{K} \theta_k = 1$
- 采用拉格朗日乘子法求极值：$J(\lambda, \boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(X_i = c) \ln(\theta_k) + \lambda\left(1 - \sum_{k=1}^{K} \theta_k\right)$

$$\frac{\partial J(\lambda, \boldsymbol{\theta})}{\partial \theta_k} = \frac{1}{\theta_c}\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{I}(X_i = k) - \lambda = 0$$

$$\frac{\partial J(\lambda, \boldsymbol{\theta})}{\partial \lambda} = 1 - \sum_{k=1}^{K} \theta_k = 0$$

取值为k的样本占所有样本的比例

$$\hat{\theta}_k = \frac{\sum_{i=1}^{N} \mathbb{I}(X_i = k)}{N} = \frac{N_k}{N}$$

多项分布的贝叶斯估计

- $\text{Multinoulli}(x|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{\mathbb{I}(X=k)}$
- 似然为：$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{\mathbb{I}(X_i=k)} = \prod_{k=1}^{K} \theta_k^{N_k}$
  - $N_k = \sum_{i=1}^{N} \mathbb{I}(X_i = k)$为取值为$k$的样本的数目
- 该似然对应的共轭先验为Dirichlet分布 ：$\text{Dia}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$
- 则后验为： $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$

$$\propto \prod_{k=1}^{K} \theta_k^{N_k} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{N_k + \alpha_k - 1}$$

$$= \text{Dia}(\boldsymbol{\theta}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$$

多项分布的贝叶斯估计

- 多项分布的的贝叶斯后验估计为：$\mathrm{Dia}(\boldsymbol{\theta}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$
  - 将超参数α_k加到经验计数N_k上
- 参数的点估计可取最大后验估计（后验的众数，MAP）：$\hat{\theta}_k = \frac{\alpha_c + N_k - 1}{\alpha_0 + N - K}$
  - 其中 $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
  - 当所有的$\alpha_k = 1$（均匀先验），MAP退化为MLE。
- 或者后验的均值：$\hat{\theta}_k = \frac{\alpha_k + N_k}{\alpha_0 + N}$
  - 当所有的$\alpha_k = 1$，得到Laplace平滑：$\hat{\theta}_k = \frac{N_k + 1}{N + K}$

单变量高斯分布的极大似然估计

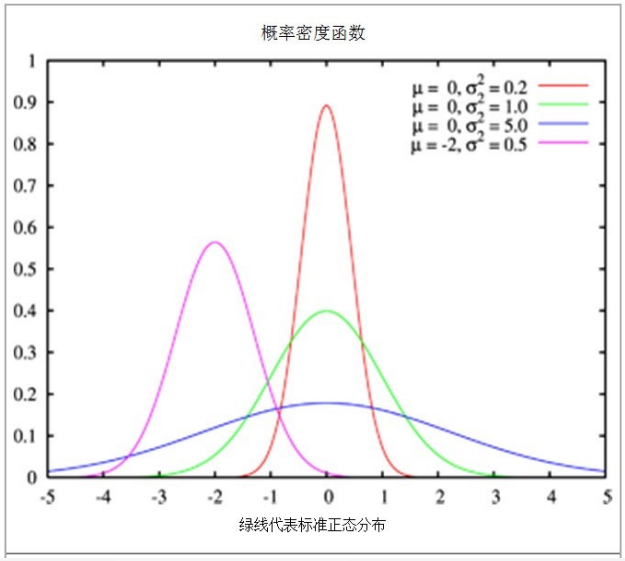
- $X \sim N(\mu, \sigma^2)$：高斯分布或正态分布

$$p(x|\mu,\sigma) = N(\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- log似然函数为

$$\ln p(\mathcal{D}|\mu,\sigma) = \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)$$

$$= -\frac{N}{2}\ln(2\pi) - N\ln(\sigma) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}$$



单变量高斯分布的极大似然估计

- log似然函数为 $\ln p(\mathcal{D}|\mu,\sigma) = -\frac{N}{2}\ln(2\pi) - N\ln(\sigma) - \sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2}$
- 对参数求偏导数：

$$\frac{\partial \ln p(\mathcal{D}|\mu,\sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \quad \Longrightarrow \quad \hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

样本均值

$$\frac{\partial \ln p(\mathcal{D}|\mu,\sigma)}{\partial \sigma} = -\frac{N}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0 \quad \Longrightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$$

$$= \frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{N}$$

样本的经验方差

单变量高斯分布的贝叶斯估计

- 我们在此只讨论参数σ已知，μ的贝叶斯估计， σ的贝叶斯估计请参看相关资料
- 似然函数为

$$p(\mathcal{D}|\mu)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)=\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

$$=a\exp\left(-\frac{1}{2\sigma^2}\left(N\mu^2-2\sum_{i=1}^{N}x_i\,\mu\right)\right)=a\exp\left(-\frac{1}{2N\sigma^2}\left(\mu^2-2\frac{1}{N}\sum_{i=1}^{N}x_i\,\mu\right)\right)$$

$$=a\exp\left(-\frac{1}{2N\sigma^2}(\mu^2-2\bar{x}\mu)\right)=N\left(\bar{x},\frac{\sigma^2}{N}\right)$$

- 其中$\bar{x}=\frac{1}{N}\sum_{i=1}^{N}x_i$
- 共轭先验为$p(\mu)=N(\mu_0,\sigma_0^2)$

单变量高斯分布的贝叶斯估计

- 贝叶斯后验为 $p(\mu|\mathcal{D}) = N(\hat{\mu}_N, \hat{\sigma}_N^2)$

$$\hat{\sigma}_N^2 = \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \qquad \hat{\mu}_N = \hat{\sigma}_N^2 \left(\frac{N}{\sigma^2} \bar{x} + \frac{\mu_0}{\sigma_0^2} \right) = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \bar{x} + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

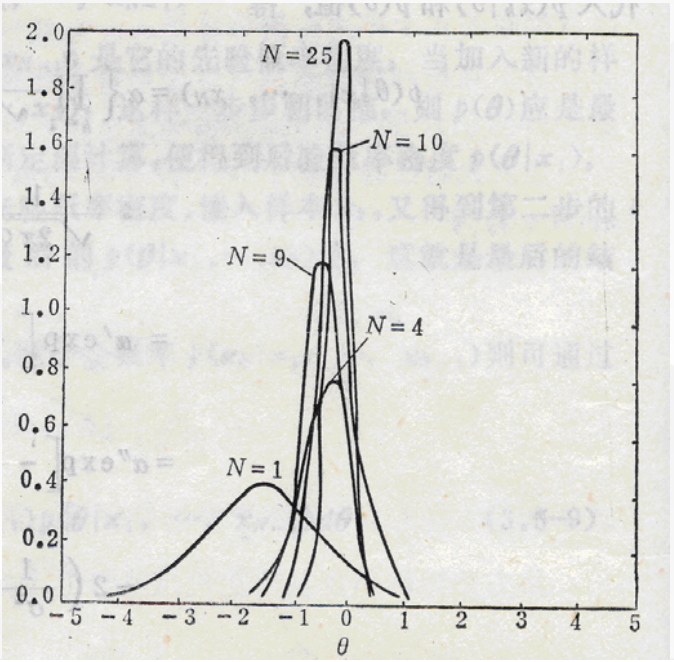
- $\hat{\mu}_N$是先验信息（μ_0、σ_0^2、σ^2）与训练样本所给信息（N、$\bar{x}$）的组合，用N个训练样本对均值的先验μ_0进行补充
- $\hat{\sigma}_N^2$是估计$\hat{\mu}_N$的不确定性的度量，σ_N^2随N的增加而减小，当$N \to \infty$ 时，$\hat{\sigma}_N^2 \to 0$。
- $\hat{\mu}_N$是$\bar{x}$（MLE）和μ_0的线性组合，两者的系数非负且其和为1。因此只要$\sigma_0^2 \neq 0$，当$N \to \infty$时，$\hat{\mu}_N \to \bar{x}$。

单变量高斯分布的贝叶斯估计

- 贝叶斯后验为 $p(\mu|\mathcal{D}) = N(\hat{\mu}_N, \hat{\sigma}_N^2)$

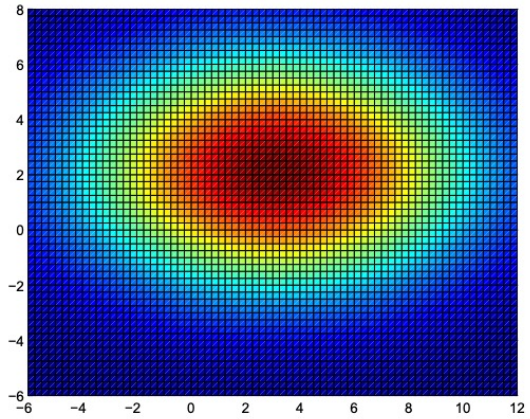
$$\hat{\sigma}_N^2 = \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \qquad \hat{\mu}_N = \sigma_N^2\left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_0}{\sigma_0^2}\right) = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\bar{x} + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_0$$

高斯分布概率密度的均值μ的学习过程：每增加一个样本，$\hat{\mu}_N$估计的不确定性减小，所以$p(\mu|\mathcal{D})$ 的峰变得越来越突起，且其均值$\hat{\mu}_N$与$\bar{x}$之间的偏差的绝对值越来越小。



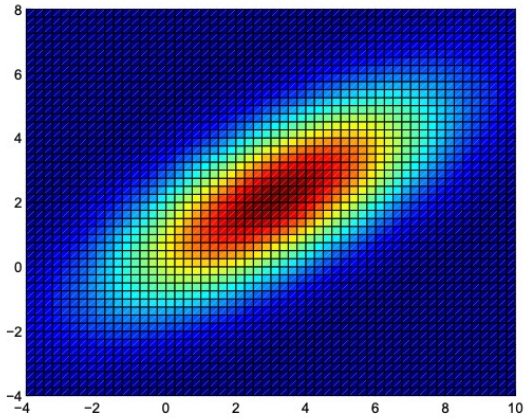
多元高斯分布（正态分布）的概率密度函数

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}),\ p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$$



$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$$



$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 15 & 5 \\ 5 & 5 \end{bmatrix}$$

概率密度函数的参数：

(1) 均值向量 $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x})$

(2) 协方差矩阵

$$\boldsymbol{\Sigma} = \mathbb{E}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\right)$$

对称的正定矩阵

对角线上的元素$\sigma_{k,k}$随机向量$\boldsymbol{x}$第k个元素的方差。

非对角线上的元素$\sigma_{j,k}$是$\boldsymbol{x}$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，$\sigma_{j,k} = 0$。

当协方差矩阵的全部非对角线上的元素均为零时，多元高斯分布的概率密度函数可简化为D个单变量高斯分布的概率密度函数的乘积

多元高斯分布的极大似然估计

■ $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ， $p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right)$

■ log似然函数：

$$\ln p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \ln\left(p(\boldsymbol{x}_i)\right)$$

$$= \sum_{i=1}^{N} \ln\left(\frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_c)\right)\right)$$

$$= -\frac{N \times D}{2} \ln(2\pi) - \frac{N}{2} \ln(|\boldsymbol{\Sigma}|) - \frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})$$

多元高斯分布的极大似然估计

■去掉log似然函数中与参数无关的项，得到：

$$\ln p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = -\frac{N}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})$$

■似然函数求导并置**0**：

$$\frac{\partial \ln p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}) = 0$$

$$\left[\frac{\partial(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{y})}{\partial \boldsymbol{y}} = (\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{A})\boldsymbol{y}\right]$$

$$\widehat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{N}\boldsymbol{x}_i}{N} = \overline{\boldsymbol{x}}$$

样本的均值

多元高斯分布的极大似然估计

- log似然函数：

$$\ln p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{N}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

- 我们将参数$\boldsymbol{\Sigma}$用它的逆矩阵（精度矩阵$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$），则似然函数变为

$-\ln(|\boldsymbol{\Sigma}|) = \ln(\boldsymbol{\Sigma}^{-1})$

$$\ln p(\mathcal{D}|\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \frac{N}{2}\ln(|\boldsymbol{\Lambda}|) - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A} = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A})$

$$= \frac{N}{2}\ln(|\boldsymbol{\Lambda}|) - \frac{1}{2}\sum_{i=1}^{N}\mathrm{tr}\left((\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Lambda}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)$$

$\mathrm{tr}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{C}) = \mathrm{tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B})$

$$= \frac{N}{2}\ln(|\boldsymbol{\Lambda}|) - \frac{1}{2}\mathrm{tr}\left(\left(\underbrace{\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^{\mathrm{T}}}_{\boldsymbol{S}}\boldsymbol{\Lambda}\right)\right)$$

多元高斯分布的极大似然估计

■ 令$\boldsymbol{S}=(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$为数据的散度矩阵，则似然函数：

$$\ln p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Lambda})=\frac{N}{2}\ln(|\boldsymbol{\Lambda}|)-\frac{1}{2}\mathrm{tr}(\boldsymbol{S}\boldsymbol{\Lambda})$$

$$\frac{\partial(\ln|\boldsymbol{X}|)}{\partial\boldsymbol{X}}=\boldsymbol{X}^{-\mathrm{T}}$$

$$\frac{\partial\left(\mathrm{tr}(\boldsymbol{A}\boldsymbol{X})\right)}{\partial\boldsymbol{X}}=\boldsymbol{A}^{\mathrm{T}}$$

■ 似然函数求导并置$\mathbf{0}$：

$$\frac{\partial\ln p(\mathcal{D}|\boldsymbol{\mu},\boldsymbol{\Lambda})}{\partial\boldsymbol{\Lambda}}=\frac{N}{2}\boldsymbol{\Lambda}^{-\mathrm{T}}-\frac{1}{2}\boldsymbol{S}^{\mathrm{T}}=\mathbf{0}$$

样本的经验协方差矩阵

$$\boldsymbol{\Lambda}^{-\mathrm{T}}=\boldsymbol{\Lambda}^{-1}=\boldsymbol{\Sigma}=\frac{1}{N}\boldsymbol{S}^{\mathrm{T}}=\frac{1}{N}\boldsymbol{S}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}}$$

大纲

- 贝叶斯最优分类器
- 生成式分类器
- 概率密度函数的参数估计
- 朴素贝叶斯分类器
  - 朴素贝叶斯的类条件独立假设
  - 模型训练：
    - 类先验分布$p(y)$的概率密度参数估计
    - 在给定类别下，每维特征的条件概率密度估计$p(x_j|Y=c)$
- 高斯判别分析

朴素贝叶斯 (Naive Bayes Classifier, NBC)

- 假设共有 C 个类别$y \in \{1,2,\ldots,C\}$ ，类别的先验分布
  - 两类： $Y \sim \text{Bernoulli}(\theta)$
  - 多类 ：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$

- 每个样本的特征为：$\boldsymbol{x} = (x_1, x_2, \ldots, x_D)^{\mathrm{T}}$
- 朴素贝叶斯分类器：假设各维特征在给定类别标签的情况下条件独立

$$p(\boldsymbol{x}|Y = c) = \prod_{j=1}^{D} p\left(x_j|Y = c\right)$$

  - 实际应用中即使特征条件独立的假设不严格满足，NBC性能也不错
  - 因为NBC比较简单，不容易过拟合。

朴素贝叶斯模型训练：参数估计

- 朴素贝叶斯模型的训练过程就是估计模型的参数

- 类先验分布：$p(y)$
- 类条件分布：$p(x_j|Y=c)$

朴素贝叶斯模型训练：参数估计

- 贝叶斯分类器的log似然函数为

$$\text{Multinoulli}(y|\boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(Y=c)}$$

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i, y_i) = \sum_{i=1}^{N} \ln(p(\boldsymbol{x}_i|y_i)p(y_i))$$

$$= \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i|y_i) + \sum_{i=1}^{N} \ln p(y_i)$$

$$= \sum_{i=1}^{N} \ln \left(\prod_{c=1}^{C} \left(p(\boldsymbol{x}_i|Y_i = c)\right)^{\mathbb{I}(Y_i=c)} \right) + \sum_{i=1}^{N} \ln \left(\prod_{c=1}^{C} \theta_c^{\mathbb{I}(Y_i=c)} \right)$$

- 类先验分布的参数只与似然函数中第2项有关
- 类条件分布$p(x_{i,j}|Y_i = c)$只与似然函数中第1项有关
- 下面我们分别讨论两部分的模型参数的估计

朴素贝叶斯模型训练：参数估计

- 朴素贝叶斯模型的训练过程就是估计模型的参数

- 类先验分布$p(y)$：
  - 两类分类： $Y \sim \text{Bernoulli}(\theta)$
  - 多类分类：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$

- 类条件分布 ：$p(x_j|Y = c)$

两类分类任务中类先验分布的参数估计

- 对两类分类任务，类先验分布为贝努利分布：$Y \sim \text{Bernoulli}(\theta)$
- 根据之前贝努利分布参数估计的结论
- 参数θ极大似然估计为第1类样本占所有样本的比例：$\hat{\theta} = \frac{N_1}{N}$
- 其中$N_1 = \sum_{i=1}^{N} y_i$为第1类的样本的数目

- 贝努利分布的共轭先验为$\text{Beta}(\theta|\alpha, \beta)$，则$\theta$的贝叶斯后验分布为

$$p(\theta|\mathcal{D}) = \text{Beta}(\theta|\alpha + N_1, \beta + N_0)$$

- 点估计可取最大后验估计或后验均值估计

最大后验估计 $\hat{\theta} = \frac{\alpha + N_1 - 1}{\alpha + \beta + N - 2}$

$\hat{\theta} = \frac{\alpha + N_1}{\alpha + \beta + N}$ 后验均值估计

当$\alpha = \beta = 1$为均匀先验，
后验均值估计已被称为Laplace平滑

多类分类任务中类先验分布的参数估计

- 对C类分类任务，类先验分布为多项分布：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$
- 根据参数估计部分多项分布参数估计的结论，
- θ_c极大似然估计为第c类样本占所有样本的比例

$$\hat{\theta}_c = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)}{N} = \frac{N_c}{N}$$

$$\text{Multinoulli}(y|\boldsymbol{\theta}) = \prod_{c=1}^{C}\theta_c^{\mathbb{I}(Y=c)}$$

- 共轭先验为$\text{Dia}(\boldsymbol{\theta}|\boldsymbol{\alpha})$，贝叶斯后验为

$$p(\boldsymbol{\theta}|\mathcal{D}) = \text{Dia}(\boldsymbol{\theta}|\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_C + N_C)$$

- 参数的Laplace平滑估计为 $\hat{\theta}_c = \frac{N_c + 1}{N + C}$

朴素贝叶斯模型训练：参数估计

- 朴素贝叶斯模型的训练过程就是估计模型的参数
- 类先验分布的极大似然估计和贝叶斯估计：
  - 两类分类：$Y \sim \text{Bernoulli}(\theta)$
  - 多类分类：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$
- 类条件分布 ：$p(x_j|Y=c)$
  - 对每个类别c的每维特征x_j分别估计参数，与类先验分布的参数和其他类别、其他维特征分布的参数无关：将属于第c类的样本的第j维特征挑出来即可
  - 根据x_j的类型和取值范围，常用的概率分布包括
    - 二值：贝努利分布
    - 多个离散值： Multinoulli分布（类别分布）
    - 计数：多项分布
    - 连续值：高斯分布

NBC——二值特征

■二值特征：特征取值只有两种可能

- 例：在文档分类中，如每个词语在文档中是否出现
- $P(x_j|Y=c)$可用贝努利分布Bernoulli$(x_j|Y=c,\phi_{c,j})$表示，其中参数$\phi_{c,j}$表示在类别$Y=c$的情况下，特征$X_j=1$的概率。

$$P(x_{i,j}|y_i=c,\phi_{c,j})=(\phi_{c,j})^{x_{i,j}}(1-\phi_{c,j})^{1-x_{i,j}}$$

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious	Y
D1	1	1	1	1	0	1	1	1	0	0	0	1
D2	1	1	0	0	1	1	1	0	1	0	0	1
D3	0	0	0	0	0	1	0	0	0	1	1	0

D 个二值特征

NBC —— 二值特征

■朴素贝叶斯大大减少了模型的参数量

- 无独立假设：

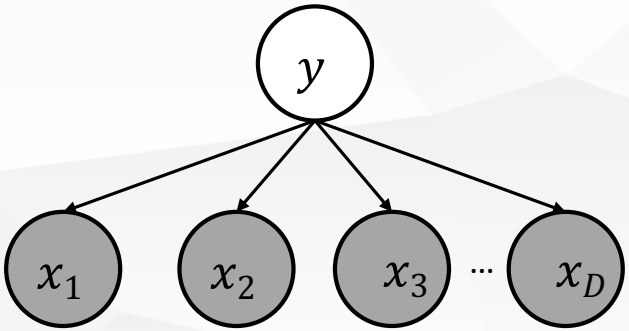
$$P(Y=c|\boldsymbol{x})=P(\boldsymbol{x}|Y=c)P(Y=c)$$

类条件密度参数：$C\times(2^D-1)$

类先验参数：$C-1$

- 条件独立假设：

$$P(Y=c|\boldsymbol{x})=P(\boldsymbol{x}|Y=c)P(Y=c)$$



$$=\prod_{j=1}^{D}P(x_j|Y=c)\;P(Y=c)$$

类条件密度参数：$C\times D$

类先验参数：$C-1$

NBC —— 二值特征

■参数$\phi_{c,j}$的极大似然估计为：

$$\hat{\phi}_{c,j} = \frac{\sum_{i=1}^{N} \mathbb{I}(y_i = c) x_{i,j}}{\sum_{i=1}^{N} \mathbb{I}(y_i = c)} = \frac{N_{c,j}}{N_c}$$

$\frac{\text{第}c\text{类的所有样本，第}j\text{维特征值为1的样本数目}}{\text{第}c\text{类的样本数目}}$

■Laplace平滑估计为：

$$\hat{\phi}_{c,j} = \frac{N_{c,j} + 1}{N_c + 2}$$

朴素贝叶斯模型训练：参数估计

- 朴素贝叶斯模型的训练过程就是估计模型的参数
- 类先验分布的极大似然估计和贝叶斯估计：
  - 两类分类：$Y \sim \text{Bernoulli}(\theta)$
  - 多类分类：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$
- 类条件分布 ：$p(x_j|Y=c)$
  - 对每个类别c的每维特征x_j分别估计参数，与类先验分布的参数和其他类别、其他维特征分布的参数无关：将属于第c类的样本的第j维特征挑出来即可
  - 根据x_j的类型和取值范围，常用的概率分布包括
    - 二值：贝努利分布
    - 多个离散值： Multinoulli分布（类别分布）
    - 计数：多项分布
    - 连续值：高斯分布

NBC —— 类别型特征

- 对离散型特征（类别型特征），假设特征有M种取值
  - $P(x_j|Y=c)$可用分布$\text{Multinoulli}(x_j|Y=c,\phi_{c,j,m})$表示，其中参数$\phi_{c,j,m}$表示在类别$Y=c$的情况下，特征$x_j=m$的概率。

亦被称为Categorical分布

$$P(X_{i,j}=m|Y_i=c,\phi_{c,j,m})=\prod_{m=1}^{M}(\phi_{c,j,m})^{\mathbb{I}(X_{i,j}=m)}$$

- 极大似然法估计：$\hat{\phi}_{c,j,m}=\frac{N_{c,j,m}}{N_c}$

$\frac{\text{第}c\text{类样本中，第}j\text{维特征值为}m\text{的样本数目}}{\text{第}c\text{类的样本数目}}$

- Laplace平滑估计：$\hat{\phi}_{c,j,m}=\frac{N_{c,j,m}+1}{N_c+M}$

例：社区账号的真实性判断

- 两类分类任务：Y = yes表示真实账号，Y = no表示不真实账号
- 选择三个特征属性：
  - X_1: 日志密度，有3种取值：s, m, l
  - X_2: 好友密度，有3种取值：s, m, l
  - X_3：是否使用真实头像，有2种取值：yes，no

例：社区账号的真实性判断 —— 模型训练

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

1. 类先验：共$N = 10$个样本，其中$Y = \text{yes}$的样本共有$N_{\text{yes}} = 7$

$$\hat{\theta}_{\text{yes}} = P(Y = \text{yes}) = \frac{N_{yes}}{N} = \frac{7}{10}$$

$$\hat{\theta}_{\text{no}} = P(Y = \text{no}) = 1 - P(Y = \text{yes})$$

$$= 1 - \frac{7}{10} = \frac{3}{10}$$

2. 类条件：

当类别$Y = \text{yes}$时，共有$N_{\text{yes}} = 7$个样本，

特征日志密度X_1有3种取值：s, l, m，样本数分别为：1，3，3，再加入平滑计数$\alpha = 1$，得到 $\hat{\phi}_{\text{yes},1,s} = \frac{1+1}{7+3} = \frac{1}{5}$，$\hat{\phi}_{\text{yes},1,l} = \frac{2}{5}$，$\hat{\phi}_{\text{yes},1,m} = \frac{2}{5}$

例：社区账号的真实性判断 —— 模型训练

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

1. 类先验

$$\hat{\theta}_{\text{yes}} = \frac{7}{10},\ \hat{\theta}_{\text{no}} = \frac{3}{10}$$

2. 类条件：

当类别 $Y = \text{yes}$ 时，对X_1特征

$$\hat{\phi}_{\text{yes},1,s} = \frac{1+1}{7+3} = \frac{1}{5},\ \hat{\phi}_{\text{yes},1,l} = \frac{2}{5},\ \hat{\phi}_{\text{yes},1,m} = \frac{2}{5}$$

当类别 $Y = \text{no}$ 时，共有$N_{\text{no}} = 3$个样本

特征日志密度X_1有3种取值：s, l, m，样本数分别为：2，1，0，再加入平滑计数$\alpha = 1$，得到

$$\hat{\phi}_{\text{no},1,s} = \frac{2+1}{3+3} = \frac{1}{2},\ \hat{\phi}_{\text{no},1,l} = \frac{1}{3},\ \hat{\phi}_{\text{no},1,m} = \frac{1}{6}$$

例：社区账号的真实性判断 —— 模型训练

日志密度	好友密度	是否使用真实头像	账号是否真实
s	s	no	no
s	l	yes	yes
l	m	yes	yes
m	m	yes	yes
l	m	yes	yes
m	l	no	yes
m	s	no	no
l	m	no	yes
m	s	no	yes
s	s	yes	no

1. 类先验分布的参数

$$\hat{\theta}_{\text{yes}} = \frac{7}{10},\ \hat{\theta}_{\text{no}} = \frac{3}{10}$$

2. 类条件分布的参数：

当类别 $Y = \text{yes}$ 时，对X_1特征

$$\hat{\phi}_{\text{yes},1,s} = \frac{1}{5},\ \hat{\phi}_{\text{yes},1,l} = \frac{2}{5},\ \hat{\phi}_{\text{yes},1,m} = \frac{2}{5}$$

当类别 $Y = \text{no}$时，对X_1特征

$$\hat{\phi}_{\text{no},1,s} = \frac{1}{2},\ \hat{\phi}_{\text{no},1,l} = \frac{1}{3},\ \hat{\phi}_{\text{no},1,m} = \frac{1}{6}$$

当类别 $Y = \text{yes}$ 时，对X_2特征

$$\hat{\phi}_{\text{yes},2,s} = \frac{1}{5},\ \hat{\phi}_{\text{yes},2,l} = \frac{3}{10},\ \hat{\phi}_{\text{yes},2,\text{m}} = \frac{1}{2}$$

当类别 $Y = \text{no}$时，对X_2特征

$$\hat{\phi}_{\text{no},2,s} = \frac{2}{3},\ \hat{\phi}_{\text{no},2,l} = \frac{1}{6},\ \hat{\phi}_{\text{no},2,m} = \frac{1}{6}$$

当类别 $Y = \text{yes}$ 时，对X_3特征

$$\hat{\phi}_{\text{yes},3,\text{no}} = \frac{4}{9},\ \hat{\phi}_{\text{yes},3,\text{yes}} = \frac{5}{9}$$

当类别 $Y = no$时，对X_3特征

$$\hat{\phi}_{\text{no},3,\text{no}} = \frac{2}{5},\ \hat{\phi}_{\text{no},3,\text{yes}} = \frac{3}{5}$$

例：社区账号的真实性判断 —— 测试

- 模型训练好后，对新的用户
  - 日志密度为m，好友密度为m，使用真实头像
- 判断该用户的社区帐号是否真实

$$P(Y=\text{yes}|X_1=\text{m},X_2=\text{m},X_3=\text{yes})$$
$$\propto P(X_1=\text{m}|Y=\text{yes})P(X_2=\text{m}|Y=\text{yes})\ P(X_3=\text{yes}|Y=\text{yes})\ \ P(Y=\text{yes})$$
$$=\hat{\phi}_{\text{yes},1,\text{m}}\times\hat{\phi}_{\text{yes},2,\text{m}}\times\hat{\phi}_{\text{yes},3,\text{yes}}\times\hat{\theta}_{\text{yes}}=\frac{2}{5}\times\frac{1}{2}\times\frac{5}{9}\times\frac{7}{10}=\frac{7}{90}$$

$$P(Y=\text{no}|X_1=\text{m},X_2=\text{m},X_3=\text{yes})$$
$$\propto P(X_1=\text{m}|Y=\text{no})P(X_2=\text{m}|Y=\text{no})\ P(X_3=\text{yes}|Y=\text{no})\ \ P(Y=\text{no})$$
$$=\hat{\phi}_{\text{no},1,\text{m}}\times\hat{\phi}_{\text{no},2,\text{m}}\times\hat{\phi}_{\text{no},3,\text{yes}}\times\hat{\theta}_{\text{no}}=\frac{1}{3}\times\frac{1}{6}\times\frac{3}{5}\times\frac{3}{10}=\frac{1}{100}$$

$P(Y=\text{yes}|X_1=\text{m},X_2=\text{m},X_3=\text{yes})>P(Y=\text{no}|X_1=\text{m},X_2=\text{m},X_3=\text{yes})$ 是真实账号

朴素贝叶斯模型训练：参数估计

- 朴素贝叶斯模型的训练过程就是估计模型的参数
- 类先验分布的极大似然估计和贝叶斯估计：
  - 两类分类：$Y \sim \text{Bernoulli}(\theta)$
  - 多类分类：$Y \sim \text{Multinoulli}(\boldsymbol{\theta})$
- 类条件分布 ：$p(x_j|Y=c)$
  - 对每个类别c的每维特征x_j分别估计参数，与类先验分布的参数和其他类别、其他维特征分布的参数无关：将属于第c类的样本的第j维特征挑出来即可
  - 根据x_j的类型和取值范围，常用的概率分布包括
    - 二值：贝努利分布
    - 多个离散值： Multinoulli分布（类别分布）
    - 计数：多项分布
    - 连续值：高斯分布

NBC —— 多项分布特征

- 多项分布的朴素贝叶斯主要用于文本分类任务
  - 多项分布：在多次Multinoulli试验中，每种结果出现的次数
  - 例：在文档分类中，如每个词语在文档中出现的次数（或TF-IDF）
  - 令字典中有D个单词，则对每个类别c，分布参数为1个D维向量：$\boldsymbol{\phi}_c = (\phi_{c,1}, \phi_{c,2}, \dots, \phi_{c,D})$

- 参数的极大似然估计

$$\hat{\phi}_{c,d} = \frac{N_{c,d}}{N_c}$$

$$\frac{\text{第}c\text{类的所有样本/文档中，第}d\text{个单词出现的次数和}}{\text{第}c\text{类文档中的总单词数目}}$$

朴素贝叶斯模型训练：参数估计

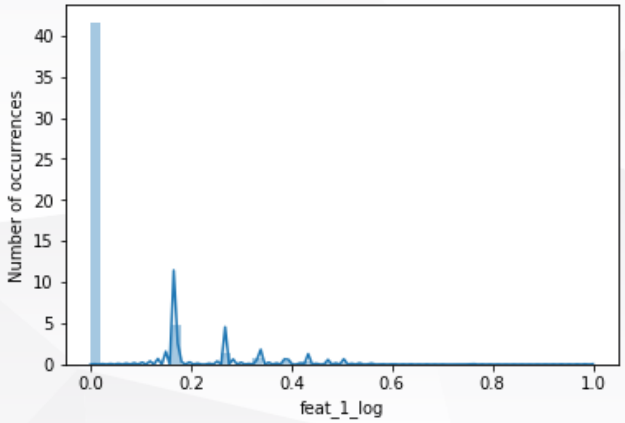
- 朴素贝叶斯模型的训练过程就是估计模型的参数
- 类先验分布的极大似然估计和贝叶斯估计：
  - 两类分类：$y \sim \text{Bernoulli}(\theta)$
  - 多类分类：$y \sim \text{Multinoulli}(\boldsymbol{\theta})$
- 类条件分布 ：$p(x_j|y=c)$
  - 对每个类别c的每维特征x_j分别估计参数，与类先验分布的参数和其他类别、其他维特征分布的参数无关：将属于第c类的样本的第j维特征挑出来即可
  - 根据x_j的类型和取值范围，常用的概率分布包括
    - 二值：贝努利分布
    - 多个离散值： Multinoulli分布（类别分布）
    - 计数：多项分布
    - 连续值：高斯分布

NBC —— 连续特征（高斯分布）

- 当特征取值为连续值，且在类边缘分布为高斯分布时，

$$p(x_{i,j}|Y_i = c, \mu_{c,j}, \sigma_{c,j}) = N(\mu_{c,j}, \sigma_{c,j}) = \frac{1}{\sqrt{2\pi}\sigma_{c,j}} \exp\left(-\frac{(x_{i,j} - \mu_{c,j})^2}{2(\sigma_{c,j})^2}\right)$$

- 注意：不是所有的连续特征都可假设为高斯分布



NBC —— 连续特征（高斯分布）

- 参数的极大似然估计为

$$\hat{\mu}_{c,j} = \frac{\sum_{i=1}^{N} \mathbb{I}(Y_i = c) x_{i,j}}{\sum_{i=1}^{N} \mathbb{I}(Y_i = c)}$$

第c类样本中，第j维特征值的均值

$$\left(\sigma_{c,j}\right)^2 = \frac{\sum_{i=1}^{N} \sum_{m=1}^{M} \mathbb{I}(Y_i = c) \left(x_{i,j} - \mu_{j,c}\right)^2}{\sum_{i=1}^{N} \mathbb{I}(Y_i = c)}$$

第c类的样本中，
第j维特征值的经验方差

Sklearn中的朴素贝叶斯实现

- Scikit-Learn中提供5种朴素贝叶斯的分类算法：
  - GaussianNB：特征值为连续值且为高斯分布
  - BernoulliNB：特征值为二值
  - CategoricalNB：特征值为多个离散值
  - MultinomialNB：特征值表示某种事件出现的次数
  - ComplementNB： MultinomialNB的一种改进，特别适用于不平衡数据集。 ComplementNB使用来自每个类的补数的统计数据来计算模型的权重，这样参数估计更稳定，在文本分类任务上，性能通常更好。

案例：新闻文档分类

- 20newsgroups数据集
  - 11,314个新闻组文档，分档分为20个不同主题
  - 80%作为训练数据，20%为测试数据
  - 特征：TF-IDF（2元语法模型），特征向量为155,785维
  - MultinomialNB：正确率为0.894388

chp2_news_nb.ipynb

案例：新闻文档分类

新闻类别	Precision	Recall	F1-score	Support
alt.atheism	0.97	0.94	0.95	89
comp.graphics	0.79	0.84	0.81	99
comp.os.ms-windows.misc	0.89	0.91	0.90	130
comp.sys.ibm.pc.hardware	0.86	0.80	0.83	109
comp.sys.mac.hardware	0.92	0.92	0.92	117
comp.windows.x	0.92	0.89	0.91	118
misc.forsale	0.85	0.95	0.90	117
rec.autos	0.95	0.94	0.95	121
rec.motorcycles	0.96	0.97	0.97	119
rec.sport.baseball	0.98	1.00	0.99	113
rec.sport.hockey	1.00	0.96	0.98	129
sci.crypt	0.96	0.99	0.98	109
sci.electronics	0.91	0.93	0.92	120
sci.med	0.99	0.91	0.95	123
sci.space	0.97	0.97	0.97	119
soc.religion.christian	0.93	0.98	0.95	128
talk.politics.guns	0.97	0.96	0.97	120
talk.politics.mideast	1.00	1.00	1.00	100
talk.politics.misc	0.94	0.94	0.94	106
talk.religion.misc	0.96	0.84	0.90	77
avg / total	0.94	0.94	0.94	2263

大纲

- 贝叶斯最优分类器
- 生成式分类器
- 概率密度函数的参数估计
- 朴素贝叶斯分类器
- 高斯判别分析
  - 高斯判别分析的假设：类条件分布为多元高斯分布
  - 一般情况下，判别函数为输入x的二次函数
  - 当两个类别的协方差矩阵相等时，判别函数为线性函数
  - 高斯判别模型的训练/参数估计：极大似然估计、贝叶斯估计

高斯判别分析

- 高斯判别分析假设每类数据由一个多元高斯分布产生，即 $p(\boldsymbol{x}|Y=c)=N(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$

- 类别的先验分布为$P(Y=c)$
  - 两类： $Y\sim\text{Bernoulli}(\theta)$
  - 多类 ：$Y\sim\text{Multinoulli}(\boldsymbol{\theta})$

- 根据贝叶斯公式，可计算给定特征时类别的后验概率为：

$$P(Y=c|\boldsymbol{x})=\frac{p(\boldsymbol{x}|Y=c)P(Y=c)}{\sum_{c'}p(\boldsymbol{x}|Y=c')P(Y=c')}$$

高斯判别分析的判别函数

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

- 类别c的判别函数：

$$f_c(\boldsymbol{x}) = \ln\big(P(\boldsymbol{x}|Y=c)\big) + \ln\big(P(Y=c)\big)$$

$$= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_c|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \ln P(Y=c)$$

- 去掉与参数$\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$无关的项（不影响分类结果），得到

$$f_c(\boldsymbol{x}) = -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_c|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^{\mathrm{T}}\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_c) + \ln P(Y=c)$$

高斯判别分析的判别函数：两类

■两个类别的判别函数分别为：

$$f_1(\boldsymbol{x}) = -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_1|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \ln P(Y=1)$$

$$f_2(\boldsymbol{x}) = -\frac{1}{2}\ln(|\boldsymbol{\Sigma}_2|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln P(Y=2)$$

$$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \ln\frac{P(Y=1)}{P(Y=2)}$$

$$= -\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + b + \ln\frac{P(Y=1)}{P(Y=2)}$$

判别函数$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})$为$\boldsymbol{x}$的二次函数，因此高斯判别分析亦被称为二次判别分析（Quadratic Discriminant Analysis, QDA）

其中$b = -\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$与$\boldsymbol{x}$无关

高斯线性判别分析：两类且协方差矩阵相等

■当两类的协方差矩阵$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$时：

$$b = -\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

$$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1})\boldsymbol{x} + \left(\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\right)\boldsymbol{x} + b + \ln\frac{P(Y=1)}{P(Y=2)}$$

$$= (\boldsymbol{\mu}_1^{T} - \boldsymbol{\mu}_2^{T})\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + b + \ln\frac{P(Y=1)}{P(Y=2)}$$

判别函数$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})$为$\boldsymbol{x}$的一次函数/线性函数
亦被称为线性判别分析（Linear Discriminant Analysis, LDA）

其中$b = -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$与$\boldsymbol{x}$无关

线性判别分析

■更特别的，当两个高斯分布为各向同性，即$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{I}$时，

$$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T)\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{P(Y=1)}{P(Y=2)}$$

$$= \sigma^{-2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{I}\boldsymbol{x} - \frac{\sigma^{-2}}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{I}\boldsymbol{\mu}_1 + \frac{\sigma^{-2}}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{I}\boldsymbol{\mu}_2 + \ln\frac{P(Y=c_1)}{P(Y=c_2)}$$

$$= \sigma^{-2}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\mu}_2\right] + \ln\frac{P(Y=c_1)}{P(Y=c_2)}$$

■若两类的先验相等 ，即$P(Y=1) = P(Y=2)$时

$$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = \sigma^{-2}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\mu}_2\right]$$

线性判别分析

■当$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\boldsymbol{I}$，且两类的先验相等 ，即$P(Y=1) = P(Y=2)$时，

$$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = \sigma^{-2}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\mu}_2\right]$$

■若$\boldsymbol{x}$位于决策边界上，$f_1(\boldsymbol{x}) - f_2(\boldsymbol{x}) = 0$，则

$$\sigma^{-2}\left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\mu}_2\right] = 0$$

两边同乘以σ^2

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\mu}_2 = 0$$

$a^2 - b^2 = (a+b)(a-b)$

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = 0$$

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathrm{T}}\left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right) = 0$$

$\boldsymbol{x}$位于两个类中心的垂直平分线上

高斯判别模型的训练

- 类先验：Bernoulli分布或Multiltinoulli分布
  - 同朴素贝叶斯分类器

- 类条件分布的参数：每个类别的均值向量$\boldsymbol{\mu}_c$和协方差矩阵$\boldsymbol{\Sigma}_c$
- 根据概率密度函数参数估计部分关于多元高斯分布参数估计的结论：

$$\hat{\boldsymbol{\mu}}_c = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)\,\boldsymbol{x}_i}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)} = \bar{\boldsymbol{x}}_c$$

第c类样本的均值

$$\hat{\boldsymbol{\Sigma}}_c = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)\,\boldsymbol{S}_c}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)\,(\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^{\mathrm{T}}}{\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(Y_i = c)}$$

第c类的样本的经验协方差矩阵

高斯判别模型的训练 —— 正则

- 当训练样本数量N_c相比特征维度D较小时，上述协方差矩阵$\hat{\boldsymbol{\Sigma}}_c$是奇异的
- 可采用收缩（Shrinkage）提升的协方差矩阵预测准确性：

$$\hat{\boldsymbol{\Sigma}}_c = (1-\delta)\boldsymbol{S}_c + \delta\boldsymbol{F}$$

- 其中δ是收缩因子，$\boldsymbol{F}$为一个高度结构化的矩阵，如对角矩阵$\boldsymbol{I}$

Scikit-learn中的高斯判别分析

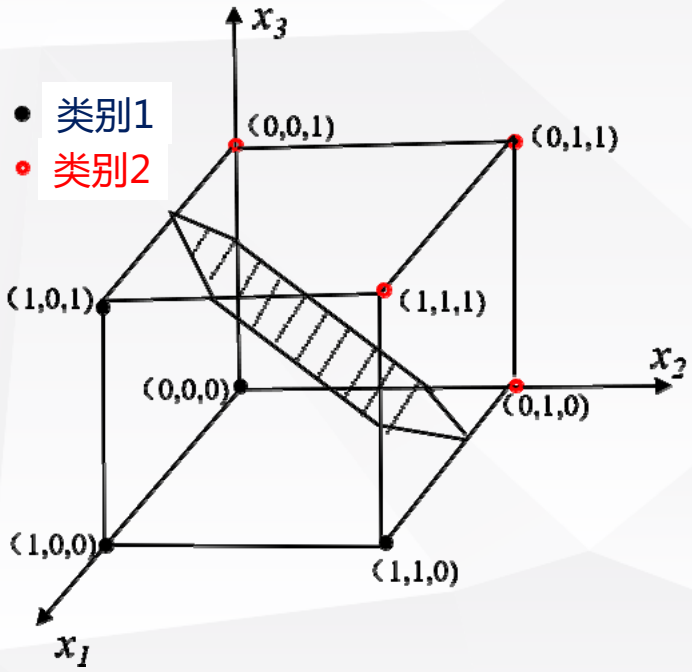
class sklearn.discriminant_analysis.QuadraticDiscriminantAnalysis(*priors=None*, *reg_param=0.0*, *store_covariances=False*, *tol=0.0001*)

参数	说明
priors	每个类别的先验
reg_param	协方差的正则参数：$\hat{\Sigma}_c = (1-\delta)\boldsymbol{S}_c + \delta \boldsymbol{I}$

属性	说明
priors_	每个类别的先验
means_	每个类别的均值向量
covariances_	每个类别的协方差矩阵
rotations_	高斯分布的旋转，即主轴
scalings_	旋转高斯分布的各主轴的方差

例：线性判别分析

■设有两个类别的模式，每个类别的特征为高斯分布



(1) 类先验分布：$P(Y=1)=P(Y=2)=1/2$

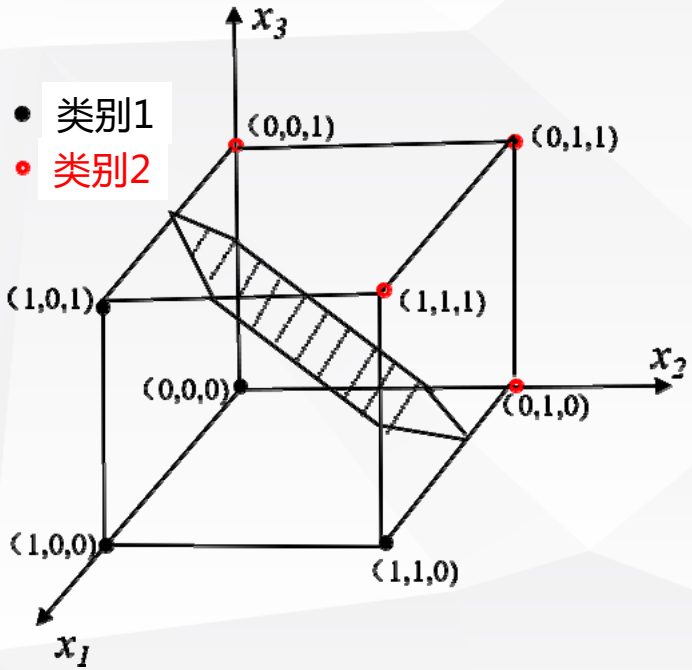
(2) 类条件分布

$$\hat{\boldsymbol{\mu}}_1=\frac{1}{4}\left(\begin{pmatrix}0\\0\\0\end{pmatrix}+\begin{pmatrix}1\\0\\0\end{pmatrix}+\begin{pmatrix}1\\1\\0\end{pmatrix}+\begin{pmatrix}1\\0\\1\end{pmatrix}\right)=\frac{1}{4}\begin{pmatrix}3\\1\\1\end{pmatrix}$$

$$\hat{\boldsymbol{\mu}}_2=\frac{1}{4}\left(\begin{pmatrix}0\\0\\1\end{pmatrix}+\begin{pmatrix}0\\1\\1\end{pmatrix}+\begin{pmatrix}0\\1\\0\end{pmatrix}+\begin{pmatrix}1\\1\\1\end{pmatrix}\right)=\frac{1}{4}\begin{pmatrix}1\\3\\3\end{pmatrix}$$

例：线性判别分析

- 设有两个类别的模式，每个类别的特征为高斯分布



(1) 类先验分布：$P(Y=1)=P(Y=2)=1/2$

(2) 类条件分布：

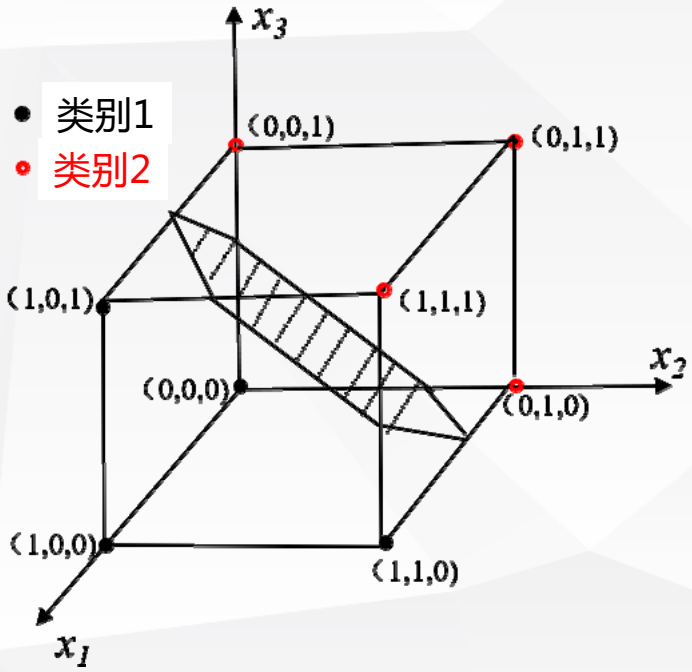
$$\hat{\boldsymbol{\mu}}_1=\frac{1}{4}\begin{pmatrix}3\\1\\1\end{pmatrix} \qquad \hat{\boldsymbol{\mu}}_2=\frac{1}{4}\begin{pmatrix}1\\3\\3\end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1=\hat{\boldsymbol{\Sigma}}_2=\frac{1}{16}\begin{bmatrix}3 & 1 & 1\\ 1 & 3 & -1\\ 1 & -1 & 3\end{bmatrix}$$

$$\hat{\Sigma}_1=\frac{1}{4}\left[\begin{pmatrix}-3/4\\-1/4\\-1/4\end{pmatrix}(-3/4 \quad -1/4 \quad -1/4)+\begin{pmatrix}1/4\\-1/4\\-1/4\end{pmatrix}(1/4 \quad -1/4 \quad -1/4)+\begin{pmatrix}1/4\\3/4\\-1/4\end{pmatrix}(1/4 \quad 3/4 \quad -1/4)+\begin{pmatrix}1/4\\-1/4\\3/4\end{pmatrix}(1/4 \quad -1/4 \quad 3/4)\right]$$

例：线性判别分析

■ 设有两个类别的模式，每个类别的特征为高斯分布



(1) 类先验分布：$P(Y=1)=P(Y=2)=1/2$

(2) 类条件分布：

$$\hat{\boldsymbol{\mu}}_1=\frac{1}{4}\begin{pmatrix}3\\1\\1\end{pmatrix}\qquad \hat{\boldsymbol{\mu}}_2=\frac{1}{4}\begin{pmatrix}1\\3\\3\end{pmatrix}$$

$$\hat{\boldsymbol{\Sigma}}_1=\hat{\boldsymbol{\Sigma}}_2=\boldsymbol{\Sigma}=\frac{1}{16}\begin{bmatrix}3&1&1\\1&3&-1\\1&-1&3\end{bmatrix}\quad \boldsymbol{\Sigma}^{-1}=4\begin{bmatrix}2&-1&-1\\-1&2&1\\-1&1&2\end{bmatrix}$$

(3) 判别函数为x线性函数

$$d_1(\boldsymbol{x})-d_2(\boldsymbol{x})=(\boldsymbol{\mu}_1^{\mathrm{T}}-\boldsymbol{\mu}_2^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$$

$$=(2\quad -2\quad -2)\boldsymbol{x}+1=2x_1-2x_2-2x_3+1=0$$

案例：鸢尾花分类

- 朴素贝叶斯分类器：高斯朴素贝叶斯
- QDA

chp2_iris_bayesian.ipynb

总结

- 最小错误率/最小风险分类器：贝叶斯分类器

- 产生式分类器
  - 类先验：$P(Y=c)$
  - 类条件：$P(\boldsymbol{x}|Y=c)$
    - 朴素贝叶斯：在给定类别的情况下，各特征之间独立

$$p(\boldsymbol{x}|Y=c)=\prod_{j=1}^{D}p(x_j|Y=c)$$

    - 多元正态分布

$$p(\boldsymbol{x}|Y=c)=N(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)$$