# 第二章

## 本章目的：

**假设样本之间都是独立同分布的，对于给定的样本集X = (x1, x2, x3,....)，估计这个样本集服从的概率分布，称为密度估计density estimation。**

- 参数分布 parametric distribution:

  当样本是离散随机变量时，使用 二项分布或多项分布来建模；

  当样本是连续随机变量时，使用高斯分布建模。

  这些是参数分布的典型例子，因为模型由一小部分参数来控制，如高斯分布中的均值和方差。

- 参数分布中参数的求解：

  在频率论的角度，我们为参数选取某个特定的值来优化评价标准的结果，如用极大似然估计；

  在贝叶斯的角度，首先估计参数的先验分布，通过贝叶斯理论，估计出基于给定数据集的参数的后验分布，然后使后验分布最大求解最优参数(?)

  - 共轭先验，即后验分布和先验分布具有同样的形式，如多项分布的共轭先验是狄利克雷分布。共轭先验分布极大地简化了使用贝叶斯分析的方法。
- 非参数的密度估计

  有的样本集的分布未必是服从某种函数，另一种方法是 使用非参数的密度估计，即样本集的分布形式取决于样本集的大小，这样的模型也有参数，但参数是用于控制模型的复杂度，而不是决定样本集的分布形式。常见的有最近邻法、核函数法等等。

# 二值变量

## 抛硬币问题

假设该硬币存在破损，记正面朝上的概率为$\mu$，当x=1时正面朝上，否则x=0；

则$P\{x=1|\mu\}=\mu$，$P\{x=0|\mu\}=1-\mu$。

合并可写成 $Bern\{x|\mu\}=\mu^x*(1-\mu)^{(1-x)}$，即$x$的概率分布可写成关于$\mu$的函数。显然这是一个伯努利分布。

假设进行了n次实验，得到一个样本集D，$P\{D|\mu\}=\prod_n P\{X_n|\mu\}$

**1. 频率论的角度：**

$\ln p(D|\mu)=\sum_{n=1}^{N}\ln p(x_n|\mu)=\sum_{n=1}^{N}\{x_n\ln\mu+(1-x_n)\ln(1-\mu)\}$

对$\mu$ 求导，令导数为0，得到 $\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n$

此即**极大似然估计法**求出的参数$\mu$ 。 若正面朝上m 次，则 $\mu_{ML}=m/N$。

然而，这一方法存在偏差，考虑当N=3时，若正好3次实验都是正面朝上，则 $\mu_{ML}=1$ ，意味着它预测将来所有的实验都是正面朝上，这是不合常理的，是极大似然估计**存在过拟合**的极端例子。

**2. 贝叶斯的角度：**

为了防止过拟合，引入 $\mu$ 的先验分布来矫正。假设 $\mu$ 的先验分布服从Beta分布

*解释：*由于$\mu$ 的后验分布 正比于 先验*似然， 而似然函数里面含有 $\mu^x*(1-\mu)^{1-x}$这一项，那么如果我们选择的先验分布也含有类似的形式，即正比于$\mu$ 和 $1-\mu$ 的幂次，相应的后验分布也会含有这样的幂次的形式，那么此时后验分布和先验分布共轭。

$\mu$ 的先验分布: $p(\mu)=Beta(\mu|a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$

- 参数为a,b，系数为gamma函数，系数的形式是为了保证Beta分布的归一化
- a,b称为整个模型的**超参数**，因为它们控制着参数 $\mu$ 的取值分布

由此得到 $\mu$ 的后验分布: $p(\mu|m,l,a,b)=\frac{\Gamma(m+l+a+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$

- $m$ 为正面朝上次数，$l=N-m$ 为反面朝上次数。可解释为后验分布中，正面朝上的次数由$m$增加到$m+a$，反面朝上的次数由 $l$ 增加到 $l+b$，而这一次

的后验分布，是下一次的先验分布。
- 因此，可以看做每次试验都增加一个新样本，通过将当前的先验分布与关于新增加样本的似然函数相乘，得到新的后验分布。

于是，当我们使用贝叶斯的角度时，相当于使用了**序列化**的学习方法。序列化的学习，能够每次只新增一个或一小批样本，并在每次计算完后可以丢弃，非常适用于实时学习的场景，也非常适合用于大规模的数据集，因为并不是全部的数据都需要存储或加载。

前一个实验的后验会作为后一个实验的先验，逐步提高准确性。并且这种顺序方法只依赖于数据的独立性，不必存储数据，只需要流水线地处理数据修正参数即可。

**3. 随着数据集增大，模型参数$\mu$ 的方差会越来越小**

可以从频率论的角度来说明。在平均意义下，这一结论成立。
$var_\theta[\theta] = E_D[var_\theta[\theta|D]] + var_\theta[E_\theta[\theta|D]] \geq E_D[var_\theta[\theta|D]]$

即先验参数方差 大于 后验参数方差的均值

# 多值变量

若样本一共有K种取值，则可以用一个K维向量表示，取第i个值则在向量的对应下标记为1，其它下标记为0。 其它求解与二项分布类似，只是对于模型参数多了条件约束 $\sum_{k=1}^{K} \mu_k = 1$

# 高斯分布

## 一维变量的高斯分布

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$$

## 多维变量的高斯分布

$$\mathcal{N}(\mathbf{x}|\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}}exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\}$$

其中 $\Sigma$ 为协方差矩阵

中心极限定理的验证见matlab文件 CLT.m ，即画出课本的图2.6

## 1. 高斯分布的几何形式

高斯分布对输入x的依赖, 体现在指数部分： $\Delta^2=(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)$

$\Delta^2$ 称为马氏距离，当协方差矩阵为单位阵时，退化为欧氏距离。

由于协方差矩阵是实对称矩阵，故可以找到特征方程：$\Sigma u_i=\lambda_i u_i$

取特征向量为单位正交，则**协方差矩阵**可以改写成对角矩阵，对角线在每一行的元素值 $\lambda_i u_i u_i^T$

代入得到$\Delta^2=\sum_{i=1}^{D}\frac{y_i^2}{\lambda_i}$， $y_i=u_i^T(\mathbf{x}-\mu)$

这相当于将原来的X坐标系，经过旋转和平移 变为Y坐标系

关于椭圆的理解https://blog.csdn.net/weixin_37895339/article/details/80351541

椭圆，如对二维正态分布，等概率切面就是一个椭圆，若是标准正态分布则是圆

在Y坐标系下的椭圆，中心为X坐标系下的$(\mu_1,\mu_2)$，长轴为$\sqrt{\lambda_1}$，短轴为$\sqrt{\lambda_2}$

要将X变化为服从标准正态分布，记$A_w=U\Lambda^{-1/2}$

则$Y=A_w^T(X-\mu)$， 得到的$Y\sim N(0,I)$

*验证*: $E[Y]=A_w^TE(X-\mu)=0, E[YY^T]=E[A_w^T(X-\mu)(X-\mu)^TA_w]$,

代入$A_w$ 以及$\Sigma=U\Lambda U^{-1}$, 得到$E[YY^T]=I$, 则
$cov[Y]=E[YY^T]-E^2[Y]=I$

更多理解见活页笔记本

## 2. 多维高斯分布的矩

$E[X] = \mu$ (D维列向量),

$E[XX^T] = \mu\mu^T + \Sigma$

$cov[X] = E[XX^T] - E[X^2] = \mu\mu^T + \Sigma - \mu\mu^T = \Sigma = E[(X - E[X]) \cdot (X - E[X])^T]$

## 3. 高斯分布的局限性

- 参数较多。一般的对称协方差矩阵有D(D+1)/2个参数，均值向量有D个参数，总共就是D(D+3)/2个参数，随着D的增大，未知量呈D^2的速度增长。若协方差矩阵是对角阵，则总共的参数只有2D个。若协方差矩阵正比于单位阵，称为各向同性的协方差，则总共的参数只有D+1。

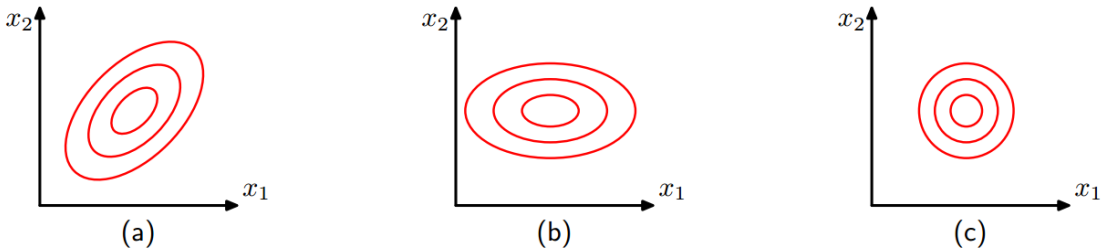


图 2.8: 二维高斯分布的常数概率密度轮廓线，其中，(a)图对应的协方差矩阵为一般形式，(b)图对应的协方差矩阵为对角矩阵，图中椭圆的轮廓线与坐标轴对齐，(c)图对应的协方差矩阵正比于单位矩阵，图中的轮廓线是同心圆。

- 分布是单峰的，无法近似多峰的分布。可通过引入潜在变量或未观察变量解决

## 4. 条件高斯分布

https://blog.csdn.net/hubin232/article/details/70335847

对于多变量高斯分布，存在这样的性质，以两个变量为例，如{A,B}的联合分布是高斯分布，则P(A|B)或P(B|A)是高斯分布。

则利用这样的性质，可以对多维变量进行分块 partition

协方差矩阵的逆 记作 精准矩阵 precision matrix

## 5. 边缘高斯分布

可以基于条件分布求积分

## 6.高斯分布的极大似然估计

给定的N个IID的数据集$X$，其中每个X服从D维高斯分布

$\ln p(X|\mu,\Sigma)=-\frac{ND}{2}\ln(2\pi)-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)$

充分统计量 $\sum_{n=1}^{N}x_n$，$\sum_{n=1}^{N}x_nx_n^T$

对$\mu$ 求导，$\frac{\partial}{\partial\mu}\ln p(X|\mu,\Sigma)=\sum_{n=1}^{N}\Sigma^{-1}(x_n-\mu)=0$

$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n$, $\Sigma_{ML}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})(x_n-\mu_{ML})^T$

求期望，$E[\mu_{ML}]=\mu, E[\Sigma_{ML}]=\frac{N-1}{N}\Sigma$ 与真实值相差$1/N$

顺序估计，每次新增一个样本：

$$\begin{aligned}\mu_{ML}^{N}&=\frac{1}{N}\sum_{n=1}^{N}x_n\\&=\frac{1}{N}x_n+\frac{1}{N}\sum_{n=1}^{N-1}x_n\\&=\frac{1}{N}x_n+\frac{N-1}{N}\mu_{ML}^{N-1}\\&=\mu_{ML}^{N-1}+\frac{1}{N}(x_N-\mu_{ML^{N-1}})\end{aligned}$$

其中，第一项为上一次的估计，第二项为估计的修正值，1/N可看作新样本起的修正作用的权值

## 7.高斯分布的贝叶斯推断

对于IID的数据集，关于 $\mu$ 的似然函数

$p(X|\mu)=\prod_{n=1}^{N}p(x_n|\mu)=\frac{1}{(2\pi\sigma^2)^{N/2}}exp\{\frac{-1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\}$

指数项部分可以写成$\mu^2N-2\mu\sum_{n=1}^{N}x_n+\sum_{n=1}^{N}x_n^2$ ，

*这是关于$\mu$ 的函数，形状为高斯（因为整个形式与高斯类似，只是指数项的部分有个$\mu^2$）。而不是一个参数为 $\mu$ 的X的分布*

假设$\mu$ 的先验分布为高斯分布 $p(\mu) = N(\mu|\mu_0, \sigma_0^2)$

则 后验分布 正比于 先验分布*似然 $p(\mu|X) = N(\mu|\mu_N, \sigma_N^2)$

其中$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$，$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$

当N趋于无穷，$\mu_N \to \mu_{ML}$，$\sigma_N^2 \to 0$

序列化学习：

$$
\begin{aligned}
p(\mu|X) &\propto p(\mu)p(X|\mu) \\
&= [p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)]p(x_N|\mu) \\
&\propto N(\mu|\mu_{N-1}, \sigma_{N-1}^2)p(x_N|\mu)
\end{aligned}
$$

总结*:

- 对于一维的高斯随机变量：

  **1.假设方差已知，求均值** $\mu$ 可观察到似然函数关于 $\mu$ 的形式类似于高斯的形式，可假设均值的先验分布服从高斯分布（与似然函数共轭），对应的后验概率是两个$\mu$ 的二次函数的指数的乘积（指数的相乘可转换为相加），故后验概率也是高斯分布的形式

  **2.假设均值已知，求方差，**同理，选择共轭分布作为先验分布可大大简化计算，在这里选择先验分布为Gamma分布

  **3.最后假设均值和方差都是未知的**，那么可以将先验分布选为高斯-gamma分布

- 对于D维的高斯随机变量：

  与一维类似，其中Gamma分布替换为Wishart分布

# 学生t分布

将精度积分，求出x的边缘分布（相对于精度来说的边缘），将结果定义为学生t分布

# Student's t-Distribution

$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathrm{Gam}(\tau|a,b)\,\mathrm{d}\tau \\
&= \int_0^\infty \mathcal{N}\left(x|\mu,(\eta\lambda)^{-1}\right)\mathrm{Gam}(\eta|\nu/2,\nu/2)\,\mathrm{d}\eta \\
&= \frac{\Gamma(\nu/2+1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left[1+\frac{\lambda(x-\mu)^2}{\nu}\right]^{-\nu/2-1/2} \\
&= \mathrm{St}(x|\mu,\lambda,\nu)
\end{aligned}
$$

where

$$\lambda = a/b \qquad \eta = \tau b/a \qquad \nu = 2a.$$

Infinite mixture of Gaussians.

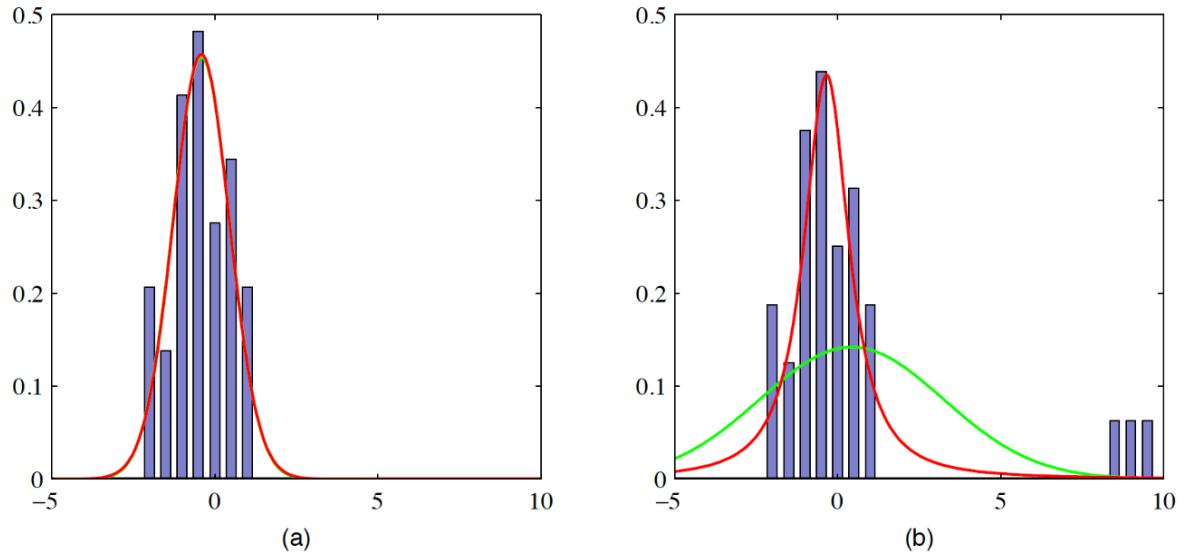


图 2.16: 与高斯分布相比，学生t分布具有鲁棒性的例子。(a)从一个高斯分布中抽取的30个数据点的直方图，以及得到的最大似然拟合。红色曲线表示使用t分布进行的拟合，绿色曲线（大部分隐藏在了红色曲线后面）表示使用高斯分布进行的拟合。由于t分布将高斯分布作为一种特例，因此它给出了与高斯分布几乎相同的解。(b)同样的数据集，但是多了三个异常数据点。这幅图展示了高斯分布（绿色曲线）是如何被异常点强烈地干扰的，而t分布（红色曲线）相对不受影响。



## 周期变量

应用于周期性变量的高斯分布。高斯分布对周期性变量的概率分布拟合效果差，可以推广高斯分布的形式，得到von Mises分布(或称环形正态分布)。这个分布考虑的变量的概率分布需要满足

$$
\begin{gathered}
p(\theta) \geq 0 \\
\int_0^{2\pi} p(\theta) d\theta = 1 \\
p(\theta + 2\pi) = p(\theta)
\end{gathered}
$$

极大似然估计法求解von-Mises分布的参数

# 混合高斯分布

有的数据可能由不止一个高斯分布生成，比如一部分数据由分布1生成，另一部分由分布2生成。即数据存在多峰，使用普通的高斯分布具有局限性。考虑多个高斯分布的线性叠加可解决。

**极大似然求解参数，形式比较复杂**，在对数项里面包含有求和，没有闭式解。

**解决方法**：使用迭代的数值优化方法，或是使用EM算法

# 指数族分布

伯努利分布可以写成指数的形式，则某一项可以由sigmoid函数表示

多项分布写成指数的形式，某一项可以由softmax函数表示

# 非参数化方法

之前的假设是建立在变量的概率密度分布符合某种特定的函数形式，且形式是由少量的参数控制的，但这严重依赖于所选分布是否准确。另一种方法是非参数化的方法，对概率密度的形式很少进行估计，而是从频率的角度。

**1. 直方图法：**

优点：一旦直方图被计算出来，数据本身就被丢弃了。适用于数据量很大的场景。

缺点：受所划分箱子的宽度大小的影响。若箱子宽度很小时，最终概率模型有很多尖刺，属于噪声；若箱子宽度过大，则最终概率模型会过于平滑，无法描述细节。

**2.核估计、近邻估计：**

假设x落在某个区域R内的概率为P,则N次试验中，有K次试验的x落在了区域R内的概率服从二项分布。。。

KNN的分类方法：

点，我们把$K$近邻概率密度估计方法分别应用到每个独立的类别中，然后使用贝叶斯定理。假设我们有一个数据集，其中$N_k$个数据点属于类别$\mathcal{C}_k$，数据点的总数为$N$，因此$\sum_k N_k = N$。如果我们想对一个新的数据点$\boldsymbol{x}$进行分类，那么我们可以画一个以$\boldsymbol{x}$为中心的球体，这个球体精确地包含$K$个数据点（无论属于哪个类别）。假设球体的体积为$V$，并且包含来自类别$\mathcal{C}_k$的$K_k$个数据点。这样公式（2.246）提供了与每个类别关联的一个概率密度的估计

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \frac{K_k}{N_k V} \tag{2.253}$$

p(x|Ck)的意思是，已知这个点是Ck类别，求这个点是x的概率

*非参数方法需要存储并计算所有数据，而参数化方法（如果拟合效果好）在存储和计算上相比非参数方法更高效。*