模式识别与机器学习

概率论

张晓平

武汉大学数学与统计学院

Table of contents

高斯分布

高斯分布

高斯分布 (Gaussian distribution)，又名正态分布 (Normal distribution)，是一个在数学、物理及工程等领域都非常重要的概率分布，在统计学的许多方面有着重大的影响力。

- 对于一元变量 x，高斯分布的形式为

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

  其中 μ 为均值，σ^2 为方差。

- 对于 D 维向量 $\boldsymbol{x}$，多元高斯分布的形式为

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

  其中 $\boldsymbol{\mu} \in \mathbb{R}^D$ 为均值，$\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$ 为协方差矩阵。

高斯分布

- 对于一元随机变量，最大熵对应的分布是高斯分布；该性质对多元高斯分布也成立。
- 正态分布的的普遍性可从中心极限定理得到。直白地说，如果一个指标受到若干独立因素的共同影响，且每个因素不能产生支配性的影响，那么这个指标就服从中心极限定理，收敛到正态分布，这就是林德伯格-费勒中心极限定理的意思。

高斯分布

由(2)可看出，高斯分布对 $\boldsymbol{x}$ 的依赖由以下二次型确定:

$$\Delta^2 = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}), \tag{3}$$

其中 Δ 被叫做 $\boldsymbol{x}$ 与 μ 之间的马氏距离 (Mahalanobis distance)。当 $\boldsymbol{\Sigma}$ 为单位阵时，Δ 退化为欧式距离。

不失一般性，设 $\boldsymbol{\Sigma}$ 为对称矩阵，其特征值分解为

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \tag{4}$$

其中 $\boldsymbol{U} = (\boldsymbol{u}_1, \cdots, \boldsymbol{u}_D)$ 为正交矩阵，$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_D)$。若 $\boldsymbol{\Sigma}$ 可逆，则

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T. \tag{5}$$

将(5)代入(3)，有

$$\Delta^2 = (\boldsymbol{x} - \mu)^T \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T(\boldsymbol{x} - \mu). \tag{6}$$

令

$$\boldsymbol{y} = \boldsymbol{U}^T(\boldsymbol{x} - \mu), \tag{7}$$

则

$$\Delta^2 = \boldsymbol{y}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y}.$$

高斯分布

以下考虑在新坐标系 $\boldsymbol{y}$ 下高斯分布的形式。注意到

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y} + \boldsymbol{\mu},$$

从 $\boldsymbol{x}$ 坐标系到 $\boldsymbol{y}$ 坐标系，Jacobi 矩阵为

$$\boldsymbol{J} = \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} = \boldsymbol{U}.$$

由 $\boldsymbol{U}$ 的正交性可知

$$|\boldsymbol{J}^2| = |\boldsymbol{U}^2| = |\boldsymbol{U}^T||\boldsymbol{U}| = |\boldsymbol{U}^T\boldsymbol{U}| = |\boldsymbol{I}| = 1,$$

因此 $|\boldsymbol{J}| = 1$。

又因为

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2},$$

故在 $\boldsymbol{y}$ 坐标系下，高斯分布的形式为

$$p(\boldsymbol{y}) = p(\boldsymbol{x})|\boldsymbol{J}| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} = \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \exp\left\{-\frac{1}{2}\boldsymbol{y}^T\Lambda\boldsymbol{y}\right\},$$

且它满足

$$\int p(\boldsymbol{y})\, d\boldsymbol{y} = \prod_{j=1}^{D} \int_{-\infty}^{\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\left\{-\frac{y_j^2}{2\lambda_j}\right\} dy_j = 1,$$

这表明在新坐标系 $\boldsymbol{y}$ 下，联合概率分布可分解为独立分布的乘积。

高斯分布

定理

高斯分布的数学期望与方差分别为

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}, \tag{8}$$

和

$$\mathrm{var}[\boldsymbol{x}] = \boldsymbol{\Sigma}. \tag{9}$$

高斯分布 i

证明：一方面，

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{x}] \quad &= \quad \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \boldsymbol{x}\, d\boldsymbol{x} \\
&\overset{\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu}}{=\!=\!=\!=} \quad \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}(\boldsymbol{z}+\boldsymbol{\mu})\, d\boldsymbol{x} \\
&= \quad \boldsymbol{\mu} \cdot \underbrace{\frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\} d\boldsymbol{x}}_{=1} \\
&= \quad \boldsymbol{\mu},
\end{aligned}
$$

这里用到了积分区间的对称性与被积函数 $\exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}\boldsymbol{z}$ 为奇函数的性质。

高斯分布 ii

另一方面，

$$\begin{aligned}
\mathrm{var}[\boldsymbol{x}] &= \mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right] \\
&= \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x} \\
&\xlongequal{\boldsymbol{z}=\boldsymbol{x}-\boldsymbol{\mu}} \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}\boldsymbol{z}\boldsymbol{z}^T\,d\boldsymbol{x}.
\end{aligned}$$

由 $\{\boldsymbol{u}_j\}_{j=1}^D$ 的正交性可知，$\boldsymbol{z}$ 可表示为

$$\boldsymbol{z}=\sum_{j=1}^{D} y_j\boldsymbol{u}_j,\quad y_j=(\boldsymbol{z},\boldsymbol{u}_j)$$

高斯分布 iii

从而有

$$
\begin{aligned}
\boldsymbol{z}\boldsymbol{z}^T &= \sum_{i=1}^{D} y_i \boldsymbol{u}_i \sum_{j=1}^{D} y_j \boldsymbol{u}_j^T = \sum_{i,j=1}^{D} y_i y_j \boldsymbol{u}_i \boldsymbol{u}_j^T \\
\boldsymbol{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{z} &= \sum_{i=1}^{D} y_i \boldsymbol{u}_i^T \boldsymbol{\Sigma}^{-1} \sum_{j=1}^{D} y_j \boldsymbol{u}_j = \sum_{i,j=1}^{D} y_i y_j \boldsymbol{u}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{u}_j = \sum_{i=k}^{D} \frac{y_k^2}{\lambda_k}.
\end{aligned}
$$

因此，

$$
\mathrm{var}[\boldsymbol{x}] = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \sum_{i,j=1}^{D} \int \exp\left\{ -\frac{1}{2} \sum_{k=1}^{D} \frac{y_k^2}{\lambda_k} \right\} y_i y_j \, d\boldsymbol{y} \cdot \boldsymbol{u}_i \boldsymbol{u}_j^T
$$

由对称性知，当 $i \neq j$ 时，和式中的积分为 0，故上式可简化为

$$\begin{aligned}
\mathrm{var}[\boldsymbol{x}] &= \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \sum_{i=1}^{D} \int \exp\left\{-\frac{1}{2}\sum_{k=1}^{D}\frac{y_k^2}{\lambda_k}\right\} y_i^2 \, d\boldsymbol{y} \cdot \boldsymbol{u}_i \boldsymbol{u}_i^T \\
&= \sum_{i=1}^{D} \underbrace{\frac{1}{(2\pi\lambda_i)^{1/2}} \int \exp\left\{-\frac{y_i^2}{2\lambda_i}\right\} y_i^2 \, dy_i}_{\mathrm{var}[y_i]=\lambda_i} \cdot \boldsymbol{u}_i \boldsymbol{u}_i^T \\
&= \sum_{i=1}^{D} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \boldsymbol{\Sigma}.
\end{aligned}$$

证毕。

注

高斯分布的局限

1. D 维高斯分布的协方差有 $\frac{D(D+1)}{2}$ 个参数，均值有 D 个参数，共计 $\frac{D(D+3)}{2}$ 个参数。当 D 很大时，总的参数个数以 D^2 递增，求 $\boldsymbol{\Sigma}^{-1}$ 无法进行。解决方法有：
   - 考虑对角的协方差，即
     $$\boldsymbol{\Sigma} := \mathrm{diag}(\sigma_i^2),$$
     此时，参数有 $D + D = 2D$ 个。
   - 考虑各向同性的协方差，即
     $$\boldsymbol{\Sigma} := \sigma \boldsymbol{I},$$
     此时，参数有 $D + 1$ 个。
2. 高斯分布本质上是单峰的，不能很好地近似多峰分布。

高斯分布

条件高斯分布

条件高斯分布

设 $\boldsymbol{x} \in \mathbb{R}^D$ 服从高斯分布 $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$。将 $\boldsymbol{x}$ 和对应的均值 $\boldsymbol{\mu}$ 做如下划分：

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_a \\ \boldsymbol{x}_b \end{pmatrix}, \quad \boldsymbol{x}_a \in \mathbb{R}^M, \ \boldsymbol{x}_b \in \mathbb{R}^{D-M},$$

和

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \quad \boldsymbol{\mu}_a \in \mathbb{R}^M, \ \boldsymbol{\mu}_b \in \mathbb{R}^{D-M}.$$

将协方差矩阵 $\boldsymbol{\Sigma}$ 划分为

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

由 $\boldsymbol{\Sigma}$ 的对称性知 $\boldsymbol{\Sigma}_{aa}$ 和 $\boldsymbol{\Sigma}_{bb}$ 也对称，且 $\boldsymbol{\Sigma}_{ba} = \boldsymbol{\Sigma}_{ab}^T$。

条件高斯分布

定义精度矩阵

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1},$$

并将其划分为

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}.$$

由于对称矩阵的逆仍然对称，故 $\boldsymbol{\Lambda}_{aa}$ 和 $\boldsymbol{\Lambda}_{bb}$ 也对称，且 $\boldsymbol{\Lambda}_{ba} = \boldsymbol{\Lambda}_{ab}^{T}$。

条件高斯分布

注意到一个事实：对于一个一般的高斯分布 $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$，其指数项可写成

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} - 2\boldsymbol{x}^T\boldsymbol{\Lambda}\boldsymbol{x} + C, \tag{10}$$

其中 C 为常数，且这里用到了 $\boldsymbol{\Lambda}$ 的对称性。由上式的右端可以看出，$\boldsymbol{x}$ 的二阶项系数为精度矩阵 $\boldsymbol{\Lambda}$，一阶项系数的形式为 $\boldsymbol{\Lambda}\boldsymbol{\mu}$，即精度矩阵 $\boldsymbol{\Lambda}$ 与均值 $\boldsymbol{\mu}$ 的乘积。

条件高斯分布

定理

条件高斯分布 $p(\boldsymbol{x}_a \mid \boldsymbol{x}_b)$ 的均值和协方差分别为

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_a \mid \boldsymbol{x}_b] &= \boldsymbol{\mu}_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\boldsymbol{x}_b - \boldsymbol{\mu}_b), \quad (11) \\ \operatorname{cov}[\boldsymbol{x}_a \mid \boldsymbol{x}_b] &= \Lambda_{aa}^{-1}. \quad (12) \end{aligned}$$

条件高斯分布

证明.
因

$$
\begin{aligned}
& (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) \\
= & \begin{bmatrix} (\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T & (\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \begin{bmatrix} (\boldsymbol{x}_a-\boldsymbol{\mu}_a) \\ (\boldsymbol{x}_b-\boldsymbol{\mu}_b) \end{bmatrix} \\
= & (\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{aa}(\boldsymbol{x}_a-\boldsymbol{\mu}_a) + 2(\boldsymbol{x}_a-\boldsymbol{\mu}_a)^T\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b) + (\boldsymbol{x}_b-\boldsymbol{\mu}_b)^T\boldsymbol{\Lambda}_{bb}(\boldsymbol{x}_b-\boldsymbol{\mu}_b),
\end{aligned}
\tag{13}
$$

这里用到了对称性 $\boldsymbol{\Lambda}_{ba}^T = \boldsymbol{\Lambda}_{ab}$。固定 $\boldsymbol{x}_b$，把上式看做时 $\boldsymbol{x}_a$ 的函数，则它仍是一个二次型，故对应的条件分布 $p(\boldsymbol{x}_a \mid \boldsymbol{x}_b)$ 仍是一个高斯分布。

改写(13)，可得

$$
\begin{aligned}
(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{x}-\boldsymbol{\mu}) &= \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a - 2\boldsymbol{x}_a^T[\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)] \\
&= \boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\boldsymbol{x}_a - 2\boldsymbol{x}_a^T\boldsymbol{\Lambda}_{aa}\left[\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\boldsymbol{x}_b-\boldsymbol{\mu}_b)\right],
\end{aligned}
\tag{14}
$$

对比(10)可知结论成立。 □

利用关于分块矩阵的恒等式

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix}, \tag{15}$$

其中

$$\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}. \tag{16}$$

由 $\boldsymbol{\Lambda}$ 的定义，即

$$\begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix} \tag{17}$$

利用(15)可知

$$\begin{aligned} \boldsymbol{\Lambda}_{aa} &= \left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)^{-1} \\ \boldsymbol{\Lambda}_{ab} &= -\left(\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}\right)^{-1}\boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}, \end{aligned}$$

条件高斯分布

推论

条件高斯分布 $p(\boldsymbol{x}_a \mid \boldsymbol{x}_b)$ 的均值和协方差矩阵分别为

$$\mathbb{E}[\boldsymbol{x}_a \mid \boldsymbol{x}_b] = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\boldsymbol{x}_b - \boldsymbol{\mu}_b), \tag{18}$$

$$\operatorname{var}[\boldsymbol{x}_a \mid \boldsymbol{x}_b] = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}. \tag{19}$$

条件高斯分布

注

对于条件分布 $p(\boldsymbol{x}_a \mid \boldsymbol{x}_b)$，

- 用精度矩阵表示其均值和方差更加简洁。
- 均值是 $\boldsymbol{x}_b$ 的线性函数，而方差与 $\boldsymbol{x}_b$ 无关。

高斯分布

边缘高斯分布

边缘高斯分布

定理

设 $p(\boldsymbol{x}_a, \boldsymbol{x}_b)$ 为联合高斯分布，则其边缘概率分布

$$p(\boldsymbol{x}_a) = \int p(\boldsymbol{x}_a, \boldsymbol{x}_b)\, d\boldsymbol{x}_b \tag{20}$$

也是一个高斯分布，且其均值和协方差矩阵分别为

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_a] &= \boldsymbol{\mu}_a, \\ \mathrm{var}[\boldsymbol{x}_a] &= \left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1}. \end{aligned} \tag{21}$$

边缘高斯分布

由于(20)涉及到关于 $\boldsymbol{x}_b$ 的积分，故需将(13)中涉及到 $\boldsymbol{x}_b$ 的项挑选出来，得

$$\boldsymbol{x}_b\boldsymbol{\Lambda}_{bb}\boldsymbol{x}_b - 2\boldsymbol{x}_b^T\left[\boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\right], \tag{22}$$

记

$$\boldsymbol{m} = \boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a),$$

对(22)进行配方可得

$$\left(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right)^T\boldsymbol{\Lambda}_{bb}\left(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right) + \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}.$$

边缘高斯分布

于是，(20)中关于 $\boldsymbol{x}_b$ 的积分为

$$\int \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right)^T \boldsymbol{\Lambda}_{bb}\left(\boldsymbol{x}_b - \boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right)\right\} \exp\left\{\frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right\} d\boldsymbol{x}_b \propto \exp\left\{\frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{m}\right\}$$

于是，边缘高斯分布中关于 $\boldsymbol{x}_a$ 的指数项为

$$\begin{aligned} & \frac{1}{2}\boldsymbol{m}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{m} - \frac{1}{2}\boldsymbol{x}_a^T\Lambda_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T(\Lambda_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b) \\ = \quad & \frac{1}{2}\left[\boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\right]^T \boldsymbol{\Lambda}^{-1}\left[\boldsymbol{\Lambda}_{bb} - \boldsymbol{\Lambda}_{ba}(\boldsymbol{x}_a - \boldsymbol{\mu}_a)\right] - \frac{1}{2}\boldsymbol{x}_a^T\Lambda_{aa}\boldsymbol{x}_a + \boldsymbol{x}_a^T(\Lambda_{aa}\boldsymbol{\mu}_a + \boldsymbol{\Lambda}_{ab}\boldsymbol{\mu}_b) \end{aligned}$$

整理可得

$$-\frac{1}{2}\boldsymbol{x}_a^T(\Lambda_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{x}_a + \boldsymbol{x}_a^T(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba})\boldsymbol{\mu}_a + C$$

其中 C 为与 $\boldsymbol{x}_a$ 无关常数。

边缘高斯分布

利用精度矩阵与协方差矩阵的关系

$$\begin{bmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$$

并结合(15)，可得

$$\boldsymbol{\Sigma}_{aa} = \left(\boldsymbol{\Lambda}_{aa} - \boldsymbol{\Lambda}_{ab}\boldsymbol{\Lambda}_{bb}^{-1}\boldsymbol{\Lambda}_{ba}\right)^{-1},$$

从而以下推论成立。

推论

边缘概率 $p(\boldsymbol{x}_a)$ 的均值和协方差为

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_a] &= \boldsymbol{\mu}_a, && (23) \\ \mathrm{var}[\boldsymbol{x}_a] &= \boldsymbol{\Sigma}_{aa}. && (24) \end{aligned}$$

边缘高斯分布

小结

给定一个联合高斯分布 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$，其中 $\boldsymbol{\Lambda}=\boldsymbol{\Sigma}^{-1}$，且

$$\mathbf{x}=\begin{pmatrix}\mathbf{x}_a\\ \mathbf{x}_b\end{pmatrix},\quad \boldsymbol{\mu}=\begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix},\quad \boldsymbol{\Sigma}=\begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix},\quad \boldsymbol{\Lambda}=\begin{pmatrix}\boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab}\\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb}\end{pmatrix},$$

则条件概率分布为

$$p(\mathbf{x}_a \mid \mathbf{x}_b)=\mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_{a|b},\boldsymbol{\Lambda}_{aa}^{-1}),\quad \boldsymbol{\mu}_{a|b}=\boldsymbol{\mu}_a-\boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b-\boldsymbol{\mu}_b),$$

边缘概率分布为

$$p(\mathbf{x}_a)=\mathcal{N}(\mathbf{x}_a \mid \boldsymbol{\mu}_a,\boldsymbol{\Sigma}_{aa}).$$

高斯分布

高斯变量的贝叶斯定理

高斯变量的贝叶斯定理

由之前的分析可知，边缘高斯概率分布与条件高斯概率分布可表示为

$$\begin{aligned} p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x} \mid \mu, \boldsymbol{\Lambda}^{-1}), && (25) \\ p(\boldsymbol{y} \mid \boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) && (26) \end{aligned}$$

其中 $\mu, \boldsymbol{A}$ 和 $\boldsymbol{b}$ 为控制均值的参数，$\boldsymbol{\Lambda}, \boldsymbol{L}$ 为精度矩阵。若 $\boldsymbol{x} \in \mathbb{R}^M, \boldsymbol{y} \in \mathbb{R}^D$，则 $\boldsymbol{A} \in \mathbb{R}^{D\times M}$。

高斯变量的贝叶斯定理

由之前的分析可知，边缘高斯概率分布与条件高斯概率分布可表示为

$$
\begin{aligned}
p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), &(25)\\
p(\boldsymbol{y} \mid \boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) &(26)
\end{aligned}
$$

其中 $\boldsymbol{\mu}, \boldsymbol{A}$ 和 $\boldsymbol{b}$ 为控制均值的参数，$\boldsymbol{\Lambda}, \boldsymbol{L}$ 为精度矩阵。若 $\boldsymbol{x} \in \mathbb{R}^M, \boldsymbol{y} \in \mathbb{R}^D$，则 $\boldsymbol{A} \in \mathbb{R}^{D\times M}$。

定理

给定边缘分布(25)和条件分布(26)，则联合分布 $p(\boldsymbol{x}, \boldsymbol{y})$ 也是一个高斯分布，且

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{z}] &= \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}, &(27)\\
\operatorname{cov}[\boldsymbol{z}] &= \begin{pmatrix} \boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T \end{pmatrix}. &(28)
\end{aligned}
$$

高斯变量的贝叶斯定理

证明： 定义

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix},$$

考虑联合概率分布的对数

$$\begin{aligned} \ln p(\boldsymbol{z}) &= \ln p(\boldsymbol{x}, \boldsymbol{y}) = \ln p(\boldsymbol{x}) + \ln p(\boldsymbol{y} \mid \boldsymbol{x}) \\ &= -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{x} - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{L} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + C \end{aligned} \tag{29}$$

其中 C 为与 $\boldsymbol{x}, \boldsymbol{y}$ 无关的常数。

高斯变量的贝叶斯定理

挑选出(29)中的二次项，得

$$
\begin{aligned}
& -\frac{1}{2}\left[\boldsymbol{x}^T(\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})\boldsymbol{x}^T+\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{L}\boldsymbol{A}\boldsymbol{x}-\boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{y}\right] \\
= \quad & -\frac{1}{2}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}^T\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}=-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{R}\boldsymbol{z}
\end{aligned}
$$

由此可知，$p(\boldsymbol{z})$ 的精度矩阵为

$$
\boldsymbol{R}=\begin{pmatrix}\boldsymbol{\Lambda}+\boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A} & -\boldsymbol{A}^T\boldsymbol{L}\\ -\boldsymbol{L}\boldsymbol{A} & \boldsymbol{L}\end{pmatrix} \tag{30}
$$

协方差矩阵为

$$
\mathrm{cov}[\boldsymbol{z}]=\boldsymbol{R}^{-1}=\begin{pmatrix}\boldsymbol{\Lambda}^{-1} & \boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\\ \boldsymbol{A}\boldsymbol{\Lambda}^{-1} & \boldsymbol{L}^{-1}+\boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T.\end{pmatrix}
$$

高斯变量的贝叶斯定理

挑选出(29)中的一次项，得

$$
\begin{aligned}
& \boldsymbol{x}^T \boldsymbol{\Lambda} \boldsymbol{\mu} - \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{b} + \boldsymbol{y}^T \boldsymbol{L} \boldsymbol{b} \\
= & \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda} \boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{b} \\ \boldsymbol{L} \boldsymbol{b} \end{pmatrix} \\
= & \boldsymbol{z}^T \boldsymbol{R} \boldsymbol{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda} \boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{b} \\ \boldsymbol{L} \boldsymbol{b} \end{pmatrix}
\end{aligned}
$$

由此可见，$\boldsymbol{z}$ 的均值为

$$
\mathbb{E}[\boldsymbol{z}] = \boldsymbol{R}^{-1} \begin{pmatrix} \boldsymbol{\Lambda} \boldsymbol{\mu} - \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{b} \\ \boldsymbol{L} \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{A} \boldsymbol{\mu} + \boldsymbol{b} \end{pmatrix}
$$

证毕。

高斯变量的贝叶斯定理

定理
边缘分布 $p(\boldsymbol{y})$ 的均值和协方差矩阵分别为

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{A}\mu + \boldsymbol{b}, \tag{31}$$

$$\mathrm{cov}[\boldsymbol{y}] = \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^{T}. \tag{32}$$

证明.
注意到 $\boldsymbol{y}$ 为 $\boldsymbol{z}$ 的一个分量，结论可由定理??和推论1直接推出。 □

高斯变量的贝叶斯定理

定理

条件分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$ 的均值和协方差矩阵分别为

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1} \left[\boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right], \tag{33}$$

$$\mathrm{cov}[\boldsymbol{x} \mid \boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1}. \tag{34}$$

高斯变量的贝叶斯定理

定理

条件分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$ 的均值和协方差矩阵分别为

$$\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1} \left[\boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu} \right], \tag{33}$$

$$\mathrm{cov}[\boldsymbol{x} \mid \boldsymbol{y}] = (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1}. \tag{34}$$

证明

将(12)应用于(30)即得(34)；由(11)与(27)可知

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x} \mid \boldsymbol{y}] &= \boldsymbol{\mu} + (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1} \boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b}) \\
&= (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1} \left[(\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})\boldsymbol{\mu} + \boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu} - \boldsymbol{b}) \right] \\
&= (\boldsymbol{\Lambda} + \boldsymbol{A}^T \boldsymbol{L} \boldsymbol{A})^{-1} \left[\boldsymbol{A}^T \boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda} \boldsymbol{\mu} \right]
\end{aligned}$$

证毕。

高斯变量的贝叶斯定理

这个条件分布的估计可看做是贝叶斯定理的一个例子。可把 $p(\boldsymbol{x})$ 看成 $\boldsymbol{x}$ 的先验分布，若变量 $\boldsymbol{y}$ 被观测到了，则条件分布 $p(\boldsymbol{x} \mid \boldsymbol{y})$ 表示 $\boldsymbol{x}$ 对应的后验分布。找到了边缘分布 $p(\boldsymbol{x})$ 和条件分布 $p(\boldsymbol{y} \mid \boldsymbol{x})$，就可以求出联合分布 $p(\boldsymbol{z}) = p(\boldsymbol{x})p(\boldsymbol{y} \mid \boldsymbol{x})$。

高斯变量的贝叶斯定理

小结

给定 $\boldsymbol{x}$ 的一个边缘高斯分布，及给定 $\boldsymbol{x}$ 条件下 $\boldsymbol{y}$ 的条件高斯分布，形式为

$$\begin{aligned} p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), & (35)\\ p(\boldsymbol{y} \mid \boldsymbol{x}) &= \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}), & (36) \end{aligned}$$

则 $\boldsymbol{y}$ 的边缘分布为

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^T\right), \tag{37}$$

给定条件 $\boldsymbol{y}$ 下 $\boldsymbol{x}$ 的条件分布为

$$p(\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{\Sigma}\left[\boldsymbol{A}^T\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\right], \boldsymbol{\Sigma}\right), \quad \boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^T\boldsymbol{L}\boldsymbol{A})^{-1}. \tag{38}$$

高斯分布

高斯分布的极大似然估计

高斯分布的极大似然估计

给定一个数据集 $\mathcal{D} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)^T$，假定观测 $\{\boldsymbol{x}_n \in \mathbb{R}^D\}$ 是独立地从多元高斯分布中抽取的。以下使用极大似然法来估计分布的参数。

对数似然函数为

$$\ln p(\mathcal{D} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2} \ln 2\pi - \frac{N}{2} |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}). \quad (39)$$

令

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

可得均值的极大似然估计

$$\boldsymbol{\mu}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n. \quad (40)$$

关于 $\boldsymbol{\Sigma}_{ML}$ 的推导相对来说比较麻烦，这里我们直接给出结果：

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{ML}) (\boldsymbol{x}_n - \boldsymbol{\mu}_{ML})^T. \quad (41)$$

高斯分布的极大似然估计

定理

极大似然解的期望分别为

$$\mathbb{E}\left[\boldsymbol{\mu}_{ML}\right] = \boldsymbol{\mu}, \tag{42}$$

$$\mathbb{E}\left[\boldsymbol{\Sigma}_{ML}\right] = \frac{N-1}{N}\boldsymbol{\Sigma}. \tag{43}$$

由该定理可知，均值极大似然估计的期望等于实际的均值，而协方差极大似然估计的期望是有偏的。可重新定义一个估计值来修正这个偏差：

$$\widetilde{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}\left(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML}\right)\left(\boldsymbol{x}_n - \boldsymbol{\mu}_{ML}\right)^T. \tag{44}$$

高斯分布

高斯分布的贝叶斯推断

高斯分布的贝叶斯推断

极大似然框架给出了对于参数 μ 和 $\mathbf{\Sigma}$ 的点估计，现在通过引入这些参数的先验分布，介绍一种贝叶斯方法。这一节着重考虑一元随机变量的情形。

已知方差 σ^2，推断均值

假设数据集 $\mathcal{D} = \{x_1, \cdots, x_N\}$ 是从方差为 σ^2 的高斯分布中抽样得到的，我们的目标是推断方差 μ。似然函数为

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

注意，可以把 $p(\mathcal{D} \mid \mu)$ 看作是 μ 的函数，但它不是 μ 的概率密度函数。

已知方差 σ^2，推断均值

令 μ 的共轭先验分布为

$$p(\mu) = \mathcal{N}(\mu \mid \mu_0, \sigma_0^2) = \frac{1}{(2\pi\sigma_0^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\},$$

故后验分布为

$$p(\mu \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mu)p(\mu).$$

其指数项为

$$\begin{aligned}
& -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 \\
= \quad & -\frac{1}{2}\left[\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu\right] + C \\
= \quad & -\frac{1}{2\frac{1}{\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)}}\left(\mu - \frac{N\sigma_0^2\mu_{ML} + \sigma_0^2\mu_0}{N\sigma_0^2 + \sigma^2}\right)^2 + C
\end{aligned}$$

高斯分布的贝叶斯推断

于是，后验分布的形式为

$$p(\mu \mid \mathcal{D}) = \mathcal{N}(\mu \mid \mu_N, \sigma_N^2),$$

其中

$$\begin{aligned} \mu_N &= \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}, && (45) \\ \frac{1}{\sigma_N^2} &= \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}, && (46) \end{aligned}$$

其中

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

已知方差 σ^2，推断均值

- 由(45)可知，后验分布的均值 μ_N 介于先验均值 μ_0 与极大似然解 μ_{ML} 之间。
  - 当 $N = 0$ 时，$\mu_N = \mu_0$，即后验均值退化为先验均值。
  - 当 $N \to \infty$ 时，$\mu_N = \mu_{ML}$，即后验均值由极大似然解给出。
- 由(46)可知精度可以相加，即后验分布的精度等于先验分布的精度加上每一个观测数据所贡献的精度。
  - 增加观测数据时，精度会持续增加，对应的后验分布的方差持续减少。
  - 没有观测数据点，我们有先验的方差，而如果数据点的数量 $N \to \infty$，方差 σ_N^2 趋于零，从而后验分布在在最大似然解附近变成了无限大的尖峰。

已知均值 μ，推断方差

假设数据集 $\mathcal{D} = \{x_1, \cdots, x_N\}$ 是从均值为 μ 的高斯分布中抽样得到的，我们的目标是推断均值 σ^2。定义精度

$$\tau = \sigma^{-2},$$

则 τ 的似然函数为

$$p(\mathcal{D} \mid \tau) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \tau^{-1}) \propto \tau^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{n}(x_n - \mu)\right\} \tag{47}$$

对应的共轭先验为 Gamma 分布:

$$\mathrm{Gam}(\tau \mid a, b) = \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp(-b\tau), \tag{48}$$

其均值与方差分别为

$$\mathbb{E}[\tau] = \frac{a}{b}, \tag{49}$$

$$\mathrm{var}[\tau] = \frac{a}{b^2}, \tag{50}$$

已知均值 μ，推断方差

考虑一个先验分布 $\mathrm{Gam}(\tau \mid a_0, b_0)$，乘上(47)即得后验分布

$$P(\tau \mid \mathcal{D}) \propto \tau^{\frac{N}{2}+a_0-1} \exp\left\{-b_0\tau - \frac{\tau}{2}\sum_{n=1}^{n}(x_n-\mu)\right\} \tag{51}$$

它仍是一个 Gamma 分布，其中

$$\begin{aligned} a_N &= a_0 + \frac{N}{2}, & (52) \\ b_N &= b_0 + \frac{1}{2}\sum_{n=1}^{n}(x_n-\mu) = b_0 + \frac{N}{2}\sigma_{ML}^2. & (53) \end{aligned}$$

已知均值 μ，推断方差

1. 由(52)可知，观测到 N 个数据的效果是使得 a 增加 $\frac{N}{2}$。
2. 由(53)可知，N 个观测数据对参数 b 的贡献为 $\frac{N}{2}\sigma_{ML}^2$。

高斯分布

高斯混合模型 (Mixtures of Gaussians)

高斯混合模型 (Mixtures of Gaussians)

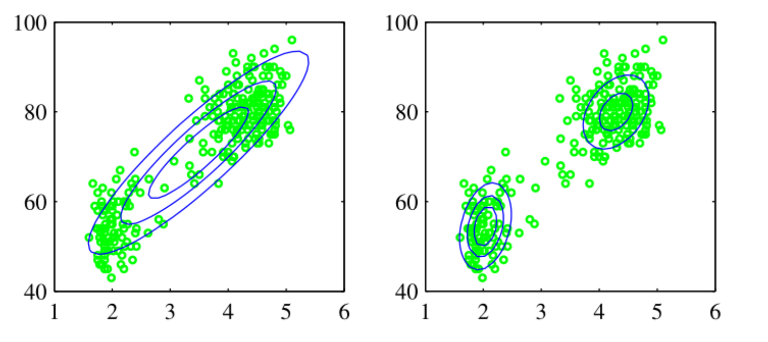

图 1: “old faithful” 数据集：由 272 次喷泉喷发的测量数据组成，其中横轴记录了喷发的持续时间（分钟），纵轴记录了距离下一次喷发的间隔时间（分钟），蓝色曲线为概率密度函数的等高线。左图为单峰高斯分布，通过极大似然估计确定分布参数。注意，该概率分布不能描述数据中的两个聚集区域，它把大部分的概率质量放在了中心区域，而这个区域的数据相对稀疏。右图使用高斯混合模型，它给出了一个更好的关于数据的描述。

高斯混合模型 (Mixtures of Gaussians)

混合分布 (mixture distributions)

通过基本的概率分布（如高斯分布）进行线性组合的叠加方法，称为混合分布。

高斯混合模型 (Mixtures of Gaussians)

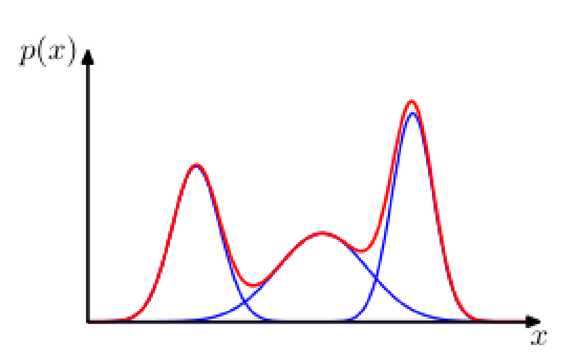

图 2: 一维高斯混合分布的例子：蓝色曲线给出了三个高斯分布，红色曲线表示它们的线性组合。

由上图可以看出，高斯分布的线性组合可以给出非常复杂的概率密度形式。通过足够多的高斯分布，并调节它们的均值、方差以及组合系数，几乎所有的连续概率密度都能以任意的精度近似。

高斯混合模型 (Mixtures of Gaussians)

以下考虑 K 个高斯概率密度的叠加，亦即混合高斯 (mixture of Gaussians)，形如

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{54}$$

其中每一个高斯概率密度 $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 被称为混合分布的一个分量 (component)，都有自己的均值 $\boldsymbol{\mu}_k$ 和协方差 $\boldsymbol{\Sigma}_k$，π_k 为混合系数 (mixture coefficients)。

高斯混合模型 (Mixtures of Gaussians)

对(54)两端积分可得

$$1 = \int p(\boldsymbol{x})\, d\boldsymbol{x} = \sum_{k=1}^{K} \pi_k \int \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\, d\boldsymbol{x} = \sum_{k=1}^{K} \pi_k,$$

即

$$\sum_{k=1}^{K} \pi_k = 1. \tag{55}$$

由 $p(\boldsymbol{x}) \geq 0$ 及 $\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geq 0$ 知

$$\pi_k \geq 0, \quad k = 1, \cdots, K.$$

联立(55)可得

$$0 \leq \pi_k \leq 1, \quad k = 1, \cdots, K, \tag{56}$$

由此可知混合系数满足概率的要求。

高斯混合模型 (Mixtures of Gaussians)

- 边缘概率 $p(\boldsymbol{x})$

  由加法原理和乘法原理知，边缘概率密度为

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k) p(\boldsymbol{x} \mid k), \tag{57}$$

  若把 $p(k)$ 看做混合高斯第 k 个分量的先验概率，$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 看做是以 k 为条件 $\boldsymbol{x}$ 的概率，则(57)等价于(54)。

高斯混合模型 (Mixtures of Gaussians)

- 边缘概率 $p(\boldsymbol{x})$

  由加法原理和乘法原理知，边缘概率密度为

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(k)p(\boldsymbol{x} \mid k), \tag{57}$$

  若把 $p(k)$ 看做混合高斯第 k 个分量的先验概率，$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 看做是以 k 为条件 $\boldsymbol{x}$ 的概率，则(57)等价于(54)。

- 后验概率 $p(k \mid \boldsymbol{x})$

  由贝叶斯公式，

$$p(k \mid \boldsymbol{x}) = \frac{p(k)\cdot(\boldsymbol{x} \mid k)}{\sum_l p(l) \cdot p(\boldsymbol{x} \mid l)} = \frac{\pi_k \cdot \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_l \pi_l \cdot p(x \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{58}$$

高斯混合模型 (Mixtures of Gaussians)

高斯混合分布由参数 $\boldsymbol{\pi}, \boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 控制，其中

$$\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_K\}, \quad \boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}, \quad \boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\Sigma}_K\}.$$

估计参数的两种方式：

- 极大似然估计(Maximum Likelihood Estimate, MLE)

  由(54)可知，高斯混合分布的对数似然函数为

  $$\ln p(\mathcal{D} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}, \tag{59}$$

  其中 $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$。这比一元高斯分布要复杂很多，参数的极大似然解不再有解析解。极大化(59)可使用数值优化方法。
- 最大化期望(Expectation Maximum, EM)

指数族分布

指数族分布

指数族分布

参数为 η，变量为 $\boldsymbol{x}$ 的指数族分布定义为

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = g(\boldsymbol{\eta}) h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\}, \tag{60}$$

其中

- $\boldsymbol{x}$ 可以为标量或向量，离散或连续；
- $\boldsymbol{\eta}$ 为分布的自然参数 (natural parameters)；
- $\boldsymbol{u}(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 的某个函数；
- $g(\boldsymbol{\eta})$ 为归一化系数，满足

$$g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} d\boldsymbol{x} = 1, \tag{61}$$

若 $\boldsymbol{x}$ 为离散变量，则上式中的积分替换为求和。

伯努利分布

伯努利分布

概率密度函数为

$$p(x \mid \mu) = \mathrm{Bern}(x \mid \mu) = \mu^{x}(1-\mu)^{1-x}, \quad x \in \{0,1\}. \tag{62}$$

也可表示成

$$p(x \mid \mu) = \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} = (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}$$

伯努利分布

与(60)进行比较可知

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

从而有

$$\mu = \frac{1}{1+\exp(-\eta)} := \sigma(\eta) \tag{63}$$

其中$\sigma(\eta)$ 被称为 logistic sigmoid 函数。容易验证

$$1-\sigma(\eta) = \sigma(-\eta), \tag{64}$$

于是伯努利分布可改写为

$$p(x \mid \mu) = \sigma(-\eta)\exp(\eta x). \tag{65}$$

再次与(60)进行比较可知

$$\begin{cases} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= \sigma(-\eta) \end{cases} \tag{66}$$

多项式分布

多项式分布

概率密度函数为

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{K} x_k \ln \mu_k\right\} \tag{67}$$

其中

$$\boldsymbol{x} = (x_1, \cdots, x_M)^T, \quad \boldsymbol{\mu} = (\mu_1, \cdots, \mu_M)^T.$$

改写为(60)的标准形式:

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \exp(\boldsymbol{\eta}^T \boldsymbol{x}) \tag{68}$$

其中

$$\boldsymbol{\eta} = (\eta_1, \cdots, \eta_M)^T, \quad \eta_k = \ln \mu_k.$$

与(60)比较可知

$$\begin{cases} \boldsymbol{u}(\boldsymbol{x}) & = & \boldsymbol{x} \\ h(\boldsymbol{x}) & = & 1 \\ g(\boldsymbol{\eta}) & = & 1. \end{cases} \tag{69}$$

由于 μ_k 满足约束

$$\sum_{k=1}^{M} \mu_k = 1, \tag{70}$$

故参数 η_k 不是相互独立的。某些情况下，去掉这个约束会比较方便，此时只用 $M-1$ 个参数就可以表示这个分布。

多项式分布

令

$$\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k,$$

则 μ_k 满足

$$0 \le \mu_k \le 1, \quad \sum_{k=1}^{M-1} \mu_k \le 1, \quad k = 1, \cdots, M-1.$$

由(70)，多项式分布可改写为

$$\begin{aligned} \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} &= \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \\ &= \exp\left\{\sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \end{aligned}$$

多项式分布

令

$$\ln\left(\frac{\mu_k}{1-\sum_{j=1}^{M-1}\mu_j}\right)=\eta_k, \quad k=1,\cdots,M-1. \tag{71}$$

可解得

$$\mu_k=\frac{\exp(\eta_k)}{1+\sum_{j=1}^{M-1}\exp(\eta_j)} \tag{72}$$

它被称为softmax 函数。

多项式分布

于是，多项式分布的形式为

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \exp(\boldsymbol{\mu}^T \boldsymbol{x}). \tag{73}$$

与(60)比较，有

$$\begin{cases} \boldsymbol{u}(\boldsymbol{x}) & = & \boldsymbol{x} \\ h(x) & = & 1 \\ g(\boldsymbol{\eta}) & = & \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} \end{cases} \tag{74}$$

其中 $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_{M-1}, 0)^T$.

高斯分布

一元高斯分布

概率密度函数为

$$
\begin{aligned}
p(x \mid \mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}\right\}
\end{aligned} \tag{75}
$$

高斯分布

与(60)做比较可得

$$\left\{\begin{array}{rcl} \boldsymbol{\eta} & = & \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \\ \boldsymbol{u}(\boldsymbol{x}) & = & \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ h(x) & = & (2\pi)^{-1/2} \\ g(\boldsymbol{\eta}) & = & (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right) \end{array}\right. \tag{76}$$

指数族分布

极大似然与充分统计量

极大似然估计

对(61)两端关于 η 求梯度，有

$$\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} d\boldsymbol{x} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x})\, d\boldsymbol{x} = 0 \tag{77}$$

极大似然估计

对(61)两端关于 η 求梯度，有

$$\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} d\boldsymbol{x} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x})\, d\boldsymbol{x} = 0 \tag{77}$$

整理可得

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x})\, d\boldsymbol{x} = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right], \tag{78}$$

即

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]. \tag{79}$$

对(61)两端关于 η 求梯度，有

$$\nabla g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\}\, d\boldsymbol{x} + g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x})\, d\boldsymbol{x} = 0 \tag{77}$$

整理可得

$$-\frac{\nabla g(\boldsymbol{\eta})}{g(\boldsymbol{\eta})} = g(\boldsymbol{\eta}) \int h(\boldsymbol{x}) \exp\left\{\boldsymbol{\eta}^T \boldsymbol{u}(\boldsymbol{x})\right\} \boldsymbol{u}(\boldsymbol{x})\, d\boldsymbol{x} = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right], \tag{78}$$

即

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}\left[\boldsymbol{u}(\boldsymbol{x})\right]. \tag{79}$$

同理，$\boldsymbol{u}(\boldsymbol{x})$ 的协方差可由 $g(\eta)$ 的二阶导数表达，高阶矩也是类似的。因此，如果能归一化指数族分布，则总能通过简单的微分求得它的各阶矩。

极大似然与充分统计量

设有一组独立同分布的数据 $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$，服从一个参数为 $\boldsymbol{\eta}$ 的指数族分布，其似然函数为

$$p(\mathcal{D} \mid \boldsymbol{\eta}) = g(\eta)^N \left(\prod_{n=1}^{N} h(\boldsymbol{x}_n) \right) \exp\left\{ \boldsymbol{\eta}^T \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n) \right\}, \tag{80}$$

故对数似然函数为

$$\ln p(\mathcal{D} \mid \boldsymbol{\eta}) = N \ln g(\eta) + \sum_{n=1}^{N} \ln h(\boldsymbol{x}_n) + \boldsymbol{\eta}^T \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n). \tag{81}$$

令 $\nabla_{\boldsymbol{\eta}} \ln p(\mathcal{D} \mid \boldsymbol{\eta}) = 0$，可得极大似然解 μ_{ML} 满足的条件为

$$-\nabla \ln g(\eta) = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n), \tag{82}$$

原则上可通过上述方程求得 $\boldsymbol{\eta}_{ML}$。

极大似然与充分统计量

由上式可以看出，极大似然解只通过 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 对数据产生依赖，故称 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 为指数族分布的充分统计量 (sufficient statistics)。编写程序时，不需要存储数据集本身，而只需要存储充分统计量即可。

极大似然与充分统计量

由上式可以看出，极大似然解只通过 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 对数据产生依赖，故称 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 为指数族分布的**充分统计量** (sufficient statistics)。编写程序时，不需要存储数据集本身，而只需要存储充分统计量即可。

例如，

- 对于伯努利分布，$u(x) = x$，只需存储数据积 $\{x_n\}$ 的和即可。
- 对于高斯分布，$\boldsymbol{u}(x) = \left(x, x^2\right)^T$，只需存储 $\{x_n\}$ 的和与 $\{x_n^2\}$ 的和。

极大似然与充分统计量

由上式可以看出，极大似然解只通过 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 对数据产生依赖，故称 $\sum_{n=1}^{N} \boldsymbol{u}(\boldsymbol{x}_n)$ 为指数族分布的充分统计量 (sufficient statistics)。编写程序时，不需要存储数据集本身，而只需要存储充分统计量即可。

例如，

- 对于伯努利分布，$u(x) = x$，只需存储数据积 $\{x_n\}$ 的和即可。
- 对于高斯分布，$\boldsymbol{u}(x) = \left(x, x^2\right)^T$，只需存储 $\{x_n\}$ 的和与 $\{x_n^2\}$ 的和。

若 $N \to \infty$，则(82)的右端变成了 $\mathbb{E}[\boldsymbol{u}(\boldsymbol{x})]$，通过与(78)可知，在此极限情况下，$\boldsymbol{\eta}_{ML}$ 等于 $\boldsymbol{\eta}$。

指数族分布

共轭先验

共轭先验

回顾一下共轭先验:

- 在伯努利分布中，共轭先验是 Beta 分布
- 在一元高斯分布中，均值的共轭先验是高斯分布，精度的共轭先验是 Gamma 分布

一般情况下，对于一个给定的概率分布 $p(\mathbf{x} \mid \boldsymbol{\mu})$，我们能寻找一个先验 $p(\boldsymbol{\eta})$，使其与似然函数共轭，从而后验分布的函数形式与先验分布相同。

共轭先验

指数族分布的共轭先验

指数族分布(60)的共轭先验为

$$p(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu \boldsymbol{\eta}^T \boldsymbol{\chi}\right\}, \tag{83}$$

其中 $f(\boldsymbol{\chi}, \nu)$ 为归一化系数。

证明.
将(83)与似然函数(80)相乘，得到后验概率（忽略归一化系数）：

$$\begin{aligned}
& p(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) p(\mathcal{D} \mid \boldsymbol{\eta}) \\
= \quad & f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu \boldsymbol{\eta}^T \boldsymbol{\chi}\right\} \cdot g(\eta)^N \left(\prod_{n=1}^N h(\boldsymbol{x}_n)\right) \exp\left\{\boldsymbol{\eta}^T \sum_{n=1}^N \boldsymbol{u}(\boldsymbol{x}_n)\right\} \\
\propto \quad & g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^T \left(\sum_{n=1}^N \boldsymbol{u}(\boldsymbol{x}_n) + \nu \boldsymbol{\chi}\right)\right\}
\end{aligned}$$

它与先验分布(83)的函数形式相同，从而证明了共轭性。 □

非参数化方法 (Nomparametric Methods)

非参数化方法 (Nomparametric Methods)

研究随机变量的过程中，随机变量的概率密度函数的作用是描述随机变量的特性。但是在实际应用中，总体概率密度函数通常是未知的，那么如何来估计总体概率密度呢？一般，我们通过抽样或者采集一定的样本，可以根据统计学知识从样本集合中推断总体概率密度。这种方法统称为概率密度估计，即根据训练样本来确定随机变量的概率分布。

非参数化方法 (Nomparametric Methods)

概率密度估计方法大致分为两类:

- 参数估计 (Parametric Estimation)
- 非参数估计 (Nonparametric Estimation)

非参数化方法 (Nomparametric Methods)

参数估计： 根据对问题的经验知识，假设问题具有某种数学模型，随机变量服从某种分布，即假定概率密度函数的形式，然后通过训练数据估计出分布函数的参数。常用的参数估计方法有

- 极大似然法
- 贝叶斯推断

非参数化方法 (Nomparametric Methods)

非参数估计 在已知样本所属的类别不假定总体分布形式下，基于大样本的性质，直接利用样本估计出整个函数。

非参数化方法 (Nomparametric Methods)

非参数估计 在已知样本所属的类别不假定总体分布形式下，基于大样本的性质，直接利用样本估计出整个函数。

在很多情况下，我们对样本的分布并没有充分的了解，无法事先给出密度函数的形式，而且有些样本分布的情况也很难用简单的函数来描述。在这种情况下，就需要用到非参数估计。但是，并不是非参数估计一定优于参数估计，因为非参数估计受训练样本影响，其完备性或者说是泛化能力不会很好；且这种估计只能用数值方法取得，无法得到完美的封闭函数图形。

非参数化方法 (Nomparametric Methods)

非参数估计 在已知样本所属的类别不假定总体分布形式下，基于大样本的性质，直接利用样本估计出整个函数。

在很多情况下，我们对样本的分布并没有充分的了解，无法事先给出密度函数的形式，而且有些样本分布的情况也很难用简单的函数来描述。在这种情况下，就需要用到非参数估计。但是，并不是非参数估计一定优于参数估计，因为非参数估计受训练样本影响，其完备性或者说是泛化能力不会很好；且这种估计只能用数值方法取得，无法得到完美的封闭函数图形。

常用的非参数估计方法有：

- 直方图法 (Histogram)
- 核密度估计 (Kernel density Estimators)
- 近邻方法 (Nearest-neighbour methods)

非参数化方法 (Nomparametric Methods)

直方图方法

直方图方法

直方图是密度函数估计中被广泛应用的一种方法。

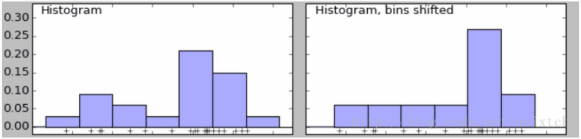

图 3: 右边直方图的划分区间比左图稍宽，但展示出的密度函数看起来却差异很大，如左图是双峰的，右图是单峰的。

直方图方法

直方图是密度函数估计中被广泛应用的一种方法。

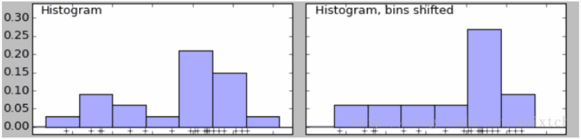

图 3: 右边直方图的划分区间比左图稍宽，但展示出的密度函数看起来却差异很大，如左图是双峰的，右图是单峰的。

该方法简单易懂，但其缺点也很明显：

- 密度函数不连续
- 密度函数受划分区间的宽度影响很大。同样的原始数据如果划分区间取不同宽度，那么展示的结果可能完全不同（见上图）。
- 直方图最多只能展示二维数据，若维度更多则无法有效展示。

非参数化方法 (Nomparametric Methods)

核密度估计

假设观测值 $\boldsymbol{x}$ 服从 D 维空间的某个未知的概率密度分布 $p(\boldsymbol{x})$。考虑包含 $\boldsymbol{x}$ 的某个小区域 $\mathcal{R}$，则 $\boldsymbol{x}$ 落入该区域的概率为：

$$P = \int_{\mathcal{R}} p(\boldsymbol{x}) \, d\boldsymbol{x}$$

假设 n 个样本 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ 是根据概率密度函数 $p(\boldsymbol{x})$ 独立同分布抽取而得到的，k 个样本落在区域 $\mathcal{R}$ 中的概率服从二项分布：

$$\mathrm{Bin}(k \mid n, P) = \frac{n!}{k!(n-k)!} P^k (1-P)^{n-k}$$

由二项分布的性质可知

$$\mathbb{E}[k] = nP, \tag{84}$$

故 k/n 就是对 P 的一个很好的估计。

核密度估计

假设 $p(\boldsymbol{x})$ 连续，并且区域 $\mathcal{R}$ 充分小，以至于在这个区域中 p 几乎没有变化，则

$$\int_{\mathcal{R}} p(\boldsymbol{x})\, d\boldsymbol{x} = p(\boldsymbol{x})V \tag{85}$$

其中 V 是区域 $\mathcal{R}$ 的体积。综上，可得 $p(\boldsymbol{x})$ 的一个估计：

$$p(\boldsymbol{x}) \approx \frac{k}{nV}. \tag{86}$$

核密度估计

为估计点 $\boldsymbol{x}$ 处的概率密度函数，构造一系列包含点 $\boldsymbol{x}$ 的区域 $\mathcal{R}_1, \cdots, \mathcal{R}_n$，$V_n$ 为 $\mathcal{R}_n$ 的体积，k_n 为落在区间 $\mathcal{R}_n$ 的样本个数，$p_n(\boldsymbol{x})$ 表示对 $p(\boldsymbol{x})$ 的第 n 次估计，即

$$p_n(\boldsymbol{x}) = \frac{k_n/n}{V_n} \tag{87}$$

$p_n(x)$ 收敛到 $p(\boldsymbol{x})$ 的条件为：

$$\begin{aligned}
\lim_{n\to\infty} V_n &= 0, \\
\lim_{n\to\infty} k_n &= \infty, \\
\lim_{n\to\infty} k_n/n &= 0.
\end{aligned} \tag{88}$$

核密度估计

由此可知，推导(86)的两个假设相互矛盾，即：一方面，$\mathcal{R}$ 要充分小，以使得该区域内的概率密度近似为常数；另一方面，$\mathcal{R}$ 也要足够大，以使得落在该区域内的样本个数 k 能让二项分布达到尖峰。有两种方式利用(86)的结果：

- 固定 k，从数据中确定 V，这就是k 近邻方法。
- 固定 V，从数据中确定 k，这就是核方法。

核密度估计：Parzen 窗口

将区域 $\mathcal{R}$ 取成以 $\boldsymbol{x}$ 为中心的超立方体，以此来确定概率密度。为统计落在该区域内的样本点个数 k，定义函数

$$\kappa(\boldsymbol{u}) = \begin{cases} 1, & |u_i| \leq \frac{1}{2}, \quad i = 1, \cdots, D \\ 0, & \text{otherwise}, \end{cases} \tag{89}$$

它表示一个以原点为中心的单位立方体，通常称它为 Parzen 窗 (Parzen window)。

核密度估计：Parzen 窗口

记 $\Omega_{\boldsymbol{x},h}$ 为以 $\boldsymbol{x}$ 为中心，边长为 h 的超立方体，则有

$$\kappa\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right)=\begin{cases}1, & \boldsymbol{x}_i\in\Omega_{\boldsymbol{x},h},\\ 0, & \boldsymbol{x}_i\notin\Omega_{\boldsymbol{x},h}.\end{cases} \tag{90}$$

于是落在 $\Omega_{\boldsymbol{x},h}$ 的样本点个数为

$$k=\sum_{i=1}^{n}\kappa\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right)$$

将其代入(86)即得 $\boldsymbol{x}$ 处的概率密度估计

$$p(\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^D}\kappa\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right), \tag{91}$$

这里用到了 $V=h^D$，即边长为 h 的 D 维超立方体。利用 $\kappa(\boldsymbol{u})$ 的对称性，可以重新描述(91)：前面我们把(91)描述为一个以 $\boldsymbol{x}$ 为中心的超立方体，现在可把(91)表述为以 $\boldsymbol{x}_i$ 为中心的 n 个超立方体。

核密度估计：高斯核函数

核密度估计(91)有一个问题，即人为引入的非连续性。若选择一个平滑的核函数，则可得到一个更加平滑的模型。如高斯核函数：

$$p(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(2\pi h^2)^{D/2}}\exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2h^2},\right\} \tag{92}$$

其中 h 表示高斯分布的标准差。

核密度估计：高斯核函数

概率密度模型(92)可通过如下方式获得：令每个样本点都服从高斯分布，然后把每个样本点的贡献相加，再求平均，使得概率密度能被正确地归一化。

参数 h 对平滑起着重要的作用：小的 h 使模型对噪声过于敏感，而大的 h 使得模型过度平滑。

核密度估计：一般的核函数

选择核函数时，只要求它满足下面的条件

$$
\begin{aligned}
\kappa(\boldsymbol{u}) &\geq 0 && (93)\\
\int \kappa(\boldsymbol{u})\, d\boldsymbol{u} &= 1. && (94)
\end{aligned}
$$

这确保了最终求得的概率密度处处非负，且积分等于 1。

非参数化方法 (Nomparametric Methods)

近邻方法

近邻方法

用核方法进行概率密度估计时，控制核宽度的参数 h 是固定的。在高密度的区域，大的 h 值可能会造成过度平滑，并且破坏了本应从数据中提取出的结构。而减小 h 的值可能导致数据空间中低密度区域估计的噪声。因此，h 的最优选择依赖于数据空间的位置，这个问题可以通过概率密度的近邻方法解决。

近邻方法

与之前固定 V 然后从数据中确定 k 不同，这里考虑固定 k 然后使用数据来确定合适的 V。为此，考虑一个以 $\boldsymbol{x}$ 为中心的球，来估计概率密度 $p(\boldsymbol{x})$，这里允许球的半径可以自由增长，直到它精确地包含 k 个数据点。这样，概率密度 $p(x)$ 可由

$$p(\boldsymbol{x}) \approx \frac{k}{nV}$$

估计，其中 V 为最终球体的体积。这种方法称为k 近邻方法。

近邻方法

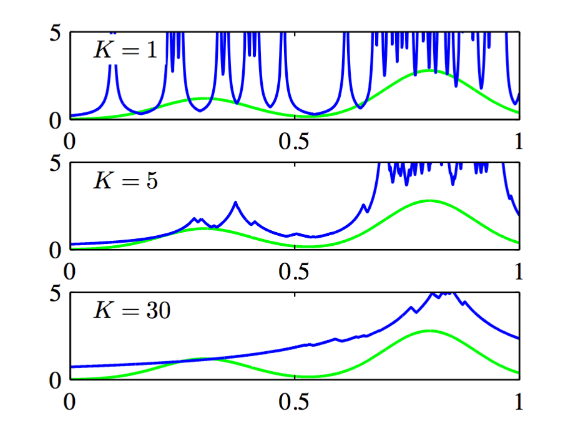

图 4: k 近邻密度估计：参数 k 控制平滑程度，小的 k 值会产生一个噪声很大的密度模型（上图），而一个大的 k（下图）平滑掉了真实概率密度（绿色曲线）的双峰性质。

k 近邻方法用于分类

工作原理

假设有一个样本数据集（也称训练集），其中每个数据都有已知的标签（即类别）。输入没有标签的数据（也称测试数据）后，将测试数据中的特征与训练集中数据对应的特征进行比较，提取出样本集中特征最相似数据（最近邻）的分类标签。

k 近邻方法用于分类

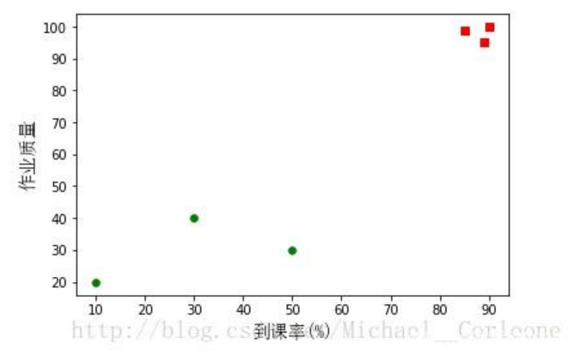

图 5: 图中每个点代表一个样本（即学生），横纵坐标代表了特征（到课率，作业质量），不同的形状代表了类别（即：红色代表成绩 A，绿色代表成绩 D）。

k 近邻方法用于分类

问题

如果一个学生想知道自己能考的怎么样，他在老师那里查到了自己的到课率 85%，作业质量是 90，那么怎么实现预测呢？

k 近邻方法用于分类

问题

如果一个学生想知道自己能考的怎么样，他在老师那里查到了自己的到课率 85%，作业质量是 90，那么怎么实现预测呢？

可以把某学生（张三）看做是点 $(85, 90)$，即他的到课率为 85%、作业质量为 90，称它为测试样本。

k 近邻方法用于分类

问题

如果一个学生想知道自己能考的怎么样，他在老师那里查到了自己的到课率 85%，作业质量是 90，那么怎么实现预测呢？

可以把某学生（张三）看做是点 $(85, 90)$，即他的到课率为 85%、作业质量为 90，称它为测试样本。

- 首先计算张三到其他 6 位同学（训练样本）的距离（一般采用欧氏距离）；
- 然后选取前 k 个最近的距离（如 $k = 3$）；
- 接着找出距离最近的三个样本分别属于哪个类别（都为 A）；
- 最后挑选三个类别中最多的类别作为预测结果。

于是可预测出张三的期末成绩可能是 A。

k 近邻方法用于分类

kNN 算法的流程

1. 计算测试数据与各个训练数据之间的距离；
2. 按照距离从小到大进行排序；
3. 选取距离最小的 k 个点；
4. 确定前 k 个点所在类别的出现频率；
5. 返回前 k 个点中出现频率最高的类别作为测试数据的预测分类。

k 近邻方法用于分类

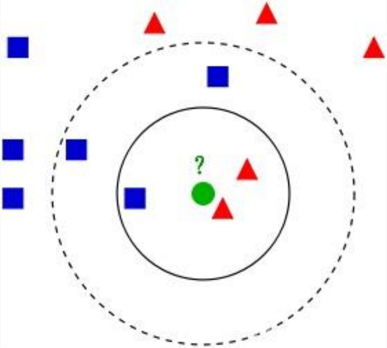

图 6: k 值对分类结果的影响：蓝色点和红色点为训练数据，绿色点为测试数据。当 $k=3$ 时，范围内红色三角形多，待分类点属于红色三角形；当 $k=5$ 时，范围内蓝色正方形多，待分类点属于蓝色正方形。

k 近邻方法用于分类

如何选择一个最佳的 k 值取决于数据。一般情况下，在分类时较大的 k 值能够减小噪声的影响，但会使类别之间的界限变得模糊。因此 k 的取值一般比较小 ($k < 20$)。

k 近邻方法用于分类

设有一个训练数据集，其中 n_i 个数据点属于类别 $\mathcal{C}_i$，样本总数为 n，故 $\sum_k n_k = n$。如果想对一个新的样本 $\boldsymbol{x}$ 进行分类，可以画一个以 $\boldsymbol{x}$ 为中心的球，包含 k 个数据点 (不论类别)。设球的体积为 V，并且包含来自类别 $\mathcal{C}_i$ 的 k_i 个数据点，由(86)可知与每个类别关联的概率密度估计为

$$p(\boldsymbol{x} \mid \mathcal{C}_i) = \frac{k_i}{n_i V}$$

无条件概率密度为

$$p(\boldsymbol{x}) = \frac{k}{nV}$$

类别的先验概率为

$$p(\mathcal{C}_i) = \frac{n_i}{n}$$

由贝叶斯公式可得后验概率

$$p(\mathcal{C}_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \mathcal{C}_i) p(\mathcal{C}_i)}{p(\boldsymbol{x})} = \frac{k_i}{k}$$

k 近邻方法用于分类

若想最小化错误分类的概率，可把测试样本 $\boldsymbol{x}$ 分配给有着最大后验概率的类别，这对应于最大的 $\frac{k_i}{k}$。因此，为分类一个新的样本，可从训练数据中选择 k 个最近的数据点，然后把新样本分配给这个集合中数量最多的样本的类别。当 $k = 1$ 时，称为最近邻方法，因为测试样本被简单地分类为训练集中距离最近的样本的类别。