# Section 2 Math Basics

## 高斯分布

在机器学习中很重要.

例如 Linear Gaussian Model

$$Y = \underbrace{AX}_{\text{线性变换}} + B + \varepsilon$$

△从高斯分布中采样得到

**多维高斯分布的定义:**

① 若 p 维随机变量 $X=(X_1, X_2, \cdots, X_p)^T$ 的密度函数为

$$f(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

其中 $x=(x_1, x_2, \cdots, x_p)^T$. $\mu$ 是 p 维向量, $\Sigma$ 是 p 阶半正定阵.

则 $X \sim N_p(\mu, \Sigma)$.

但是当 $|\Sigma|=0$ 时, $\Sigma^{-1}$ 不存在, 上述定义失效, 则有定义②

② 设 $X_1, X_2, \cdots, X_p$ 为 p 个独立的标准正态随机变量.

$$令\ Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix} = A_{m\times p} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} + \mu_{m\times 1}$$

则称 $Y$ 为 $m$ 维正态随机向量，记 $Y \sim N_m(\mu, \Sigma)$

其中 $E(Y)=\mu$，$Var(Y)=AA^T$，显然，$\Sigma=AA^T$ 分解一般不唯一

这个定义的优点在于，表明了多元正态分布是由多个独立的标准正态分布线性组合得到的，所以多元正态分布的一些性质可以由一元正态分布得到。

## 多元高斯分布的性质：

① 设 $X=(X_1, X_2, \cdots, X_p)^T \sim N_p(\mu, \Sigma)$，若 $\Sigma$ 为对角阵，则 $X_1, X_2, \cdots, X_p$ 相互独立

② 设 $X \sim N_p(\mu, \Sigma)$，$A \in \mathbb{R}^{s \times p}$，$d \in \mathbb{R}^{s \times 1}$，则

$$AX+d \sim N_s(A\mu+d, A\Sigma A^T)$$

即正态随机变量的线性变换仍然是正态的。

③ 设 $X \sim N_p(\mu, \Sigma)$，将 $X, \mu, \Sigma$ 作如下剖分：

$$X=\begin{array}{cc}\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} & \begin{array}{c} q \\ p-q \end{array}\end{array}, \quad \mu=\begin{array}{cc}\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} & \begin{array}{c} q \\ p-q \end{array}\end{array}, \quad \Sigma=\begin{array}{cc}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} & \begin{array}{c} q \\ p-q \end{array}\end{array}$$

则 $X^{(1)} \sim N_q(\mu^{(1)}, \Sigma_{11})$，$X^{(2)} \sim N_{p-q}(\mu^{(2)}, \Sigma_{22})$

$$\Sigma_{12}=Cov(X^{(1)}, X^{(2)})$$

注意：

(1) 多元正态分布的任意元边际分布仍为正态分布，反之则不真。

(2) $\Sigma_{12}=Cov(X^{(1)}, X^{(2)})$，$\Sigma_{12}=0$ 表示 $X^{(1)}, X^{(2)}$ 不相关，由定义可知：此时 $X^{(1)}$ 与 $X^{(2)}$ 独立，故对于多元正态分布来说不相关与独立是等价的。

## 参数估计：

Data: $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}_{N\times p}$. 其中N表示样本数，p表示样本维度

$x_i \in \mathbb{R}^p$

$x_i \overset{\text{i.i.d.}}{\sim} N_p(\mu, \Sigma)$

$\theta = (\mu, \Sigma)$.

极大似然估计 (MLE)：

假设 $p=1$ (一维). $\theta = (\mu, \sigma^2)$, $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$.

$$\theta = \arg\max_\theta P(X|\theta)$$
$$= \arg\max_\theta \log P(X|\theta)$$
$$= \arg\max_\theta \sum_{i=1}^N \log P(x_i|\theta)$$
$$= \arg\max_\theta \sum_{i=1}^N \left(-\log\sqrt{2\pi} - \log\sigma - \frac{(x-\mu)^2}{2\sigma^2}\right).$$

$$\mu_{MLE} = \arg\max_\mu \sum_{i=1}^N \left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$
$$= \arg\min_\mu \sum_{i=1}^N (x_i-\mu)^2.$$

$$\frac{\partial}{\partial\mu}\sum_{i=1}^N (x_i-\mu)^2 = -\sum_{i=1}^N 2(x_i-\mu) = 0$$

$$\sum_{i=1}^N x_i - \sum_{i=1}^N \mu = 0$$

$$\boxed{\Rightarrow \mu_{MLE} = \frac{1}{N}\sum_{i=1}^N x_i}$$ → 无偏估计.

$$\sigma^2_{MLE}=\underset{\sigma^2}{\arg\max}\sum_{i=1}^{N}\left(-\frac{1}{2}\log\sigma^2-\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

$$=\underset{\sigma^2}{\arg\min}\sum_{i=1}^{N}\left(\frac{1}{2}\log\sigma^2+\frac{(x_i-\mu)^2}{2\sigma^2}\right).$$

$$\frac{\partial}{\partial\sigma^2}\sum_{i=1}^{N}\left(\frac{1}{2}\log\sigma^2+\frac{(x_i-\mu)^2}{2\sigma^2}\right)=\sum_{i=1}^{N}\left(\frac{1}{2}\cdot\frac{1}{\sigma^2}-\frac{(x_i-\mu)^2}{2\sigma^4}\right)=0.$$

$$\sum_{i=1}^{N}\left(\frac{1}{\sigma^2}-\frac{(x_i-\mu)^2}{\sigma^4}\right)=0$$

$$\sum_{i=1}^{N}\left(\sigma^2-(x_i-\mu)^2\right)=0$$

$$\boxed{\Rightarrow \sigma^2_{MLE}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2,\ \text{其中}\ \mu=\mu_{MLE}}$$

$\downarrow$

$DX=EX^2-(EX)^2$

有偏估计. $E[\sigma^2_{MLE}]=\frac{1}{N}\sum_{i=1}^{N}E[(x_i-\mu)^2]$

$$=\frac{1}{N}\sum_{i=1}^{N}D[x_i-\mu]+\left(E[x_i-\mu]\right)^2$$

$$D[x_i-\mu]=D[x_i-\tfrac{1}{N}\textstyle\sum x_i].$$

$$=D[\tfrac{N-1}{N}x_i-\tfrac{1}{N}\textstyle\sum x_j]$$

$$=\frac{(N-1)^2}{N^2}\sigma^2+\frac{N-1}{N^2}\sigma^2$$

$$=\frac{N-1}{N}\sigma^2$$

主要由于 $\mu$ 是 $\bar{X}$，也就是 $\mu_{MLE}$，而不是 $\mu$ 造成了有偏

$$E[x_i-\mu]=Ex_i-E\mu=0.$$

则 $E[\sigma^2_{MLE}]=\frac{N-1}{N}\sigma^2$

显然，将 $\sigma^2_{MLE}$ 乘上 $\frac{N}{N-1}$，就能无偏

即：$\hat{\sigma}^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2$ 是无偏估计

# 概率密度函数

如前面定义所示:

$$f(x_1, x_2, \cdots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} \underbrace{(x-\mu)^T \Sigma^{-1} (x-\mu)}_{\text{二次型}}\right\}$$

$x$ 和 $\mu$ 之间的马氏距离.

马氏距离: $D_M(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$

单个数据点的马氏距离:

$D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)}$

上面两式中, $\Sigma$ 为协方差阵, $\mu$ 为均值

可见当 $\Sigma = I$ 时, 马氏距离和欧氏距离等价

具体内容参见以下链接

https://zhuanlan.zhihu.com/p/46626607

基于马氏距离: 可以进行以下推导.

$\Sigma = U \Lambda U^T$, 其中 $UU^T = U^T U = I$, $\Lambda = \underset{i=1,\cdots,p}{\operatorname{diag}}(\lambda_i)$.

则 $$\Sigma = (u_1, u_2, \cdots, u_p) \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{pmatrix} (u_1, u_2, \cdots, u_p)^T$$

$$= \sum_{i=1}^{p} u_i \lambda_i u_i^T$$

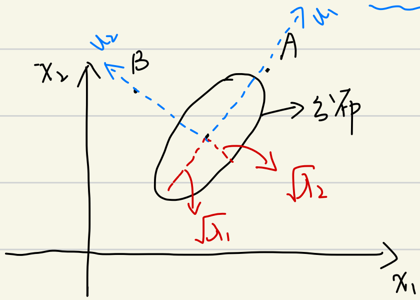
所以, $\Sigma^{-1} = (U \Lambda U^T)^{-1} = (U^T)^{-1} \Lambda^{-1} U^{-1} = U \Lambda^{-1} U^T = \sum_{i=1}^{p} u_i \frac{1}{\lambda_i} u_i^T$

则 $$(x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T\left(\sum_{i=1}^{p} u_i \frac{1}{\lambda_i} u_i^T\right)(x-\mu)$$
$$= \sum_{i=1}^{p} (x-\mu)^T u_i \frac{1}{\lambda_i} u_i^T (x-\mu)$$

令 $y_i = (x-\mu)^T u_i$ ，$Y = (x-\mu)^T U$

则 $$(x-\mu)^T\Sigma^{-1}(x-\mu) = \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}$$
$$(x-\mu)^T\Sigma^{-1}(x-\mu) = \underbrace{(x-\mu)^T U \Lambda^{-1} U^T (x-\mu)} = Y^T \Lambda^{-1} Y$$

将 $(x-\mu)$ 映射到 $u_1, u_2, \cdots, u_p$




如图，如果要判断A和B属于分布的概率，显然，A比B更符合该分布，但是如果只求 $x-\mu$，那么A和B的概率一样。
马氏距离就是将 $x_1, x_2$ 变换到 $u_1, u_2$，
然后再沿 $u_1, u_2$ 按特征值缩放。
最后再去评价距离。

这个 $\lambda_1$ 和 $\lambda_2$ 与 $\sigma_1^2$ 和 $\sigma_2^2$ 有关系吗

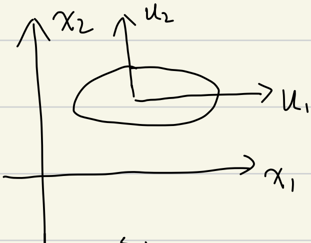
所以，多维高斯分布的概率密度相当于将协方差考虑在内，使得各维独立，再根据特征值缩放，来衡量概率密度大小。

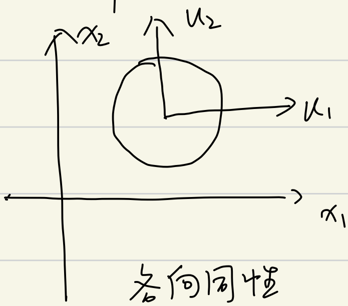
## 局限性：

① 参数过多

$\Sigma$ 如果是 $p$ 维，则参数个数：$\frac{p^2-p}{2}+p=\frac{p(p+1)}{2}$，是 $O(p^2)$ 级别的

一般来说，假设 $\Sigma$ 为对角矩阵，甚至是各向同性的.




例子：factor analysis $\longrightarrow$ $\Sigma$ 对角矩阵

P-PCA $\longrightarrow$ 各向同性.

② 一般来说，一个高斯分布无法表示模型

例子、GMM $\longrightarrow$ 多个高斯分布

## 边缘分布与条件概率分布

已知：$x=\begin{pmatrix} x_a \\ x_b \end{pmatrix}\begin{matrix} m \\ n \end{matrix}$，$m+n=p$，$\mu=\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$，$\Sigma=\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$.

求 $p(x_a)$，$p(x_b|x_a)$　　PRML中用配方法求解.

$p(x_b)$，$p(x_a|x_b)$

**性质② 证明**：$Y=AX+b$.

$$E[Y]=E[AX+b]=AE[X]+b=A\mu+b$$

$$Var[Y]=Var[AX+b]=Var[AX]=AVar[X]A^T=A\Sigma A^T$$

**性质③ 证明**：

由性质②.得：$x_a=(I_m, O_n)\begin{pmatrix} x_a \\ x_b \end{pmatrix}+0$

$$E[x_a]=(I_m, O_n)\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}+0=\mu_a$$

$$Var[x_a]=(I_m, O_n)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m^T \\ O_n^T \end{pmatrix}=\Sigma_{aa}$$

则 $x_a \sim N(\mu_a, \Sigma_{aa})$

$x_a$ 的边缘分布求解完成.

## $\boxed{\text{求条件概率密度}}$

设 $x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a$

则由性质②，得：

$$x_{b\cdot a} = (-\Sigma_{ba}\Sigma_{aa}^{-1}, I_n)\begin{pmatrix}x_a\\ x_b\end{pmatrix}$$

$$\mu_{b\cdot a} = E[x_{b\cdot a}] = \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a$$

$$\Sigma_{bb\cdot a} = Var[x_{b\cdot a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}, I_n)\begin{pmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}\begin{pmatrix}-(\Sigma_{aa}^{-1})^T\Sigma_{ba}^T\\ I_n\end{pmatrix}$$

$$= (0, \Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab})\begin{pmatrix}-(\Sigma_{aa}^{-1})^T\Sigma_{ba}^T\\ I_n\end{pmatrix}$$

$$= \Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}.$$

$\Sigma_{aa}$ 的 schur complement

有关 schur complement 的相关内容见以下链接.

https://blog.csdn.net/sheagu/article/details/115771184

则 $x_{b\cdot a} \sim N(\mu_{b\cdot a}, \Sigma_{bb\cdot a})$

$\therefore\ x_b = x_{b\cdot a} + \underbrace{\Sigma_{ba}\Sigma_{aa}^{-1}x_a}_{b}$.

这里还是套用 $Y=AX+b$.
则 $Y=x_b$, $A=I$, $X=x_{b\cdot a}$, 后半部分为 $b$

由性质①，得：

$$E[x_b|x_a] = \mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$$

$$Var[x_b|x_a] = Var[x_{b\cdot a}] = \Sigma_{bb\cdot a}.$$

则 $x_b|x_a \sim N(\mu_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a, \Sigma_{bb\cdot a})$

边缘概率密度求解完毕.

$x_b$ 和 $x_a|x_b$ 的求解方式也一样

上述过程中，有一步缺少证明：

即 $x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a$

应该写为 $x_b|x_a = x_{b\cdot a}|x_a + \Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a$

需要证明 $x_{b\cdot a}$ 与 $x_a$ 独立.

定理：若 $x\sim N(\mu,\Sigma)$，则 $Mx\perp Nx \iff M\Sigma N^T=0$.

证明：

$$
\begin{aligned}
Cov(Mx, Nx) &= E[(Mx-E[Mx])(Nx-E[Nx])^T] \\
&= E[(Mx-M\mu)(Nx-N\mu)^T] \\
&= E[M(x-\mu)(x-\mu)^T N^T] \\
&= M E[(x-\mu)(x-\mu)^T] N^T \\
&= M\Sigma N^T
\end{aligned}
$$

由于 $Mx\perp Nx \iff Cov(Mx, Nx)=0$

则 $Mx\perp Nx \iff M\Sigma N^T = 0$.

代入上式：$x_{b\cdot a} = \underbrace{\begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix}}_{M} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{x}$

$$x_a = \underbrace{\begin{pmatrix} I_m, & 0 \end{pmatrix}}_{N} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{x}.$$

$$
\begin{aligned}
\Rightarrow M\Sigma N^T &= \left(-\Sigma_{ba}\Sigma_{aa}^{-1}, I_n\right)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m \\ 0 \end{pmatrix} \\
&= \left(0,\ \Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)\begin{pmatrix} I_m \\ 0 \end{pmatrix} = 0.
\end{aligned}
$$

则 $x_{b\cdot a}$ 与 $x_a$ 独立，即 $x_{b\cdot a}|x_a = x_{b\cdot a}$.

对于 $\Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a$，可理解如下：

求期望时，$E[\Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a]$

$$= \Sigma_{ba}\Sigma_{aa}^{-1}E[x_a|x_a]$$

$$= \Sigma_{ba}\Sigma_{aa}^{-1}\underbrace{\sum p(x_a|x_a)\cdot x_a}_{x_a}$$

$$= \Sigma_{ba}\Sigma_{aa}^{-1}x_a.$$

求方差时，$Var[\Sigma_{ba}\Sigma_{aa}^{-1}x_a|x_a]$

$$= \Sigma_{ba}\Sigma_{aa}^{-1}(\Sigma_{ba}\Sigma_{aa}^{-1})^T Var[x_a|x_a]$$

$$= \Sigma_{ba}\Sigma_{aa}^{-1}(\Sigma_{ba}\Sigma_{aa}^{-1})^T \underbrace{E[(x_a-x_a|x_a)(x_a-x_a|x_a)^T]}_{0}$$

$$= 0$$

上述为个人理解，可能有误.

$\boxed{\text{已知 } x,\ y|x,\ \text{求 } y,\ x|y}$

已知 $p(x) = N(x|\mu, \Lambda^{-1})$ → precision matrix → (covariance matrix)$^{-1}$ 精度矩阵就是协方差矩阵的逆

$p(y|x) = N(y|Ax+b, L^{-1})$

求 $p(y)$, $p(x|y)$

$\boxed{\text{求解 } p(y)}$

由已知: $y = Ax+b+\varepsilon$, 其中 $\varepsilon \sim N(0, L^{-1})$ 且 $\varepsilon \perp A$ $\varepsilon$和$A$独立

$y, x, \varepsilon$ 都是 random variable, $A$和$b$看作系数

则由性质②:

$$E[y] = E[Ax+b+\varepsilon] = A\mu+b.$$

$$Var[y] = Var[Ax+b+\varepsilon] = A\Lambda^{-1}A^T + L^{-1}$$

则 $y \sim N(A\mu+b,\ A\Lambda^{-1}A^T+L^{-1})$, $p(y)$求解完毕.

$\boxed{\text{求解 } p(x|y)}$

构造 $z = \begin{pmatrix} x \\ y \end{pmatrix}$, 则 $\mu_z = \begin{pmatrix} \mu \\ A\mu+b \end{pmatrix}$, $\Sigma_z = \begin{pmatrix} \Lambda^{-1} & \Delta \\ \Delta^T & L^{-1}+A\Lambda^{-1}A^T \end{pmatrix}$

$$\begin{aligned}
\Delta = Cov(x,y) &= E[(x-E[x])(y-E[y])^T] \\
&= E[(x-\mu)(Ax+b+\varepsilon-A\mu-b)^T] \\
&= E[(x-\mu)(Ax-A\mu+\varepsilon)^T] \\
&= E[(x-\mu)(Ax-A\mu)^T + (x-\mu)\varepsilon^T] \\
&= E[(x-\mu)(Ax-A\mu)^T] + E[(x-\mu)\varepsilon^T]
\end{aligned}$$

其中，由于 $x \perp \varepsilon$，则 $E[(x-\mu)\varepsilon^T] = E[\underbrace{(x-\mu)}_{0}]E[\varepsilon^T] = 0$

$$
\begin{aligned}
\text{则 } \Delta &= E[(x-\mu)(Ax-A\mu)^T] \\
&= E[(x-\mu)(x-\mu)^T A^T] \\
&= E[(x-\mu)(x-\mu)^T] A^T \\
&= Var[x] A^T \\
&= \Lambda^{-1} A^T
\end{aligned}
$$

$$
\therefore z \sim N\left( \begin{bmatrix} \mu \\ A\mu+b \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1}+A\Lambda^{-1}A^T \end{bmatrix} \right)
$$

沿用上一节求解条件概率的方法，即可求得 $p(x|y)$.

构造 $x_{b \cdot a}$，求条件概率