淡蓝小点技术系列：PRML重点问题选讲（Draft）

淡蓝小点Bluedotdot

微信：bluedotdot_cn

2024 年 5 月 30 日

PRML Chapter1 要点墙

1-001	1-002	1-003	1-004	1-005	1-006	1-007	1-008	1-009	1-010
1-011	1-012	1-013	1-014	1-015	1-016	1-017	1-018	1-019	1-020
1-021	1-022	1-023	1-024	1-025	1-026	1-027	1-028	1-029	1-030
1-031	1-032	1-032+	1-033	1-034	1-035	1-036	1-037	1-038	1-039
1-040	1-041	1-042	1-043	1-044	1-045	1-046	1-047	1-048	1-049
1-050	1-051	1-052	1-053	1-054	1-055	1-056	1-057	1-058	1-059
1-060	1-061	1-062	1-063	1-064	1-065	1-066	1-067	1-068	1-069
1-070									

设$p_x(x)$是随机变量x的概率密度函数，现有双射变换g使得$x = g(y)$，求y的概率密度函数$p_y(y)$。

错误思路：将$g(y)$代入$p_x(x)$并替换所有的x得到$p_x(g(y))$，它实际是关于y的函数，也就是$p_y(y)$。
设x在$(0,1)$的范围内均匀分布，所以$p_x(x) = 1$。令$x = 2y$，也就是$x = g(y) = 2y$。如果直接替换的话有

$$p_y(y) = p_x(g(y)) = p_x(2y) = 1$$

注意，$p_x(x) = 1$是个常函数，跟变量的变化无关，所以变换后仍有$p_x(2y) = 1$。因为$x \in (0,1)$所以自然有$y \in (0, \frac{1}{2})$。所以上式积分有

$$\int p_y(y)\mathrm{d}y = \int_0^{\frac{1}{2}} 1\mathrm{d}y = y|_0^{\frac{1}{2}} = \frac{1}{2}$$

可见，积分并不等于1，所以利用这种方法得到的$p_y(y)$不是合法的概率密度。

概率密度函数正确的变换方式是[1]：

$$p_y(y) = p_x(x)|\frac{\mathrm{d}x}{\mathrm{d}y}| = p_x(g(y))|g'(y)|$$

根据这一式子前面例子有

$$p_y(y) = p_x(g(y))|\frac{\mathrm{d}x}{\mathrm{d}y}| = 1 * 2 = 2$$

所以积分为

$$\int p_y(y)\mathrm{d}y = \int_0^{\frac{1}{2}} 2\mathrm{d}y = 2y|_0^{\frac{1}{2}} = 1$$

[1]书上1.27式

前面给出的变化式是$x = g(y)$，现在假设$y = f(x)$，即$f = g^{-1}$（因为前面要求了g为双射）。设$F_y(Y)$是y的累积分布函数，也就是

$$F_y(Y) = P_y(y \leq Y) = \int_{-\infty}^{Y} p_y(y)\mathrm{d}y$$

所以有

$$\frac{\partial F_y(Y)}{\partial y} = p_y(y)$$

这里的$p_y(y)$就是我们要求的y的概率密度。用$f(x)$替换y得到

$$F_y(Y) = P_y(y \leq Y) = P_y(f(x) \leq Y)$$

因为$f(x)$必为单调函数[2]，若是递增函数则$f(x) \leq Y$可以表示为$x \leq f^{-1}(Y)$，因为此时有$f(x) \leq f(f^{-1}(Y)) = Y$；若是递减函数则$f(x) \leq Y$可表示为$x \geq f^{-1}(Y)$，因为此时有$f(x) \leq f(f^{-1}(Y)) = Y$，所以上式可写为

$$F_y(Y) = \begin{cases} P_x(x \leq f^{-1}(Y)) & f\text{递增} \\ P_x(x \geq f^{-1}(Y)) & f\text{递减} \end{cases} = \begin{cases} P_x(x \leq f^{-1}(Y)) & f\text{递增} \\ 1 - P_x(x \leq f^{-1}(Y)) & f\text{递减} \end{cases}$$

[2]双射函数在连续定义域内必单调

因此有

$$p_y(y) = \frac{\mathrm{d}F_y(Y)}{\mathrm{d}y} = \begin{cases} \dfrac{\mathrm{d}P_x(x \le f^{-1}(Y))}{\mathrm{d}x}\dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递增} \\ \dfrac{\mathrm{d}\{1 - P_x(x \le f^{-1}(Y))\}}{\mathrm{d}x}\dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递减} \end{cases}$$

$$= \begin{cases} p_x(x)\dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递增} \\ -p_x(x)\dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递减} \end{cases}$$

$$= p_x(x)|\frac{\mathrm{d}x}{\mathrm{d}y}|$$

注意：若$f(x)$是递增函数，则必有$\frac{\mathrm{d}y}{\mathrm{d}x} > 0$（导数大于零函数值递增），所以也有$\frac{\mathrm{d}x}{\mathrm{d}y} > 0$（可把后者看作前者的倒数）；若$f(x)$是递减函数，则必有$\frac{\mathrm{d}x}{\mathrm{d}y} < 0$，所以也有$\frac{\mathrm{d}y}{\mathrm{d}x} < 0$。所以有

$$\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递增} \\ -\dfrac{\mathrm{d}x}{\mathrm{d}y} & f\text{递减} \end{cases}$$

至此我们就得到了书上的1.27式。

也可以从另一种角度来看1.27式的推导过程。设x的取值范围为(a,b)，y的取值范围为(α,β)。根据$g(y)$是递增还是递减，我们有

$$\begin{aligned}\int_\alpha^\beta p_y(y)\mathrm{d}y = \int_a^b p_x(x)\mathrm{d}x &= \int_\alpha^\beta p_x(g(y))\mathrm{d}g(y)\\ &= \begin{cases} \displaystyle\int_\alpha^\beta p_x(g(y))g'(y)\mathrm{d}y & f\text{递增}\\ \displaystyle-\int_\alpha^\beta p_x(g(y))g'(y)\mathrm{d}y & f\text{递减}\end{cases}\\ &= \int_\alpha^\beta p_x(g(y))|g'(y)|\mathrm{d}y\end{aligned}$$

所以有

$$p_y(y) = \frac{\mathrm{d}\int_\alpha^\beta p_y(y)\mathrm{d}y}{\mathrm{d}y} = \frac{\mathrm{d}\int_\alpha^\beta p_x(g(y))|g'(y)|\mathrm{d}y}{\mathrm{d}y} = p_x(g(y))|g'(y)|$$

假设y^*为$p_y(y)$的众数[3]，是否能得出结论$x^* = g(y^*)$也为$p_x(x)$的众数了？

答案为否，通常情况下$p_x(x^*)$不一定是最大值（极值），只有特殊情况下x^*才是$p_x(x)$的极值点。这告诉我们：不能通过求$p_y(y)$的极大似然估计得到y^*，再将y^*代入g得到$x^* = g(y^*)$，并将x^*当作$p_x(x)$极大似然估计的结果。

[3]众数就是$p_y(y)$在y^*处取得最大值，也就是y^*为极值点有$p_y^{'}(y^*) = 0$

根据1.27式我们有$p_y(y) = p_x(g(y))sg'(y)$，其中$s \in \{-1, +1\}$对应去绝对值后的符号。基于这个公式有:

$$p'_y(y) = sp'_x(g(y))\{g'(y)\}^2 + sp_x(g(y))g''(y)$$

假设x^*是$p_x(x)$的极值点，根据$x^* = g(y^*)$得到的y^*，现在验证是否有$p'_y(y^*) = 0$。将x^*, y^*代入后有

$$p'_y(y^*) = sp'_x(g(y^*))\{g'(y^*)\}^2 + sp_x(g(y^*))g''(y^*)$$

上式右边第一项必为0因为有$p'_x(g(y^*)) = 0$，但第二项不一定为0。因为$p_x(g(y^*))$不一定为0（$f'(x) = 0$并不一定有$f(x) = 0$），$g''(y^*)$也不一定为0。所以此时并不一定有$p'_y(y^*) = 0$，即y^*并不一定是$p_y(y)$的极值点。

但是，当x和y是线性关系时（即g或者f为线性函数时）前述结论成立。因为当g为线性函数时$|\frac{\mathrm{dx}}{\mathrm{dy}}| = c$，而给$p_x(x)$乘以一个常数并不改变它极值点的位置。或者因为$g$为线性函数，则必有$g'' = 0$，所以上式$p'_y(y^*) = 0$。

PRML Chapter1 要点墙

1-001	1-002	1-003	1-004	1-005	1-006	1-007	1-008	1-009	1-010
1-011	1-012	1-013	1-014	1-015	1-016	1-017	1-018	1-019	1-020
1-021	1-022	1-023	1-024	1-025	1-026	1-027	1-028	1-029	1-030
1-031	1-032	1-032+	1-033	1-034	1-035	1-036	1-037	1-038	1-039
1-040	1-041	1-042	1-043	1-044	1-045	1-046	1-047	1-048	1-049
1-050	1-051	1-052	1-053	1-054	1-055	1-056	1-057	1-058	1-059
1-060	1-061	1-062	1-063	1-064	1-065	1-066	1-067	1-068	1-069
1-070									

概率统计中，参数估计（学习、训练）是基本问题之一，按估计方式不同可分为点估计、区间估计两类，在机器学习中我们讨论比较多的是点估计。

点估计的一般方法是：设x的分布函数为$F(x;\theta_1,\theta_2,\theta_3,\cdots,\theta_m)$，注意这里$F$的函数形式是已知的[4]，只是具体的参数$\boldsymbol{\theta}=(\theta_1,\cdots,\theta_m)$未知。为每个$\theta_i$构造一个关于样本的函数，并用这个函数的计算结果作为对θ_i的估计值[5]，即

$$\hat{\theta}_i = h_i(x_1,x_2,\cdots,x_n)$$

因而，点估计的重点在于如何构造一个好的、关于样本的函数，使其计算结果$\hat{\theta}_i$正好等于或近似等于该样本所服从分布的真实θ_i值[6]。

[4]也就是说分布类型是已知的

[5]例如高斯分布有$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{N}x_i, \hat{\sigma^2}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\hat{\mu})^2$

[6]注意，用估计函数$h_i(\mathbf{x})$估计得到的结果用$\hat{\theta}_i$表示，真实值用θ_i表示

点估计有两种常用的构造估计函数的方法：矩估计和极大似然估计[7]。

矩估计：矩来源于经典力学中的力矩，在概率统计中它度量的是分布“重心”所在的位置，因而它是一种期望值。矩也有多种度量方式，例如直接计算分布重心所在位置就叫原点矩，将分布重心平移到原点后再计算的矩就叫中心矩。

$\mathbf{n}$阶原点矩（简称$\mathbf{n}$阶矩）：$\mathrm{E}[\mathbf{x}^n]$

$\mathbf{n}$阶中心矩：$\mathrm{E}[(\mathbf{x}-\mathrm{E}[\mathbf{x}])^n]$

$\mathbf{n}$阶绝对矩：$\mathrm{E}[|\mathbf{x}|^n]$

$\mathbf{n}$阶绝对中心矩：$\mathrm{E}[|\mathbf{x}-\mathrm{E}[\mathbf{x}]|^n]$

要特别注意到，1阶矩（1阶原点矩）$\mathrm{E}[\mathbf{x}]$就是期望值，2阶中心矩$\mathrm{E}[(\mathbf{x}-\mathrm{E}[\mathbf{x}])^2]$就是方差。$\mathbf{n}$阶绝对矩是变量求绝对值后的$\mathbf{n}$次方的矩，$\mathbf{n}$阶绝对中心矩则是先将变量的均值移动到原点再求绝对值的$\mathbf{n}$次方的矩。

[7]显然，极大后验估计也是点估计的一种，它基于极大似然估计的

设x的密度函数为$f(x;\theta_1,\theta_2,\cdots,\theta_m)$，则它的$k$阶原点矩为:

$$\mathrm{E}[x^k]=\int_{-\infty}^{+\infty}x^k f(x;\theta_1,\theta_2,\cdots,\theta_m)\mathrm{d}x$$

若$\mathrm{E}[x^k]$的值是已知的[8]，则上式就成了关于$(\theta_1,\cdots,\theta_m)$的方程。理论上，我们只需将$m$阶矩的方程联立起来就得到了一个包含$m$个方程的$m$元方程组，求可以求出全部的$\theta$值，这就是矩估计的基本思想。

矩估计法的困难之处在于：

1. 不是每种分布的各阶矩都存在，有些分布的矩是不存在的，也就是上述积分是发散的或不可积的
2. 即使矩理论上存在，求积分的计算也可能是非常困难的，最后联立方程组求解则可能是更困难的

[8]可用其它方式去求，例如基本大量样本用蒙特卡洛法求近似值；在求矩的值之前应首先判断矩是否存在，这涉及到绝对收敛的概念

第二种方法：极大似然估计。将样本代入似然函数得到似然的连乘，求使得该连乘积最大的参数的值[9]

$$L(x_1, x_2, \cdots, x_n; \hat{\theta_1}, \hat{\theta_2}, \cdots, \hat{\theta_m}) = \max_{\theta_1, \theta_2, \cdots, \theta_m} L(x_1, x_2, \cdots, x_m; \theta_1, \theta_2, \cdots, \theta_m)$$

理论上，只有当对数似然是关于θ_i的凸函数时上述方法求得的$\hat{\theta}_i$才是全局最优解，否则可能只是局部最优解。

这里有几点请注意：

1. 这里所有参数都共用一个似然函数，而非矩估计里可为每个θ_i单独构造估计函数
2. 只有当所有样本相互独立时才能求它们的乘积（两个独立事件的联合概率等于它们各自概率的乘积）
3. 为了便于计算常借助对数函数将连乘转换为求和

[9]极大似然估计最初由伯努利提出，但主要由费希尔（Fisher）发展完善

同一个参数可以有多种不同估计方法，那如何评判估计值的好坏了？一般我们从无偏性、有效性、一致性三个方面来评价。我们这里重点介绍无偏性。

无论是矩估计还是极大似然估计，都是基于样本集$\{\mathbf{x}\}$计算得到的，因而同样的方法、不同的样本集会得到某个参数的不同估计结果。若取遍所有可能的样本集，在所有可能样本集下某个参数估计值的期望正好等于总体中该参数的真实值，我们就说此估计是无偏估计。设$\hat{\theta} = \hat{\theta}(x_1, x_2, \cdots, x_n)$是未知参数$\theta$的一个估计值，若对$\hat{\theta}$的所有可能取值有

$$\mathrm{E}(\hat{\theta}) = \theta$$

则称$\hat{\theta}(x_1, x_2, \cdots, x_n)$是$\theta$的一个无偏估计量[10]。

[10]注意，不是说此时某一个具体的$\hat{\theta}$值是无偏的，而是说这种估计方式（函数）是无偏的

PRML书上1.55式是均值的无偏估计，1.56式是方差的有偏估计，1.59式是方差的无偏估计

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n (1.55), \quad \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})(1.56), \quad \tilde{\sigma^2} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 (1.59)$$

要验证这三项是否有偏，只需要验证μ_{ML}、σ^2_{ML}、$\tilde{\sigma^2}$三项的期望是否等于μ、σ的定义

1. 验证：$\mathrm{E}[\mu_{ML}] \stackrel{?}{=} \mathrm{E}[x] = \mu$
2. 验证：$\sigma^2_{ML} \stackrel{?}{=} \mathrm{E}[(x-\mu)^2] = \sigma^2$
3. 验证：$\tilde{\sigma^2} \stackrel{?}{=} \mathrm{E}[(x-\mu)^2] = \sigma^2$

先看μ_{ML}，根据期望定义有

$$\begin{aligned}\mathrm{E}[\mu_{ML}] &= \mathrm{E}[\frac{1}{N}\sum_{n=1}^{N} x_n] \\ &= \frac{1}{N}\sum_{n=1}^{N}\mathrm{E}[x_n] \\ &= \frac{1}{N}N\mu \\ &= \mu\end{aligned}$$

注意$\mathrm{E}[x_n]$表示第n个样本取遍所有可能的取值时的期望，所以它实际上就是x的期望，因而就等于μ。可见μ_{ML}是对μ的无偏估计[11]。

[11]样本集均值是对总体均值的无偏估计

再看σ^2_{ML}，根据定义有

$$\mathrm{E}[\sigma^2_{ML}] = \frac{1}{N}\mathrm{E}[\sum_{n=1}^{N}(x_n - \mu_{ML})^2] = \frac{1}{N}\mathrm{E}[\sum_{n=1}^{N}(x_n - \bar{x})^2]$$

重点在于对$\sum_{n=1}^{N}(x_n - \bar{x})^2$的变形。这里简便起见，用$\bar{x}$代替了$\mu_{ML}$。

$$\begin{aligned}\sum_{n=1}^{N}(x_n - \bar{x})^2 &= \sum_{n=1}^{N}[(x_n - \mu) - (\bar{x} - \mu)]^2 \\ &= \sum_{n=1}^{N}\{(x_n - \mu)^2 - 2(x_n - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\} \\ &= \sum_{n=1}^{N}(x_n - \mu)^2 - 2(\bar{x} - \mu)\sum_{n=1}^{N}(x_n - \mu) + N(\bar{x} - \mu)^2\end{aligned}$$

因为

$$\sum_{n=1}^{N}(x_n-\mu)=N\bar{x}-\sum_{n=1}^{N}\mu=N\bar{x}-N\mu=N(\bar{x}-\mu)$$

代入上式有：

$$\begin{aligned}\sum_{n=1}^{N}(x_n-\bar{x})^2&=\sum_{n=1}^{N}(x_n-\mu)^2-2(\bar{x}-\mu)\cdot N(\bar{x}-\mu)+N(\bar{x}-\mu)^2\\&=\sum_{n=1}^{N}(x_n-\mu)^2-N(\bar{x}-\mu)^2\end{aligned}$$

所以有

$$\mathrm{E}[\sum_{n=1}^{N}(x_n-\bar{x})^2]=\mathrm{E}[\sum_{n=1}^{N}(x_n-\mu)^2-N(\bar{x}-\mu)^2]=\sum_{n=1}^{N}\mathrm{E}[(x_n-\mu)^2]-N\mathrm{E}[(\bar{x}-\mu)^2]$$

回顾方差的定义：设变量x的期望为μ，则x的方差为

$$D(x) = \mathrm{E}[(x-\mu)^2]$$

则上式第一项有

$$\mathrm{E}[(x_n-\mu)^2] = D(x_n) = \sigma^2$$

类似的，这里$\mathrm{E}[(x_n-\mu)^2]$是第n个样本x_n取遍所有可能的值时的期望，所以它实际就是$\mathrm{E}[(x-\mu)^2]$，也就是总体方差，这里用σ^2表示。因而

$$\sum_{n=1}^{N} \mathrm{E}[(x_n-\mu)^2] = \sum_{n=1}^{N} \sigma^2 = N\sigma^2$$

对于第二项，因为已知$\bar{x}$是对总体均值的无偏估计，所以有

$$\mathrm{E}[\bar{x}] = \mu$$

将μ用上式代入后有

$$\begin{aligned}
\mathrm{E}[(\bar{x}-\mu)^2] &= \mathrm{E}[(\bar{x}-\mathrm{E}[\bar{x}])^2] \\
&= D(\bar{x}) \quad (\text{把}\bar{x}\text{看作变量，它正好是}\bar{x}\text{的方差}) \\
&= D(\frac{1}{N}\sum_{i=1}^{N} x_i) \\
&= \frac{1}{N^2}\sum_{i=1}^{N} D(x_i) \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

所以我们有

$$\begin{aligned}\mathrm{E}[\sum_{n=1}^{N}(x_n-\bar{x})^2] &= \sum_{n=1}^{N}\mathrm{E}[(x_n-\mu)^2]-N\mathrm{E}[(\bar{x}-\mu)^2] \\ &= N\sigma^2-N\frac{\sigma^2}{N} \\ &= (N-1)\sigma^2\end{aligned}$$

所以

$$\begin{aligned}\mathrm{E}[\sigma^2_{ML}] &= \frac{1}{N}\mathrm{E}[\sum_{n=1}^{N}(x_n-\bar{x})^2] \\ &= \frac{N-1}{N}\sigma^2 \\ &\neq \sigma^2\end{aligned}$$

所以σ^2_{ML}是σ^2的有偏估计

而$\mathrm{E}[\tilde{\sigma^2}]$有

$$\begin{aligned}\mathrm{E}[\tilde{\sigma^2}] &= \mathrm{E}[\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\ &= \frac{1}{N-1}\mathrm{E}[\sum_{n=1}^{N}(x_n-\mu_{ML})^2] \\ &= \frac{1}{N-1}(N-1)\sigma^2 \\ &= \sigma^2\end{aligned}$$

所以$\tilde{\sigma^2}$是σ^2的无偏估计

PRML Chapter2 要点墙

2-001	2-002	2-003	2-004	2-005	2-006	2-007	2-008	2-009	2-010
2-011	2-012	2-013	2-014	2-015	2-016	2-017	2-018	2-019	2-020
2-021	2-022	2-023	2-024	2-025	2-026	2-027	2-028	2-029	2-030
2-031	2-032	2-033	2-034	2-035	2-036	2-037	2-038	2-039	2-040
2-041	2-042	2-043	2-044	2-045	2-046	2-047	2-048	2-049	2-050
2-051	2-052	2-053	2-053+	2-054	2-055	2-056	2-057	2-058	2-058+
2-059	2-060	2-061	2-062	2-063	2-064	2-065	2-066	2-067	2-068

高斯分布是最重要的概率分布，现对其性质做一个小结

1. 一元高斯分布是一个单峰、对称的分布，均值μ同时是该分布的期望、众数和中位数，其密度函数为

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

众数（mode）：对离散型随机变量，众数是概率质量最大的变量值；对连续型随机变量，众数是概率密度值最大的变量值（即概率密度函数峰值）。
中位数（median）：对于连续型随机变量其定义为若$P(X\leq c)\geq\frac{1}{2}$且$P(X\geq c)\geq\frac{1}{2}$，则$c$为其中位数；对离散型随机变量，若最小的$c_1$使得$P(X\leq c_1)\geq\frac{1}{2}$且最大的$c_2$使得$P(\geq c_2)\geq\frac{1}{2}$，则$\frac{c_1+c_2}{2}$为其中位数。

x	1	2	3	4	5	6	7	8
p	0.1	0.1	0.15	0.05	0.2	0.2	0.15	0.05

（此时有$P(X\leq 5)\geq 0.5$且$P(X\geq 5)\geq 0.5$，因此中位数为5）

2. 多元高斯分布的密度函数如下，$\mathbf{x}, \boldsymbol{\mu}$都是D维向量，$\boldsymbol{\Sigma}$是$D \times D$型对称矩阵

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

$\boldsymbol{\mu}$、$\boldsymbol{\Sigma}$的极大似然估计结果如下所示（无偏估计）

$$\boldsymbol{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n (2.121), \quad \boldsymbol{\Sigma}_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T (2.125)$$

均值影响概率密度函数的中心点位置，协方差阵影响其形状。
若协方差阵为单位阵，说明x_1、x_2两个变量相互独立且各自方差为1，因此等高线投影图为标准同心圆
若协方差阵为对角阵，说明x_1、x_2相互独立但各自方差不同，因此等高线投影为椭圆且长短轴与坐标轴平行
若协方差阵是一般对称阵，说明x_1、x_2有一定的相关性，等高线投影图一般表现为长短轴发生偏转的椭圆

Congrats | 淡小蓝点 bluedotdot.cn

微信号：bluedotdot_cn

More：《PRML Page-by-page》、《面向机器学习（深度学习、人工智能）的数学基础》、《DDPM原理推导及代码实现》、《OpenAI编程基础》等