1. 极大似然估计

首先要明确极大似然估计要解决的问题。假设某个随机变量的概率分布形式已知，但是要确定这个分布的具体形式还需要知道分布模型的参数。极大似然估计就是一种从观测样本来估计出分布参数的技术；其核心思想就是，分布的参数在取什么具体值的时候会使得观测以最大可能的概率发生。

例：设样本$\left(X_1,X_2,...,X_n\right)$来自正态总体$X\sim N\left(\mu,\sigma^2\right)$，$\mu,\sigma^2$未知，求$\mu,\sigma^2$的极大似然估计。

解：

设$\left(x_1,x_2,...,x_n\right)$为对应的样本观察值，则关于$\mu,\sigma^2$的似然函数为

$$L\left(\mu,\sigma^2\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}}=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2}$$

因此，

$$\ln L\left(\mu,\sigma^2\right)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$$

令$\frac{\partial\ln L}{\partial\mu}=0,\frac{\partial\ln L}{\partial\sigma^2}=0$得到

$$\begin{cases}\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)=0\\-\frac{n}{2}\frac{1}{\sigma^2}-\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\cdot\left(-\frac{1}{\sigma^4}\right)=0\end{cases}，得到\begin{cases}\widehat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i\\\widehat{\sigma^2}=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\widehat{\mu}\right)^2\end{cases}$$

例：设样本$\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\right),\mathbf{x}_i\in\mathbb{R}^{d\times1}$来自正态总体$X\sim N\left(\boldsymbol{\mu},\Sigma\right)$，$\boldsymbol{\mu},\Sigma$未知，求$\boldsymbol{\mu},\Sigma$的极大似然估计。

解：d维随机变量的高斯分布概率密度函数为，

$f\left(\mathbf{x}:|\boldsymbol{\mu},\Sigma\right)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$，其中$\boldsymbol{\mu}\in\mathbb{R}^{d\times1}$，$\Sigma\in\mathbb{R}^{d\times d}$为协方差矩阵，是一个半正定（当然也是实对称）矩阵。

则$\boldsymbol{\mu},\Sigma$的似然函数为，

$$L\left(\boldsymbol{\mu},\Sigma\right)=\prod_{i=1}^{n}\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right)$$

$$=(2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\Sigma^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\right)$$

其对数似然函数为，

$$l(\boldsymbol{\mu},\Sigma)=-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln(|\Sigma|)-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{3-1}$$

现在要计算最优的$\boldsymbol{\mu}^*,\Sigma^*$来使得$l(\boldsymbol{\mu},\Sigma)$达到最大，因此要计算$l(\boldsymbol{\mu},\Sigma)$的驻点。

$$\frac{dl}{d\boldsymbol{\mu}}=-\frac{1}{2}\frac{d\left\{\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\}}{d\boldsymbol{\mu}}=-\frac{1}{2}\sum_{i=1}^{n}\frac{d\left\{(\boldsymbol{\mu}-\mathbf{x}_i)^T\Sigma^{-1}(\boldsymbol{\mu}-\mathbf{x}_i)\right\}}{d\boldsymbol{\mu}}$$

根据第二章（矩阵论）结论 7.5 第 5 条，

$$\frac{dl}{d\boldsymbol{\mu}}=-\frac{1}{2}\sum_{i=1}^{n}(\Sigma^{-1}+\Sigma^{-T})(\boldsymbol{\mu}-\mathbf{x}_i)=-\frac{1}{2}\sum_{i=1}^{n}(\Sigma^{-1}+\Sigma^{-1})(\boldsymbol{\mu}-\mathbf{x}_i)=\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})$$

令上式等于 **0**，则我们有

$\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})=\mathbf{0}$，由于$\Sigma^{-1}$可逆，则$\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})=\mathbf{0}$，因此，

$$\boldsymbol{\mu}^*=\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

接下来要计算最优的Σ^*。如果要直接计算式（3-1）对Σ的导数不太容易，我们把（3-1）代换成Σ^{-1}的等价函数，求出可使得$l(\boldsymbol{\mu},\Sigma^{-1})$取得最大值时的$(\Sigma^{-1})^*$，当然就可以求得$\Sigma^*$。

令$A=\Sigma^{-1}$，则$A^{-1}=\Sigma$，且容易知道，$A=A^T$，式（3-1）可等价转换为，

$$l(\boldsymbol{\mu},A)=-\frac{nd}{2}\ln(2\pi)-\frac{n}{2}\ln(|A^{-1}|)-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T A(\mathbf{x}_i-\boldsymbol{\mu})$$

接下去计算$l(\boldsymbol{\mu},A)$关于A这部分的梯度为零的点。

$$\frac{d\{l(\boldsymbol{\mu},A)\}}{dA}=-\frac{n}{2}\frac{d\{\ln(|A^{-1}|)\}}{dA}-\frac{1}{2}\frac{d\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^T A(\mathbf{x}_i-\boldsymbol{\mu})}{dA}$$

$$=-\frac{n}{2}\frac{d\{-\ln|A|\}}{dA}-\frac{1}{2}\sum_{i=1}^{n}\frac{d\left\{(\mathbf{x}_i-\boldsymbol{\mu})^T A(\mathbf{x}_i-\boldsymbol{\mu})\right\}}{dA}$$

$$=\frac{n}{2}\frac{1}{|A|}|A|A^{-T}-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

$$=\frac{n}{2}A^{-1}-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

令$\dfrac{d\left\{l\left(\boldsymbol{\mu},A\right)\right\}}{dA}=\mathbf{0}$，则有

$\dfrac{n}{2}A^{-1}-\dfrac{1}{2}\sum_{i=1}^{n}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T=\mathbf{0}$，则$A^{-1}=\dfrac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T$，则

$\Sigma^*=\dfrac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T$。