最大似然可以用马氏距离估计吗

BY JIN

2019-02-15

假如有一维连续空间中的实验结果X，可能（先验概率均等地）产生自两个正态分布$\mathcal{N}_1(\mu_1,\ \sigma_1^2)$，$\mathcal{N}_2(\mu_2,\ \sigma_2^2)$之一，想用最大似然估计法推测其来源，应该比较$x$在两个分布中的马氏距离$D_i = \frac{X-\mu_i}{\sigma_i}$还是概率密度$f_i(X)=\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{1}{2}\left(\frac{X-\mu_i}{\sigma_i}\right)^2}$？由于马氏距离在标准化正态分布中的概率密度$g(d)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}d^2}(d\geqslant 0)$为严格递减函数，但却比$f_i$少了一个$\frac{1}{\sigma_i}$的因子，因而可以断定当$\sigma$不同时两种判定方式并不等价。

按最大似然估计方法，由于先验概率$P^-(x\sim\mathcal{N}_i)$均等，在X附近取极小的区间$\Delta_x\to 0$，有

$$\begin{aligned} P(x\sim\mathcal{N}_i|X-\Delta\leqslant x\leqslant X+\Delta) &= \eta P(X-\Delta\leqslant x\leqslant X+\Delta|x\sim\mathcal{N}_i)P^-(x\sim\mathcal{N}_i) \\ &= \eta\cdot 2\Delta\cdot f_i(X)\cdot P^-(x\sim\mathcal{N}_i) \\ &= \eta' f_i(X) \end{aligned}$$

因此X属于第i个分布的可能性与$f_i(X)$成正比例，确定最大似然时比较每个分布在X的概率密度即可。

另一方面，求马氏距离等于标准化了原正态分布，对于同样的$\Delta_x\to 0, \Delta_{d_i}=\frac{\Delta_x}{\sigma_i}$却是一个与方差相关的量

$$\begin{aligned} P(x\sim\mathcal{N}_i|X-\Delta_x\leqslant x\leqslant X+\Delta_x) &= P(x\sim\mathcal{N}_i|D-\Delta_{d_i}\leqslant d_i\leqslant D+\Delta_{d_i}) \\ &= \eta\cdot 2\frac{\Delta_x}{\sigma_i}\cdot g(D_i)\cdot P^-(x\sim\mathcal{N}_i) \\ &= \eta'\frac{g(D_i)}{\sigma_i} \end{aligned}$$

因此单纯比较马氏距离是错误的；应该按马氏距离求出标准正态分布的概率密度，再除以对应的标准差。而按之前的推导这正是X在每个分布中的概率密度。