MMSE估计的来龙去脉

Eason

目录

本文讲述MMSE的来龙去脉。对于一个随机变量期望的估计，频率统计是一种方法，此处略去不提。本文基于贝叶斯公式，给出估计随机变量值的另一种方式。

1 问题模型

通常，我们需要估计未知随机变量X的值，但是往往需要通过对另外一个随机变量的观测Y来对X进行推断。我们把X的概率密度分布叫做先验概率$P_X(x)$。当我们获取到观测值$Y=y$，然后对X进行估计时，此时的概率值叫做后验估计$P(X|Y=y)$。

后验估计通常通过贝叶斯公式完成。

2 贝叶斯公式

贝叶斯公式表述如下：

$$P_{X|Y}(x|y)=\frac{P_{Y|X}(y|x)p_X(x)}{P_Y(y)} \tag{1}$$

此处给一个例子，假设$P_X(x)\sim uniform(0,1)$，且$P(Y|X=x)\sim geometry(x)$，求$P(X|Y=2)$.

根据贝叶斯公式有：

$$P(X|Y=2)=\frac{P_{Y|X}(2|X=x)P(X)}{P(Y)} \tag{2}$$

因为：

$$P(Y|X=x)=x(1-x)^{y-1} \tag{3}$$

则：

$$P(2|X=x)=x(1-x) \tag{4}$$

利用全概率公式：

$$P_Y(2)=\int_{-\infty}^{+\infty}P(2|x)P(x)dx=\frac{1}{6} \tag{5}$$

所以$P(X|Y=2)=6x(1-x)$.

3 最大后验估计准则

由于后验概率密度分布$P_{X|Y}(x|y)$包含了关于X的所有信息。所以我们可以利用后验概率密度对X进行点估计。对X进行点估计的一个准则是：选择一个x使得$P_{X|Y}(x|y)$的值最大。这个估计准则叫做最大后验估计准则。

为了找到X的MAP值，我们力图找到x使得式~(6)的取值最大。

$$P(x|y)=\frac{P(y|x)P(x)}{P(y)} \tag{6}$$

注意$P(y)$不依赖于X，所以我们只需要最大化$P(y|x)P(x)$即可。更进一步，如果$P(x)$是等概分布，那么我们只需要最大化$P(y|x)$即可，即找到那个最大化$P(y|x)$的x值即可。此时，MAP准则和ML准则等效。

4 条件期望

$P(X|y)$包含$Y = y$时X的所有信息。所以我们可以利用$P(X|y)$找到关于x的多个估计，比如均值，中位数，mode。我们定义mode为最大化 $P(X|y)$的x的值，即：mode是MAP对应的x的值。另一个取值方式是取后验分布的均值，即：

$$\hat{x} = E[P(X|Y = y)] \tag{7}$$

给一个例子，感受一下$\hat{x}$的计算。假设

$$f_X(x) = \begin{cases} 2x & 0 \le x \le 1 \\ 0 & otherwise \end{cases} \tag{8}$$

$$f_{Y|X}(y|x) = \begin{cases} 2xy - x + 1 & 0 \le y \le 1 \\ 0 & otherwise \end{cases} \tag{9}$$

求$\hat{x}$.

首先利用全概率公式，我们有：

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x) dx = \frac{4}{3}y + \frac{1}{3}, 0 \le y \le 1 \tag{10}$$

我们有$f_{X|Y}(x|y) = \frac{6x(2xy-x+1)}{4y+1}, 0 \le x \le 1$,则 $\hat{x}$:

$$\hat{x} = E[X|Y = y] = \int_0^1 x f_{X|Y}(x|y) dx = \frac{6y+1}{8y+2} \tag{11}$$

5 MMSE

接下来我们证明$\hat{x}$是在均方误差准则下X的最好估计。假设给定$Y = y$我们要估计X的值$\hat{x}$。一般情况，$\hat{x}$是y的函数。估计的误差为：

$$\tilde{X} = X - \hat{x} = X - g(y) \tag{12}$$

通常，我们希望均方误差的期望最小，即：

$$\min_{g(y)} E[(X - g(y))^2 | Y = y] \tag{13}$$

式~(13)正是我们称这个方法为最小均方误差估计的原因。

为简便起见，假定我们在没有任何观察的情况下对X进行估计，我们会用一个什么样的值作为X的估计值呢？假设这个值是a，那么MSE为：

$$h(a) = E[(X-a)^2] = E[X^2] - 2aE[X] + a^2 \tag{14}$$

我们对a求导，有：

$$h^{'}(a) = -2E[X] + 2a \tag{15}$$

令$h^{'}(a) = 0$，那么有$a = E[X]$。现在，假设，我们有观测数据$Y = y$，那么此时我们会对X做什么样的估计呢？假设估计为$\hat{x}$，那么MSE有：

$$E[(X-\hat{x})^2|Y=y] = E[X^2|Y=y] - E[2\hat{x}X|Y=y] + \hat{x}^2 \tag{16}$$

就上式对$\hat{x}$求导，并令导数等于零，则：

$$\hat{x} = E[X|Y=y] \tag{17}$$

即最小均方误差准则下的最优解是条件期望。这个值我们可以通过贝叶斯公式求出。

6 MMSE的一些特性

由于$\hat{x}$是y的函数，即$\hat{x} = g(y)$。我们也可以认为 $\hat{X} = g(Y)$，同样有$\hat{X}_M = E[X|Y]$，即关于X在MSE准则下的估计$\hat{X}_M$是条件期望。

由于$E[\hat{X}_M] = E[E[X|Y]] = E[X]$，所以有$E[\tilde{X} = E[X - \hat{X}M]] = 0$。即$\hat{X}_M$是$X$的无偏估计。

定义随机变量$W = E[\tilde{X}|Y]$。令$\hat{X}_M = E[X|Y]$是MMSE估计算子，定义$\tilde{X} = X - \hat{X}$为估计误差。那么$W = 0$，并且对于任意的估计子$g(Y)$都有$E[\tilde{X}g(Y)] = 0$. 对于这个结论，我们给出证明：

$$\begin{aligned} W &= E[\tilde{X}|Y] \\ &= E[X - \hat{X}_M|Y] \\ &= E[X|Y] - E[\hat{X}_M|Y] \\ &= \hat{X}_M - \hat{X}_M \\ &= 0 \end{aligned}$$

另外，我们有$E[\tilde{X}g(Y)|Y]=g(Y)E[\tilde{X}|Y]=0$，所以$E[\tilde{X}g(Y)]=E[E[\tilde{X}g(Y)|Y]]=0$。

接下来我们证明$\tilde{X}$和$\hat{X}_M$是不相关的。我们有：

$$\begin{aligned} Cov(\tilde{X},\hat{X}_M) &= E[\tilde{X}\hat{X}_M]-E[\tilde{X}]E[\hat{X}_M] && (18)\\ &= E[\tilde{X}\hat{X}_M] && (19)\\ &= E[\tilde{X}g(Y)]=0 && (20)\end{aligned}$$

因为$\tilde{X}=X-\hat{X}_M$，所以$X=\hat{X}_M+\tilde{X}$，又因为$\tilde{X}$和$\hat{X}_M$是互不相关的，则有：

$$Var(X)=Var(\tilde{X})+Var(\hat{X}_M) \tag{21}$$

上式可以解释为X的方差有一部分是估计体现的，有一部分是估计误差体现的。如果估计值$\hat{X}_M$捕捉到了X大部分的方差，那么估计误差就会小一些。

我们把式 (21)用期望重写为：

$$E[X^2]-E[X]^2=E[\tilde{X}^2]-E[\tilde{X}]^2+E[\hat{X}_M^2]-E[\hat{X}_M]^2 \tag{22}$$

由于 $E[\tilde{X}]^2=0$，且$E[X]=E[\hat{X}_M]$，则有：

$$E[X^2]=E[\tilde{X}^2]+E[\hat{X}_M^2] \tag{23}$$

7 应用

MMSE估计在通信系统和信号处理领域诸多方向都会出现，比如信道追踪，信号检测，译码，图像重建，无线定位，频偏估计等等。在这些领域，我们通常需要基于观测估计未知的参数$\mathbf{x}\in\mathbb{R}^D$，观测方程可以表示为：

$$z=f(\mathbf{x})+n \tag{24}$$

其中n是测量噪声，测量方程$f(\mathbf{x})$可以是线性的也可以是非线性的。

有很多方法可以帮助我们从$\{z_i\}$中估计出$\mathbf{x}$。这些方法可以简单的分为基于统计的和非统计的。基于统计的方法有：最大似然估计（MLE），最大后验概率估计（MAP），最小均方误差估计（MMSE）。基于非统计的方法有：最小二乘（LS），最优线性无偏估计（BLUE）和最小方差无偏估计（MVU）。基于统计的估计通常以最小估计误差为优化目标，给出最优的参数估计结果。基于非统计的估计则提供了一种当信号统计特性未知时的简单估计方法。无论采用哪一类估计方法，估计子的无偏性和协方差都是我们要考虑的两个度量。在

一些特殊的场合，基于统计的估计算法和基于非统计的估计算法是等效的。基于我们对系统和统计信息的掌握，我们有多种估计算法可选。例如，如果我们知道系统测量是线性的，测量噪声是零均值高斯变量，$\mathbf{z}=\mathbf{A}\mathbf{x}+\mathbf{n}$，那么我们可以使用MLE来估计 $\mathbf{x}$。更进一步，如果我们知道x的先验信息$p(\mathbf{x})$，那么可以用线性的MMSE算法来估计$\mathbf{x}$。

MMSE方法的目标是最小化均方误差，因此在统计意义上，这个算法是最优的（假设已知先验信息$p(\mathbf{x})$）。MSE定义为：

$$MSE=\int_{\mathbb{X}} p(\mathbf{x}|z)(\mathbf{x}-\mathbf{x})^T(\hat{\mathbf{x}}-\mathbf{x})d\mathbf{x} \tag{25}$$

其中$p(\mathbf{x}|z)$是$\mathbf{x}$的后验分布。则MMSE的估计结果为：

$$\hat{\mathbf{x}}_{MMSE}=\arg\min_{\mathbf{x}}\int_{\mathbb{X}} p(\mathbf{x}|z)(\mathbf{x}-\mathbf{x})^T(\hat{\mathbf{x}}-\mathbf{x})d\mathbf{x} \tag{26}$$

通过对上式求导：

$$\frac{d\int_{\mathbb{X}} p(\mathbf{x}|z)(\mathbf{x}-\mathbf{x})^T(\hat{\mathbf{x}}-\mathbf{x})d\mathbf{x}}{d\mathbf{x}}=0 \tag{27}$$

最优的MMSE估计算子是：

$$\hat{\mathbf{x}}_{MMSE}=\int_{\mathbb{X}} p(\mathbf{x}|z)d\mathbf{x} \tag{28}$$

可以看到，最优MMSE估计是它的后验概率期望。通常，我们使用贝叶斯链式法则来求解后验概率：

$$p(\mathbf{x}|z)=\frac{p(z|\mathbf{x}p(\mathbf{x})}{p(z)} \tag{29}$$

其中$p(z|\mathbf{x})$是似然函数，$p(\mathbf{x})$是先验信息，$p(z)$是归一化项，可以通过全概率公式求$p(z)$：

$$p(z)=\int_{\mathbb{X}} p(z|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{30}$$

对于一个特性系统或者估计问题来说，MMSE剩下的问题就是求解式 (29)中出现的统计密度。

接下来我们以线性高斯MMSE估计子为例，其系统模型为：

$$\mathbf{z}=\mathbf{A}\mathbf{x}+\mathbf{n} \tag{31}$$

其中$\mathbf{n} \sim \mathcal{N}(\mathbf{n}|\mathbf{0}, \mathbf{W})$， $\mathbf{W}$是精度矩阵.另外假设目标随机变量$\mathbf{x}$服从高斯分布。

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\chi, \Lambda) \tag{32}$$

其中χ, Λ 是对应的均值矩阵和精度矩阵。

基于以上的公式，似然函数为：

$$p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{A}\mathbf{x}, \mathbf{W}) \tag{33}$$

因此，基于高斯分布的特性，后验分布为：

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}) &\propto p(\mathbf{z}|\mathbf{x})p(\mathbf{x}) && (34)\\ &= \mathcal{N}(\mathbf{z}|\mathbf{A}\mathbf{x}, \mathbf{W})\mathcal{N}(\mathbf{x}|\chi, \Lambda) && (35)\\ &= \mathcal{N}(\mathbf{x}|\mathbf{A}^{+}\mathbf{z}, \mathbf{W}^{'})\mathcal{N}(\mathbf{x}|\chi, \Lambda) && (36) \end{aligned}$$

其中$\mathbf{A}^{+}$是$\mathbf{A}$的广义逆，$\mathbf{W}^{'} = \mathbf{A}^{T}\mathbf{W}\mathbf{A}$ 。我们发现，后验概率 $p(\mathbf{x}|\mathbf{z})$是两个高斯分布的乘积，因此也是一个高斯分布。

$$\begin{aligned} p(\mathbf{x}|\mathbf{z}) &= \mathcal{N}(\mathbf{x}|\mathbf{A}^{+}\mathbf{z}, \mathbf{W}^{'})\mathcal{N}(\mathbf{x}|\chi, \Lambda) && (37)\\ &= \mathcal{N}(\mathbf{x}|\chi^{\dagger}, \Lambda^{\dagger}) && (38) \end{aligned}$$

其中：

$$\begin{aligned} \chi^{\dagger} &= (\Lambda^{\dagger})^{-1}(\mathbf{W}^{'}\mathbf{A}^{+}\mathbf{z} + \Lambda\mathbf{x}) && (39)\\ \Lambda^{\dagger} &= \mathbf{W}^{'} + \Lambda && (40) \end{aligned}$$

我们把$\mathbf{W}^{'}$的结果带入式 (39)，则有：

$$\begin{aligned} \chi^{\dagger} &= (\mathbf{A}^{T}\mathbf{W}\mathbf{A} + \Lambda)^{-1}(\mathbf{A}^{T}\mathbf{W}\mathbf{z} + \Lambda\chi) && (41)\\ \Lambda^{\dagger} &= \mathbf{A}^{T}\mathbf{W}\mathbf{A} + \Lambda && (42) \end{aligned}$$

之前我们知道MMSE估计结果是后验期望，所以：

$$\hat{\mathbf{x}}_{MMSE} = \chi^{\dagger} = (\mathbf{A}^{T}\mathbf{W}\mathbf{A} + \Lambda)^{-1}(\mathbf{A}^{T}\mathbf{W}\mathbf{z} + \Lambda\chi) \tag{43}$$

接下来，我们考虑通信系统的场景。在通信系统中，$\mathbf{x}$的均值通常为0，即$\chi = \mathbf{0}$，所以：

$$\hat{\mathbf{x}}_{MMSE} = \chi^{\dagger} = (\mathbf{A}^T\mathbf{W}\mathbf{A} + \Lambda)^{-1}\mathbf{A}^T\mathbf{W}\mathbf{z} \tag{44}$$

在式 (44)中$\mathbf{W}$是精度矩阵。通常，我们还会看到式 (44)使用协方差矩阵的写法。我们知道精度矩阵和协方差矩阵的关系为：

$$\Sigma_n = \mathbf{W}^{-1} \tag{45}$$

$$\Sigma_x = \Lambda^{-1} \tag{46}$$

所以式 (44)可以变为：

$$\mathbf{x}_{MMSE} = (\mathbf{A}^T\Sigma_n^{-1}\mathbf{A} + \sigma_x^{-1})^{-1}\mathbf{A}^T\Sigma_n^{-1}\mathbf{z} \tag{47}$$

$$= (\mathbf{A} + \Sigma_n(\mathbf{A}^T)^{-1}\Sigma_x^{-1})^{-1}\mathbf{z} \tag{48}$$

$$\approx ((\mathbf{A})^T\mathbf{A} + \Sigma_n\Sigma_x^{-1})\mathbf{A}^T\mathbf{z} \tag{49}$$

$$= (\mathbf{A}^T\mathbf{A} + \gamma^{-1}\mathbf{I})^{-1}\mathbf{A}^T\mathbf{z} \tag{50}$$

其中γ是接收端的信噪比。如果信号功率是归一化的，那么 $\gamma \propto \sigma_n^{-2}$ 。所以在通信系统中，线性MMSE估计为：

$$\mathbf{x}_{MMSE} = (\mathbf{A}^T\mathbf{A} + \sigma_n^2\mathbf{I})^{-1}\mathbf{A}^T\mathbf{z} \tag{51}$$

式~(51)在通信系统中经常出现，尤其是信道估计和符号检测模块。在符号检测模A代表信道。