

Maximum a posteriori estimation

Já estudamos estimadores de máxima verossimilhança, no qual o parâmetro a ser estimado ^{não} é conhecido; porém, admitimos que seu valor era fixo. Vamos agora mudar de perspectiva e admitir que o parâmetro a ser estimado é uma variável aleatória. Essa ideia representa bem os chamados métodos estatísticos bayesianos ou estatística bayesiana.

Dado um parâmetro θ , se o mesmo é uma variável aleatória, então podemos pensar em distribuições de probabilidade associadas a ele. Nesse contexto, podemos escrever

$$P(\theta|\bar{x}) = \frac{P(\bar{x}|\theta) P(\theta)}{P(\bar{x})}$$

usando o Teorema de Bayes, sendo \bar{x} um conjunto de dados. Esse teorema pode ser obtido a partir da definição de probabilidade condicional. Para isso, tome dois eventos A e B , de modo que,

$$P(A|B) = \frac{P(A,B)}{P(B)} \Rightarrow P(B) P(A|B) = P(A) P(B|A)$$

$$P(B|A) = \frac{P(A,B)}{P(A)} \quad P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

Na expressão envolvendo θ e \bar{x} , temos os seguintes termos:

$P(\bar{x}|\theta)$ - a verossimilhança;

$P(\bar{x})$ - a distribuição a priori dos dados;

$P(\theta)$ - a distribuição a priori do parâmetro;

$P(\theta|\bar{x})$ - a distribuição a posteriori do parâmetro.

Em princípio, podemos usar um processo de maximização de $P(\bar{x}|\theta)$ para obter o chamado estimador de máximo a posteriori de θ . Vejamos um exemplo concreto no caso do lançamento de moedas onde θ é a probabilidade de ocorrer caras. Nesse caso, em n lançamentos com K caras, a verossimilhança fica

$$P(\bar{x}|\theta) = l(\theta) = \theta^K (1-\theta)^{n-K}$$

Já vimos escolher a distribuição a priori para θ . Uma possibilidade é usar a distribuição beta

$$B(a, b) \sim x^{a-1} (1-x)^{b-1} \quad 0 \leq x \leq 1$$

com $a=b=6$. Nesse caso, como $P(\bar{x})$ não depende de θ podemos escrever

$$P(\theta|\bar{x}) \propto P(\bar{x}|\theta) P(\theta)$$

$$\begin{aligned} \ln P(\theta|\bar{x}) &\propto \ln P(\bar{x}|\theta) + \ln P(\theta) \\ &\propto \ln l(\theta) + \ln B(6, 6) \end{aligned}$$

Derivando com relação a θ

$$\left. \frac{\partial}{\partial \theta} \ln P(\theta|\bar{x}) \right|_{\theta=\hat{\theta}_{MAP}} = 0 \Rightarrow \hat{\theta}_{MAP} = \frac{K+5}{n+10}$$

~~seja~~

Note que esse estimador é enviesado, pois

$$E(\hat{\theta}_{MAP}) = \frac{E(K) + 5}{n+10} = \frac{5 + n\theta}{n+10} \neq \theta$$

Esse viés reflete a escolha que fizemos para $P(\theta)$, a qual é concentrada ao redor de $\theta=1/2$. Note que se $\theta=1/2$,

$$E(\hat{\theta}_{MAP}) = \frac{5 + \frac{n}{2}}{n+10} = \frac{1}{2}$$

ou seja, não temos viés nesse caso.

Podemos calcular também $E(\hat{\theta}_{MAP}^2)$ e $V(\hat{\theta}_{MAP})$, obtendo

$$E(\hat{\theta}_{MAP}) = \frac{25 - 10n\theta + n\theta(n-1)\theta + 1}{(10+n)^2}$$

$$V(\hat{\theta}_{MAP}) = \frac{n(1-\theta)\theta}{(n+10)^2} \quad \text{deliberadamente}$$

Podemos comparar esse último resultado com o caso da máxima verossimilhança

$$V(\hat{\theta}_{ML}) = \frac{\theta(1-\theta)}{n}$$

Portanto, a razão

$$\frac{V(\hat{\theta}_{MAP})}{V(\hat{\theta}_{ML})} = \frac{n^2}{(n+10)^2} < 1$$

e, assim, $V(\hat{\theta}_{MAP}) < V(\hat{\theta}_{ML})$. Desse modo, para valores não tão grandes de n , a estimativa MAP é mais precisa do que a ML se θ estiver ao redor $1/2$. Note ainda que essa vantagem desaparece quando $n \rightarrow \infty$.

Uma outra estratégia é usar cada x_k como parâmetro para estimar a distribuição ~~de~~ a

posteriori, isto é,

$$P(\theta | X_{k+1}) = \frac{P(X_{k+1} | \theta) P(\theta | \underline{\theta=x_k})}{P(X_{k+1})}$$

Note que esse caso é mais complicado de se resolver analiticamente, por que cada $P(\theta | X_{k+1})$ é uma função variável aleatória. Por outro lado, esse processo é mais em linha com a ideia da estatística Bayesiana, no sentido que o resultado do estimador é uma distribuição e não um único valor. Podemos pensar esse processo como uma atualização da distribuição $P(\theta | X_{k+1})$ a medida que vemos observações novas valores para X_k .

Ao final desse procedimento, temos a distribuição de valores de θ e podemos encontrar intervalos de confiança para os valores de θ . No caso, esses intervalos são chamados de intervalos de credibilidade.

Exemplo no notebook

Robust Statistics

No caso dos estimadores MLE e MAP, sempre assumimos uma distribuição de probabilidade da qual os dados são independentemente e identicamente distribuídos. A ideia da estatística robusta é construir estimadores que possam ser usados quando essas hipóteses não são todas válidas. Por exemplo, suponha um modelo que funcione bem exceto para alguns pontos. Esse procedimento oferece uma maneira de lidar com esses outliers. (4)

A noção de localização

Trata-se de uma espécie de generalização da ideia de medida de centralidade e pode ser definida como segue. Seja X uma variável aleatória com distribuição F e $\Theta(X)$ uma medida descritiva de F . Dizemos que $\Theta(X)$ é uma medida de localização se para quais a e b , temos

$$\Theta(X+b) = \Theta(X) + b \quad (\text{location invariance})$$

$$\Theta(-X) = -\Theta(X)$$

$$\text{se } X > 0 \text{ então } \Theta(X) > 0$$

$$\Theta(ax) = a\Theta(X) \quad (\text{scale invariance})$$

Essas propriedades capturam a ideia de centralidade da distribuição. Um exemplo desse tipo de medida é a medida amostral $\hat{\mu} = \frac{1}{n} \sum X_i$.

$$\hat{\mu}[X_i + b] = \frac{1}{n} \sum (X_i + b) = b + \frac{1}{n} \sum X_i = b + \hat{\mu}[X]$$

$$\hat{\mu}[-X] = -\hat{\mu}[X]$$

$$\hat{\mu}[ax] = \frac{1}{n} \sum ax_i = a\hat{\mu}[X]$$

Estimativa robusta e contaminação

Podemos imaginar que o dado analisado tenha origem numa distribuição que é "contaminada" por outra, isto é,

$$F(x) = \epsilon G(x) + (1-\epsilon) H(x)$$

Sendo ϵ entre 0 e 1. Assim, nosso dado é uma mistura entre $G(x)$ e $H(x)$, ainda que não sabemos como é essa mistura. Nossa ideia será procurar por um estimador que reflita a localização de $G(x)$ a parte da contaminação de $H(x)$. A situação pode ser ainda pior, pois podem haver mais de um contaminante. Assim os estimadores que estamos procurando devem ser derivados de uma família de distribuições ao invés de uma única.

Estimadores de verossimilhança generalizados

No caso da verossimilhança usual, estamos interessados em maximizar

$$L_\mu(x_i) = \prod_i f_0(x_i - \mu)$$

de modo que o estimador de verossimilhança é

$$\hat{\mu} = \operatorname{argmax}_\mu L_\mu(x_i)$$

A diferença aqui é que não vamos supor que f_0 é a distribuição do dado x_i . Usando

$$\rho = -\log f_0$$

podemos escrever

$$\hat{\mu} = \operatorname{argmin} \sum \rho(x_i - \mu)$$

Supondo que ρ é diferenciável com respeito a μ , temos

$$\sum \psi(x_i - \hat{\mu}) = 0$$

sendo $\psi = \rho'$.

Distribution of M-estimates

Dada uma F , vamos definir $\mu_0 = \mu(F)$ como a solução de

$$E_F(\psi(x - \mu_0)) = 0$$

Nesse caso, podemos mostrar que $\hat{\mu} \rightsquigarrow N(\mu_0, \sigma^2/n)$

com

$$\sigma^2 = \frac{E_F(\psi(x - \mu_0)^2)}{E_F(\psi'(x - \mu_0))^2},$$

ou seja, $\hat{\mu}$ é assintoticamente normal. Podemos definir ainda a chamada razão de eficiência

$$\text{Eff}(\hat{\mu}) = \frac{\sigma_0}{\sigma},$$

sendo σ_0 a variância assintótica do estimador MLE. Essa medida dá uma ideia de quanto a contaminação afeta a amostra. Por exemplo, se dois estimadores com variâncias assintóticas $\sigma_1^2 < \sigma_2^2$, sendo $\sigma_1 = 3\sigma_2$, então, o estimador associado a σ_1 requer 3 vezes mais dados que o segundo para obter a mesma variância do segundo. No caso da média de uma variável normal, temos

$$f_0 \sim e^{-x^2/2} \quad \left| \begin{array}{l} \psi = \rho' \sim x \\ \psi' = 1 \end{array} \right.$$

$$\rho = -\log f \sim \frac{x^2}{2}$$

Desse modo,

$$\sigma = \sqrt{V(X)}$$

Se usarmos a mediana como estimador temos

$$\sigma_{mp} = \frac{1}{4 f(\mu_0)^2}$$

No caso de F (distribuição dos dados) ser

$$F \sim N(0, 1)$$

$$\sigma_{mp} \approx 1,571$$

Assim $\sigma_{mp} \approx \sigma \cdot 1,6$, ou seja, para estimar a mediana com a mesma variância da média, devemos ter 1.6 vezes amostras.

M-estimates como médias ponderadas

Suspomos que $\psi(0) = 0$ e $\psi'(0) \neq 0$, podemos assumir que ψ é aproximadamente linear ao redor de zero:

$$W(x) = \begin{cases} \psi(x)/x & x \neq 0 \\ \psi'(x) & x = 0 \end{cases}$$

De modo que

$$\sum \psi(x_i - \hat{\mu}) = 0$$

se tem

$$\sum W(x_i - \hat{\mu})(x_i - \hat{\mu}) = 0$$

resolvendo para $\hat{\mu}$, temos

$$\hat{\mu} = \sum w_i x_i / \sum w_i$$

sendo $w_i = W(x_i - \hat{\mu})$. Note que essa expressão não é útil do ponto de vista prático, já que w_i depende de $\hat{\mu}$, além de não conhecemos ψ .

Huber functions

Uma escolha para ψ é usar as chamadas funções de Huber

$$\rho_K(x) = \begin{cases} x^2 & |x| \leq K \\ 2K|x| - K^2 & |x| > K \end{cases}$$

devendo a

$$2\psi_K(x) = \begin{cases} x & |x| \leq K \\ \operatorname{sgn}(x)K & |x| > K \end{cases}$$

na qual se $K \rightarrow \infty$ ou $K \rightarrow 0$, temos a média e a mediana. Note que $\psi_\infty(x) = x$, então $W(x) = 1$ e

$$\sum W(x_i - \hat{\mu})(x_i - \hat{\mu}) = 0$$

$$\sum (x_i - \hat{\mu}) = 0$$

$$\hat{\mu} = \frac{1}{n} \sum x_i$$

No caso $K=0$, temos a medianacinda que não seja tão direta a demonstração. Desse modo, a função de Huber representa um tipo de interpolação entre a média e a mediana. Usando essa função, temos

$$W_K(x) = \min \left\{ 1, \frac{K}{|x|} \right\}$$

Falar da figura com $K=2$.

Break down Point

Suponha que m dos pontos em

$$\bar{x} = \sum x_i/n$$

vá para infinito, então, $\bar{x} \rightarrow \infty$. Isso significa que o B.P. desse estimador é 0%. No caso da mediana, se $x_K \rightarrow \infty$ para um certo K , seu valor não diverge e é possível mostrar que B.P. é 50%, o que significa que o significado da metade do dados pode divergir sem afetar o seu valor.

Uma maneira de definir o B.P. é considerar n dados da forma $D = \{(x_i, y_i)\}$ e supor um T , de modo a:

$$T(D) = \hat{\theta}$$

sendo $\hat{\theta}$ o vetor dos parâmetros da regressão T . Tome agora todos os possíveis pontos corrompidos de D , os D' , entre a maior fiação causada é

$$\text{bias}_m = \sup_{D'} \| T(D') - T(D) \|$$

no qual $\sup_{D'}$ leva em conta todos os possíveis conjuntos de m pontos de contaminação. Nesse caso, o B.P. é definido como

$$\epsilon_m = \min \left\{ \frac{m}{n} : \text{bias}_m \rightarrow \infty \right\}$$

Ou seja, a menor fração de dados corrompidos que levará a um bias divergente.

No caso de uma regressão linear, um ponto causa $T(D) \rightarrow \infty$, logo, $\epsilon_m = 1/n$, o qual tende a zero se $n \rightarrow \infty$.

Estimando a escala

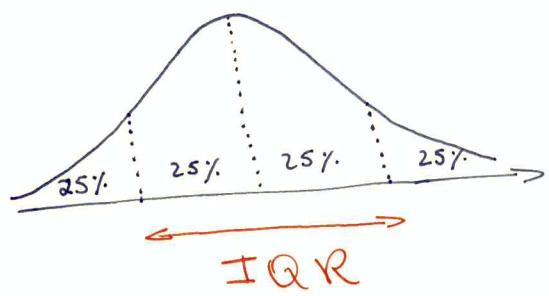
O conceito de escala se refere a uma medida da dispersão dos dados. A medida mais comum nesse caso é o desvio padrão, porém o B.P. dessa quantidade é muito baixo. Uma alternativa mais robusta é

$$MAD = \text{Med}(|x - \text{Med}(x)|)$$

a chamada desvio absoluto da mediana. Outra medida é o interquartil range.

$$IQR = x_{(n-m+1)} - x_{(m)}$$

sendo $m = n/4$, com $x_{(m)}$ n-ésimo elemento dos dados ordenados. Num gráfico teríamos



Exemplo notebook

Bootstrapping

De maneira geral pode ser muito complicado encontrar a distribuição associada a um estimador. A ideia do procedimento de bootstrapping é encontrar uma maneira

de aproximar essa distribuição.

Para apresentar o método, considere uma amostra $\{X_1, X_2, \dots, X_n\}$ com $X_i \sim F$, sendo F desconhecida.

Considerando ainda uma amostra:

$$\{x_1, x_2, x_3, \dots, x_n\}$$

A média amostral é

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

a questão é a saber o quão próximo da verdadeira média Θ , \bar{x} está. Para isso podemos calcular

$$M_2(F) = E_F(x^2) - E_F(x)^2$$

e o erro padrão da média

$$\sigma(F) = \left[M_2(F)/n \right]^{1/2}$$

Entretanto, não conhecemos F . Nesse caso, poderíamos ainda calcular

$$\bar{\sigma} = \left[\bar{M}_2/n \right]^{1/2}$$

sendo

$$\bar{M}_2 = \sum (x_i - \bar{x})^2 / (n-1)$$

o estimador para $M_2(F)$.

Portanto, uma possibilidade melhor seria obter uma estimativa para F a partir de $\{x_1, x_2, \dots, x_n\}$ supondo que cada x_i seja igualmente provável, isto é, a probabilidade seria $1/n$. Denotando por \hat{F}

essa estimativa para \bar{F} , temos

$$\hat{T}_B = [M_2(\hat{F})/n]^{1/2}$$

o chamado estimador bootstrap do erro padrão. Entretanto não há uma maneira direta para estimar \bar{F} .

Por outro lado, podemos pensar em reconstruir $T(F)$ usando o dado disponível, ou seja, produzimos um novo conjunto de n variáveis realizando sorteios aleatórios e independentes com substituição:

$$y^* = \{x_1^*, x_2^*, \dots, x_n^*\}$$

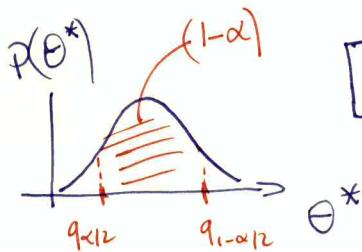
Construímos um conjunto de amostras bootstrap $\{y_k^*\}_{k=1, \dots, B}$ e calculamos a estatística de interesse para cada um (no caso a média \bar{x}). Desse modo, a estimativa bootstrap fica

$$\hat{\theta}_B^* = \frac{1}{B} \sum_{k=1}^B \theta_k^*$$

sendo θ_k^* a estimativa obtida a partir de y_k^* . Desse modo, podemos encontrar também o erro padrão

$$\hat{T}_B^2 = \frac{1}{B-1} \sum [\hat{\theta}_k^* - \hat{\theta}_B^*]^2$$

e intervalos de confiança com significância α



$$[q_{\alpha/2}(\{\theta_k^*\}), q_{1-\alpha/2}(\{\theta_k^*\})]$$

Exemplos Notebook

Bootstrap paramétrico

No exemplo anterior, usamos a amostra para calcular os valores de θ_n^* . Alternativamente poderíamos supor que os dados seguem uma distribuição particular e gerar novas amostras a partir dessa distribuição. Usando essa amostras podemos calcular θ_K^* , um procedimento conhecido como bootstrap paramétrico.

Exemplo notebook

De modo geral a abordagem bootstrap é muito útil para definir intervalos de confiança, especialmente ao lidar com distribuições multivariadas.