

Elementos de Estimação

1. Estimação de Parâmetros

Considere um conjunto de observações $\mathbf{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\}$, as quais dependem de um (ou mais) parâmetro(s) desconhecido(s) $\boldsymbol{\theta}$. O desafio em estimação de parâmetros consiste em determinar o valor de $\boldsymbol{\theta}$ a partir das amostras disponíveis. Para isto, é preciso construir um estimador, o qual pode ser visto como uma regra (função) que gera uma estimativa do valor do parâmetro $\boldsymbol{\theta}$, denotada por $\hat{\boldsymbol{\theta}}$, para cada realização dos dados (\mathbf{X}):

$$\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)})$$

Por simplicidade, considera-se que os dados $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}$ são independentes e identicamente distribuídos (i.i.d.).

Como um estimador é uma função dos dados observados, os quais são obtidos a partir de um processo aleatório, o próprio estimador também possui um caráter estocástico. Por isso, sua caracterização – para fins de análise de desempenho, por exemplo – deve ser feita com base em medidas estatísticas.

1.1. Polarização (*Bias*)

O viés (polarização ou *bias*) de um estimador é dado por*:

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

sendo $E(\cdot)$ o operador esperança (que diz respeito à estrutura probabilística dos dados). Um estimador não-polarizado (*unbiased*) tem viés igual a zero. Um estimador assintoticamente não-polarizado (*asymptotically unbiased*) é aquele para o qual vale:

$$\lim_{m \rightarrow \infty} bias(\hat{\theta}) = 0$$

* Por simplicidade, faremos a exposição das propriedades de um estimador tendo em vista o caso escalar.

1.2. Variância

A variância de um estimador $\hat{\theta}$ é dado por:

$$\text{var}(\hat{\theta}) = E \left([\hat{\theta} - E(\hat{\theta})]^2 \right)$$

Exemplos:

Consideremos uma função de massa de probabilidade Bernoulli. A variável pode assumir valores $X = 1$ (com probabilidade p) e $X = 0$ (com probabilidade $(1 - p)$) (GOODFELLOW ET AL., 2016). Naturalmente,

$$E(X) = 0(1 - p) + p = p$$

É comum expressar a função de massa da seguinte forma:

$$P(x, p) = p^x (1 - p)^{1-x}$$

Consideremos o seguinte estimador de média:

$$\hat{p} = \frac{1}{m} \sum_{k=1}^m x_k$$

Analisemos seu viés:

$$\text{bias}(\hat{p}) = E[\hat{p}] - p$$

Então,

$$\text{bias}(\hat{p}) = E\left[\frac{1}{m} \sum_{k=1}^m x_k\right] - p = \frac{1}{m} \sum_{k=1}^m E[x_k] - p = p - p = 0$$

Isso significa que o estimador é não-polarizado.

Consideremos agora uma densidade de probabilidade gaussiana:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x - \mu)^2}{2\sigma^2}\right]$$

Utilizemos o mesmo estimador de média:

$$\hat{\mu} = \frac{1}{m} \sum_{k=1}^m x_k$$

Então,

$$\text{bias}(\hat{\mu}) = E \left[\frac{1}{m} \sum_{k=1}^m x_k \right] - \mu = \frac{1}{m} \sum_{k=1}^m E[x_k] - \mu = \mu - \mu = 0$$

Portanto, o estimador também é não-polarizado.

Consideremos agora o seguinte estimador de variância da gaussiana (que utiliza o estimador de média mostrado anteriormente):

$$\hat{\sigma}^2 = \frac{1}{m} \sum_{k=1}^m (x_k - \hat{\mu})^2$$

É possível mostrar que esse estimador é polarizado:

$$E[\hat{\sigma}^2] = \frac{m-1}{m} \sigma^2 \rightarrow \text{bias}(\hat{\sigma}^2) \neq 0$$

Um estimador não-polarizado para a variância seria:

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{k=1}^m (x_k - \hat{\mu})^2$$

Após termos considerado esse estimador de variância, analisemos a variância de um estimador. Retomemos o estimador de média da variável Bernoulli.

$$\hat{p} = \frac{1}{m} \sum_{k=1}^m x_k$$

A variância do estimador seria:

$$\text{Var}(\hat{p}) = \frac{1}{m} p(1 - p)$$

A variância diminui com o aumento do número de amostras, o que indica, grosso modo, que ter mais dados leva a estimativas mais “concentradas” em torno do valor correto.

2. Técnicas de Estimação

2.1. Estimação por Máxima Verossimilhança (ML)

Consideremos que seja nosso desejo estimar um conjunto de parâmetros θ a partir de um conjunto de m exemplos i.i.d. $\mathbf{X} = \{\mathbf{x}^{(1)} \dots \mathbf{x}^{(m)}\}$. Havendo um modelo probabilístico, pode-se definir a verossimilhança $\mathcal{L}(\theta) = p(\mathbf{X}|\theta)$. O processo de estimação por máxima verossimilhança (ML, do inglês *maximum-likelihood*) tem por base a maximização de $\mathcal{L}(\theta)$ com respeito a θ .

$$\theta_{\text{ML}} = \arg \max_{\theta} p(\mathbf{X}|\theta) = \arg \max_{\theta} \left(\prod_{k=1}^m p(\mathbf{x}^{(k)}; \theta) \right)$$

Como a função logaritmo é monotonicamente crescente, é possível trabalhar com o logaritmo da verossimilhança (*log-likelihood*) no processo de otimização. Isso leva ao seguinte critério:

$$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \log p(\mathbf{X}|\boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta}} \left(\sum_{k=1}^m \log p(\mathbf{x}^{(k)}; \boldsymbol{\theta}) \right)$$

2.2. Estimação MAP (Máximo a Posteriori)

Outra abordagem possível é realizar a estimação a partir da densidade a posteriori. Nesse caso, os parâmetros também são considerados variáveis aleatórias. O problema se torna então:

$$\boldsymbol{\theta}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X})$$

Usando a regra de Bayes, aplicando o logaritmo e desconsiderando $p(\mathbf{X})$, que não depende de $\boldsymbol{\theta}$, temos o problema da seguinte forma:

$$\boldsymbol{\theta}_{\text{MAP}} = \arg \max_{\boldsymbol{\theta}} \log[p(\mathbf{X}|\boldsymbol{\theta})] + \log[p(\boldsymbol{\theta})]$$

Percebe-se que a otimização inclui a *log-likelihood* e também a densidade associada ao parâmetro (a priori). Este ponto é o grande diferencial deste método de estimação: a possibilidade de incorporar informação a priori.

3. Referência bibliográfica

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. **Deep Learning**, MIT Press, 2016.

KAY, S. M. **Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory**, Prentice Hall, 1993.