# Elementos de Estimação

## 1. Estimação de Parâmetros

Uma tarefa importante para nós será a de estimar parâmetros a partir de dados amostrados. Um estimador pontual é uma função da seguinte forma (GOODFELLOW ET AL., 2016):

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

sendo $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, ..., $\mathbf{x}^{(m)}$ um conjunto de dados independentes e identicamente distribuídos (i.i.d.). Notem que o estimador é uma variável aleatória.

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

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

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

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

A variância de um estimador $\hat{\theta}$ é simplesmente sua variância com respeito aos dados (lembrem-se de que se trata de uma variável aleatória).

**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:

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

Então,

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

Isso significa que o estimador é não-viesado.

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,

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

Portanto, o estimador também é não-viesado.

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 viesado:

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

Um estimador não-viesado 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:

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

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.

## 1.1. Estimação por Máxima Verossimilhança (ML)

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

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

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:

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

## 1.2. 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:

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

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.

# 2. Referência bibliográfica

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