Redução de Dados

Princípios de Redução de Dados

Luis Antonio Fantozzi Alvarez

Universidade de São Paulo

Panorama desta aula

- Nesta aula, nós discutiremos o conceito de suficiência e seus derivados.
  - Conceitos capturam ideia crucial de redução de dados: como reduzir a informação nos dados a um conjunto "pequeno" de estatísticas, sem perder informação relevante para o problema de interesse?
  - Estes conceitos são úteis na construção de estimadores e testes de hipóteses com propriedades desejáveis.
- Em seguida, nós veremos que, em famílias exponenciais, a redução de dados toma uma forma bastante conveniente.
  - Veremos que diversas distribuições conhecidas pertencem à família exponencial.
  - Veremos, sob quais condições estatísticas, com boas propriedades de redução de dados são facilmente recuperáveis em famílias exponenciais
- Referências desta aula: Casella e Berger (2001), com tópicos de Schervish (1995), Lehmann e Romano (2005) e Lehmann e Casella (1998).

Suficiência

Ambiente

- Nosso ponto de partida é um espaço mensurável (Ω, Σ).
  - Ω é o espaço amostral.
  - Σ é o espaço de eventos aos quais podemos atribuir probabilidades (σ-álgebra).
- Pesquisador observa uma amostra, dada pela variável aleatória $\boldsymbol{X} := \mathrm{Id}_\Omega$, cuja lei é dada por uma probabilidade P sobre (Ω, Σ) desconhecida.
- O pesquisador postula uma família de leis de probabilidade $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ candidatas a terem gerado a amostra.
  - Θ é o espaço de parâmetros que indexa a família.
- Exemplo: pesquisador observa uma variável escalar, e supõe que tenha sido amostrada de uma distribuição normal com média desconhecida e variância unitária.
  - Quem é $\mathcal{P}$?
  - E se a variância é desconhecida?
  - E se pesquisador possui n observações independentes?

Hipóteses técnicas

- No que segue, supomos que existe uma medida μ tal que todo elemento de $P_\theta \in \mathcal{P}$ admite densidade p_θ com respeito a μ.
  - Para famílias de distribuições discretas, μ é a medida de contagem e p_θ são as f.m.p.
  - Para famílias de distribuições contínuas, μ é a medida de Lebesgue e p_θ são as f.d.p.
- Uma estatística é uma transformação (mensurável) da amostra, i.e. uma função $T(\boldsymbol{X})$.
- No que segue, vamos supor que, para qualquer estatística, as distribuições condicionais de $\boldsymbol{X}|T(\boldsymbol{X})$ sob cada P_θ admitem uma densidade condicional (regular) $p_{\boldsymbol{X}|T(\boldsymbol{X}),\theta}(\cdot|\cdot)$.
  - Essa condição limita a complexidade de (Ω, Σ).
  - Na maior parte dos casos práticos é satisfeita.
  - Condições suficientes em Durrett (2019, Seção 4.1.3) (fora do escopo do curso).

Estatística suficiente

- Uma estatística é tão somente uma transformação (mensurável) da amostra, isto é, uma função $T(\boldsymbol{X})$ tal que $T^{-1}(A) \in \Sigma$ para todo A na σ-álgebra do contradomínio.

Definição

Uma estatística T é dita suficiente para θ se a distribuição condicional de $\boldsymbol{X}|T(\boldsymbol{X})$ não depende de θ, isto é, se existe H tal que:

$$P_\theta[\boldsymbol{X} \in A|T(\boldsymbol{X})] = H(A|T(\boldsymbol{X})), \quad \forall \theta \in \Theta, A \in \Sigma\,.$$

- Uma vez que conhecemos $T(\boldsymbol{X})$, não há mais informação adicional sobre θ na amostra.

- Sob nossas condições técnicas, a definição é equivalente a que a densidade condicional $p_{\boldsymbol{X}|\boldsymbol{T}(X),\theta}$ não dependa de θ.

Exemplo

Suponha que a amostra consista de duas Bernoullis independentes e identicamente distribuídas, com parâmetro $\theta \in (0,1)$ desconhecido. Neste caso, a f.m.p. é:

$$P_\theta[X_1 = x, X_2 = y] = \theta^x (1-\theta)^{1-x} \theta^y (1-\theta)^{1-y}, \quad \forall x, y \in \{0,1\},$$

de onde segue que, para todo $t \in \{0, 1/2, 1\}$.

$$P_\theta[X_1 = x, X_2 = y | X_1 + X_2 = 2t] = \frac{P_\theta[X_1 = x, X_2 = y, X_1 + X_2 = 2t]}{P_\theta[X_1 + X_2 = 2t]} =$$

$$= \frac{\theta^{x+y}(1-\theta)^{2-x-y}\mathbf{1}\{x+y=2t\}}{\binom{2}{2t}\theta^{2t}(1-\theta)^{1-2t}} = \mathbf{1}\{x+y=2t\},$$

de onde concluímos que $(X_1 + X_2)/2$ é suficiente.

Lema da fatoração de Neyman-Fisher

Lema

$T(\boldsymbol{X})$ é suficiente para θ se, e somente se, existem funções h_θ, $\theta \in \Theta$, e c tais que, para todo $\theta \in \Theta$:

$$p_\theta(\boldsymbol{x}) = h_\theta(T(\boldsymbol{x}))c(\boldsymbol{x}), \quad \mu\text{-q.t.p.}$$

- Critério conveniente para encontrar uma estatística suficiente.
    - No caso discreto, "μ-q.t.p." pode ser lido como para todo $\boldsymbol{x}$.
    - No caso contínuo, condição pode ser violada num conjunto de medida de Lebesgue zero (por exemplo, conjuntos enumeráveis de pontos).

Exemplo

Suponha que o pesquisador observe uma amostra iid $X_1, X_2, \ldots, X_n$ de uma distribuiçao exponencial com parâmetro $\lambda > 0$ desconhecido. Neste caso, observe que:

$$p_\lambda(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} (\lambda e^{-\lambda x} \mathbf{1}\{x > 0\}) =$$

$$= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right) \mathbf{1}\{\min_i x_i > 0\} = h_\lambda\left(\sum_{i=1}^{n} x_i\right) c(\boldsymbol{x}),$$

de onde concluímos que $(X_1 + X_2 + \ldots X_n)/n$ é estatística suficiente.

Estatística suficiente minimal

- Observe que o conceito de estatística suficiente só nos informa sobre a capacidade de uma transformação em condensar a informação relevante na amostra sobre θ.
- Este conceito não versa sobre o "tamanho" desta estatística.
  - De fato, a própria amostra, $T(\boldsymbol{X}) = \boldsymbol{X}$, é sempre uma estatística suficiente.
- Note que, se T é estatística suficiente e $T = M \circ U$, então U é estatística suficiente.
  - Estatísticas mais "finas" que uma estatística suficiente são estatísticas suficientes.
- Na outra direção, podemos pensar na estatística suficiente mais "grossa" possível.

Estatística suficiente minimal (cont.)

Definição

Uma estatística T é dita suficiente minimal, se:

1. T é suficiente.
2. Para qualquer outra estatística S suficiente, existe M tal que $T = M \circ S$.

Lema

Considere uma estatística T tal que:

$$T(\boldsymbol{x}) = T(\boldsymbol{y}) \iff \boldsymbol{y} \in D(\boldsymbol{x}),$$

onde

$$D(\boldsymbol{x}) = \{\boldsymbol{y} : p_\theta(\boldsymbol{y}) = p_\theta(\boldsymbol{x})h(x, y), \quad \forall\theta \in \Theta \text{ e algum } h(x, y) > 0\}.$$

Então T é suficiente minimal.

Exemplo

Suponha que o espaço amostral é $\mathbb{R}_+^n$. Considere uma amostra aleatória de $U[0, \theta]$, $\theta > 0$. Neste caso:

$$p_\theta(x_1, \ldots, x_n) = \frac{1}{\theta^n} \mathbf{1}\{\max_i x_i < \theta\}.$$

Pelo critério de fatoração, $X_{(n)} := \max_i X_i$ é suficiente. Vamos mostrar que é minimal suficiente usando o lema anterior. Observe que $D(\boldsymbol{x}) = \{\boldsymbol{y} : y_{(n)} = x_{(n)}\}$. De fato, se $y_{(n)} \neq x_{(n)}$, ao considerar $\theta' = (x_{(n)} + y_{(n)})/2$, teremos que $0 = p_{\theta'}(\boldsymbol{x}) < p_{\theta'}(\boldsymbol{y})$ ou $0 = p_{\theta'}(\boldsymbol{y}) < p_{\theta'}(\boldsymbol{x})$. Segue do lema anterior que $X_{(n)}$ é minimal.

Ancilaridade

Definição

Uma estatística é dita ancilar para θ se sua distribuição não depende de θ, i.e., se existe F tal que:

$$P_\theta[T(X) \in A] = F[A], \quad \forall \theta \in \Theta, A \in \mathcal{S},$$

onde $\mathcal{S}$ é σ-álgebra acoplada ao contradomínio de T.

Exemplo

No modelo linear Gaussiano homocedástico com regressores fixos:

$$\boldsymbol{y}_{n\times 1} \sim \mathcal{N}\left(\boldsymbol{Z}\beta, \sigma^2 \mathbb{I}_n\right),$$

onde posto$(\boldsymbol{Z}) = k$, $\beta \in \mathbb{R}^k$ desconhecido e $\sigma^2 > 0$ conhecido; veremos em Econometria I que a estatística $S = \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, onde $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{Z}\hat{\beta}_{\text{MQO}}$, é ancilar.

Suficiência completa

- Gostaríamos de que uma estatística suficiente fosse independente de estatísticas ancilares, visto que essas não nos trazem informação de θ.

- Conceito apropriado para isto é o de estatística completa suficiente.

Definição

Uma estatística T é dita completa para θ se, para qualquer f mensurável com valores reais:

$$\mathbb{E}_\theta[f(T)] = 0, \forall\theta \in \Theta \implies \mathbb{P}_\theta[f(T) = 0] = 1, \forall\theta \in \Theta$$

Teorema (Basu)

Uma estatística completa suficiente para θ é independente de qualquer estatística ancilar de θ.

Suficiência completa vs. suficiência minimal

Teorema (Bahadur)

Se U é estatística completa suficiente de dimensão finita, então é suficiente minimal.

- Recíproca **não** é verdadeira: existem estatísticas minimais suficientes de dimensão finita que não são completas (veja Lehmann e Casella, 1998).

Famílias Exponenciais

Família exponencial

Definição

Uma família de distribuições $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ sobre um espaço mensurável (Ω, Σ) é dita uma família exponencial se:

1. Existe uma medida μ tal que cada elemento $P_\theta \in \mathcal{P}$ admite densidade p_θ com respeito a μ.
2. Existem $\eta : \Theta \mapsto \mathbb{R}^s$, $T : \Omega \mapsto \mathbb{R}^s$, $B : \Theta \mapsto \mathbb{R}$ e $h : \Omega \mapsto \mathbb{R}_+$ tais que:

$$p_\theta(\boldsymbol{x}) = \exp(\eta(\theta)' T(\boldsymbol{x}) - B(\theta)) h(\boldsymbol{x})$$

- Diversas distribuições conhecidas pertecem a esta família: Gamma, Chi-Quadrado, Beta, Normal, Poisson, Negativo-Binomial.
- Propriedade útil: se $\mathcal{P}_1, \ldots, \mathcal{P}_n$ são famílias exponenciais, então $\mathcal{P}^n := \{\otimes_{i=1}^n P_i : P_i \in \mathcal{P}_i\}$ é uma família exponencial.
  - Consequência: a distribuição amostral de uma amostra aleatória de uma família exponencial também constitui uma família exponencial.

Exemplo

Sejam $\boldsymbol{X} = (X_1, X_2, \dots, X_n)'$ uma amostra aleatória de $N(\mu, \sigma^2)$, onde $\mu \in \mathbb{R}$ e $\sigma^2 > 0$ são desconhecidos. Neste caso, fazendo $\theta = (\mu, \sigma^2)'$, temos:

$$\begin{gathered} p_\theta(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) = \\ \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{\mu^2}{2\sigma^2}n\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n, \end{gathered} \tag{1}$$

de onde segue que a família é exponencial com $T(\boldsymbol{x}) = (\sum_{i=1}^{n} x_i^2, \sum_{i=1}^{n} x_i)'$ e $\eta(\theta) = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)'$.

Identificabilidade de famílias de distribuições

Definição

Para uma família $\{P_\theta : \theta \in \Theta\}$, dizemos que o parâmetro é identificável se:

$$\theta = \theta' \iff P_\theta = P_{\theta'}$$

- Identificação é o requerimento básico para estimação pontual.
  - Requer a existência de um mapa 1-1 entre parâmetro e distribuição amostral.
- No exemplo anterior, θ é identificado, visto que $f_1(P_\theta) := \mathbb{E}_\theta[X_1] = \mu$ e $f_2(P_\theta) := \mathbb{E}[X_1^2] = \sigma^2 + \mu^2$.
- Por outro lado, se a família paramétrica fosse $\{\mathcal{N}(\theta \vee 0, 1) : \theta \in \mathbb{R}\}$, θ não é identificável.

Parametrização natural de família exponencial

Note que, da definição de família exponencial:

$$p_\theta(\boldsymbol{x}) = \exp(\eta(\theta)' T(\boldsymbol{x}) - B(\theta)) h(\boldsymbol{x}) \,, \quad \theta \in \Theta \,,$$

segue uma reparametrização natural

$$p_{\boldsymbol{\eta}}(\boldsymbol{x}) = \exp(\eta' T(\boldsymbol{x}) - B^*(\boldsymbol{\eta})) h(\boldsymbol{x}) \,, \quad \boldsymbol{\eta} \in \Xi \,,$$

onde $\Xi := \{\eta(\theta) : \theta \in \Theta\} \subseteq \mathbb{R}^s$ é o espaço paramétrico natural; e $B^*(\boldsymbol{\eta}) = B(\theta)$ para algum (qualquer) $\theta \in \Theta$ tal que $\eta(\theta) = \boldsymbol{\eta}$

Definição

Uma família exponencial $\{P_\theta : \theta \in \Theta\}$ é dita uma família de posto cheio se sua reparametrização natural é tal que η é identificável e Ξ contém uma bola aberta de $\mathbb{R}^s$.

- No exemplo de amostra aleatória normal, $\Theta = (-\infty, 0) \times \mathbb{R}$, logo a família tem posto cheio.
- $\{N(\mu, \mu^2) : \mu \in \mathbb{R}\}$ não tem posto cheio (família curvada).

Suficiência em famílias exponenciais

Corolário

Numa família exponencial, $T(\boldsymbol{X})$ é uma estatística suficiente.

Corolário

Numa família exponencial de posto cheio, $T(\boldsymbol{X})$ é uma estatística suficiente minimal.

Teorema

Numa família exponencial de posto cheio, $T(\boldsymbol{X})$ é uma estatística suficiente completa.

Tópicos adicionais

Momentos de T

Teorema

Para uma família exponencial em sua parametrização natural, e $\eta \in \mathrm{int}(\Xi)$, *temos que, para toda h tal que*

$$\int_\Omega |h(\boldsymbol{x})| p_{\boldsymbol{\eta}'}(\boldsymbol{x}) \mu(d\boldsymbol{x}) < \infty \,,$$

para todo $\boldsymbol{\eta}'$ numa vizinhança de η; a função:

$$\boldsymbol{\psi} \mapsto \int_\Omega h(\boldsymbol{x}) p_{\psi}(\boldsymbol{x}) \mu(d\boldsymbol{x}),$$

é diferenciável em η, com derivada dada por diferenciação dentro da integral.

- Aplicação do resultado: momentos de T.

Teorema de Rao-Blackwell

- Considere o problema de estimar um parâmetro $\psi(\theta)$ escalar de uma família P_θ.

- Seja $L(\theta; \hat{\delta})$ a perda incorrida em utilizar o estimador $\hat{\delta}$ para estimar $\psi(\theta)$.

- A perda esperada, ou risco, é dada por $R_\theta(\hat{\delta}) = \mathbb{E}_\theta[L(\theta; \hat{\delta})]$.

Teorema (Rao-Blackwell)

Suponha que $z \mapsto L(\theta, z)$ é convexa e não negativa, para todo $\theta \in \Theta$. Dado um estimador $\hat{\delta}$ de $\psi(\theta)$ e T uma estatística suficiente integrável de θ, temos que:

1. *$S = \mathbb{E}_\theta[\hat{\delta} | T]$ define um estimador.*
2. *$R_\theta(S) \leq R_\theta(\hat{\delta})$ para todo $\theta \in \Theta$.*
3. *Se $\hat{\delta}$ é não viciado, S também o é.*

Referências

Referências

Casella, George e Roger L Berger (2001). *Statistical inference*. Duxbury.

Durrett, Rick (2019). *Probability: Theory and Examples*. 5ª ed. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. DOI: 10.1017/9781108591034.

Lehmann, E e George Casella (1998). *Theory of point estimation*. Springer Science & Business Media.

Lehmann, E e J Romano (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer New York. ISBN: 9780387276052. URL: https://books.google.com.br/books?id=K6t5qn-SEp8C.

Schervish, Mark J. (1995). *Theory of Statistics*. Springer New York. DOI: 10.1007/978-1-4612-4250-5. URL: https://doi.org/10.1007/978-1-4612-4250-5.