Introdução à Inferência Bayesiana

Prof. Caio Azevedo

Introdução

- A inferência bayesiana (IB) é baseada, essencialmente, no teorema de Bayes, que é uma consequência da definição de probabilidade condicional (e o teorema da probabilidade total).
- Com efeito, se A e B são dois eventos do mesmo espaço amostral, então:

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

- Os alicerces da IB são (slide seguinte):

Introdução

a) Assumir que $\boldsymbol{\theta}$ é aleatório, em certo sentido, (va, para modelar a incerteza que temos sobre ele), com suporte $\Theta \subseteq \mathcal{R}^k$

b) A descrição desse comportamento aleatório (incerteza) é especificado através da distribuição a priori ($\pi(.)$) (não necessariamente uma distribuição de probabilidade).

c) O modelo estatístico é especificado por $f_X(.|\boldsymbol{\theta}) \equiv f_X(.,\boldsymbol{\theta})$, cuja interpretação é a distribuição de $X|\theta$ (X dado que θ assume o valor θ. Isto é, $\forall x \in \mathcal{X}$, isto é

$$f_{X|\theta}(x|\theta) = f_{X|H=\theta}(x|\theta)$$

em que: $H : \Omega \rightarrow \Theta$.

Introdução

- Além disso, $f_H(\theta) = \pi(\theta)$ é a distribuição a priori, enquanto que a distribuição de $H|X = x$, $\pi(\theta|x) = f_{X|X=x}(\theta|x)$ é a distribuição a posteriori. Das relações anteriores, temos que

$$
\begin{aligned}
\pi(\theta|x) &= f_{H|X=x}(\theta|x) = \frac{f_{X,H}(x,\theta)}{f_X(x)} = \frac{f_{X|H=\theta}(x|\theta)f_X(\theta)}{f_X(x)} \\
&= \frac{f(x|\theta)f(\theta)}{f_X(x)} = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f_{X|H=\theta}(x|\theta)f_X(\theta)d\theta} \\
&= \frac{f(x|\theta)\pi(\theta)}{\mathcal{E}_\theta(f(x|H))}
\end{aligned}
$$

Introdução

- Portanto, $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$, em que o símbolo "$\propto$" denota que os membros da relação acima são proporcionais em relação à θ. Isto é, $\exists$ uma constante c que não depende de θ, tal que:

$$\begin{aligned} \pi(\theta|x) &= cf(x|\theta)\pi(\theta) \\ c &= \int_{\Theta} f(x|\theta)\pi(\theta)d\theta \end{aligned}$$

- Especificação do modelo bayesiano: $X|\theta \sim f_X(.|\theta)$.

Introdução

- Exemplo: Seja $X|\theta \sim$ Bernoulli(θ), $\theta \in (0,1)$ e assuma $\theta \sim$ beta(α, β), α e β conhecidos.
- A distribuição a posteriori é dada por:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Theta} f(x|\theta)\pi(\theta)d\theta}$$

Introdução

- Então,

$$\begin{aligned}\pi(\theta|x) &\propto \theta^{x}(1-\theta)^{1-x}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{(0,1)}(\theta) \\ &= \theta^{x+\alpha-1}(1-\theta)^{\beta-x+1-1}\mathbb{1}_{(0,1)}(\theta)\end{aligned}$$

- Logo, $\theta|X = x \sim \text{beta}(\alpha + x, \beta - x + 1)$.
- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X|\theta \sim \text{Bernoulli}(\theta)$, encontre a posteriori, sob $\theta \sim \text{beta}(\alpha, \beta)$. Temos que:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\theta) = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\prod_{i=1}^{n}\mathbb{1}_{\{0,1\}}(x_i)$$

Introdução

- Assim, vem que:

$$
\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto f(x|\theta)\pi(\theta) \\
&= \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}\theta^{\alpha-1}(1-\theta)^{\beta-1}\mathbb{1}_{(0,1)}(\theta) \\
&= \theta^{\alpha+\sum_{i=1}^{n} x_i-1}(1-\theta)^{\beta+n-\sum_{i=1}^{n} x_i-1} \qquad (1)
\end{aligned}
$$

que corresponde ao núcleo de uma distribuição (parte que depende de θ) beta$(\alpha+\sum_{i=1}^{n} x_i, \beta+n-\sum_{i=1}^{n} x_i)$. Logo,

$$
\theta|\boldsymbol{x} \sim \text{beta}\left(\alpha+\sum_{i=1}^{n} x_i, \beta+n-\sum_{i=1}^{n} x_i\right).
$$

Introdução

- Note, portanto, que o "estimador" bayesiano é a própria distribuição a posteriori.
- Por exemplo, dela podemos obter :
  - Estimativas pontuais (média, moda e mediana a posteriori).
  - Intervalo de credibilidade (quantis de interesse).
  - Calcular as probabilidades de ocorrência de hipóteses (testes de hipótese).

Introdução

- Por exemplo, $\widehat{\theta}_{EAP} = \mathcal{E}(\theta|\boldsymbol{X}) = \frac{\alpha+\sum_{i=1}^{n} X_i}{\alpha+\beta+n}$. Suponha que observemos $x_1 = x_2 = x_3 = 0$ e $\alpha = \beta = 1$. Assim, teríamos $\widetilde{\theta}_{MV} = 0$ e $\widetilde{\theta}_{EAP} = 1/5$. Qual estimador é melhor?
- No exemplo vimos que priori(beta(α, β)) $\rightarrow$ posteriori(beta(α^*, β^*)). Ou seja, a priori e a posterior pertecem à mesma família.
- Nesse caso, dizemos que o modelo Bernoulli (verossimilhança) é conjugado ao modelo beta.
- Isto ocorre, nesse caso, porque

$$L(\theta) \propto \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i}$$

Introdução

- Suponha que a distribuição de $\boldsymbol{X}$ é dada por:

$$f_{\boldsymbol{X}}(\boldsymbol{x};\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})t_j + d(\boldsymbol{\theta})\right\}, \boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^k$$

em que $\Omega = \left\{\boldsymbol{t} \in \mathcal{R}^{k+1}, w(\boldsymbol{t}) < \infty\right\}$, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$,
$\boldsymbol{t} = (t_1, ..., t_{k+1})'$ e $w(\boldsymbol{t}) = \int \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})t_j + t_{k+1}d(\boldsymbol{\theta})\right\} d\boldsymbol{\theta}$.

Introdução

- Teorema: A família exponencial $k+1$ paramétrica. Note que:

$$\pi(\boldsymbol{\theta}|\boldsymbol{t}) \approx \pi(\boldsymbol{\theta}) = \frac{1}{w(\boldsymbol{t})} \exp\left\{\sum_{j=1}^{k} c_j(\boldsymbol{\theta})t_j + t_{k+1}d(\boldsymbol{\theta})\right\}$$

à priori, é a distribuição conjugada de $f(\boldsymbol{x}|\boldsymbol{\theta})$ (dado anteriormente), ou seja: $f(\boldsymbol{x}|\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta}) \to \pi(\boldsymbol{\theta}|\boldsymbol{x})$

Cont.

- Exemplo: Seja $X|\theta \sim N(\theta, \sigma^2)$, σ^2 conhecido.

$$
\begin{aligned}
f(x|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \exp\left\{\frac{\theta x}{\sigma^2} + \frac{\theta^2}{2/\sigma^2}\right\} \\
&= h(x) \exp c_1(\theta) t_1(x) + d(\theta)
\end{aligned}
$$

em que $c_1(\theta) = \frac{\theta}{\sigma^2}$, $t_1(x) = x$, $d(\theta) = -\frac{\theta^2}{2\sigma^2}$.

- Aplicando o resultado do teorema acima (k=1), temos que:

$$
\begin{aligned}
\pi(\theta|t_1, t_2) &= \frac{1}{w(t_1, t_2)} \exp\left\{\frac{\theta t_1}{\sigma^2} - \frac{\theta^2 t_2}{2\sigma^2}\right\} \\
&= \exp\left\{-\frac{-t_2}{2\sigma^2}\left(\theta^2 - 2\frac{t_1}{t_2} + \frac{t_1^2}{t_2}\right) + \frac{t_1^2}{2\sigma^2 t_2}\right\} \\
&\propto \exp\left\{-\frac{-t^2}{2\sigma^2}\left(\theta - \frac{t_1}{t_2}\right)^2\right\}
\end{aligned}
$$

Cont.

- Assim,

$$\theta | t_1, t_2 \sim N\left(\frac{t_1}{t_2}, \frac{\sigma^2}{t_2}\right)$$

- Ou seja, nesse caso a família conjugada é $N(\eta, \sigma^2)$.
- Obs: a) O Teorema anterior também vale sob aa, ou seja, ou seja

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left\{\sum_{j=1}^{k} c_i(\boldsymbol{\theta}) \sum_{i=1}^{n} t_j(x_i) + nd(\theta)\right\}$$

- Então se,

$$\pi(\theta|\boldsymbol{t}) = \frac{1}{w(\boldsymbol{t})} \exp\left\{\sum_{j=1}^{n} c_j(\boldsymbol{\theta}) t_j + t_{k+1} d(\boldsymbol{\theta})\right\}$$

é a distribuição a priori (família conjugada) de $\boldsymbol{\theta}$.

Cont.

- Logo, de acordo com o teorema, a distribuição a posteriori é dada por:

$$\pi(\theta|\boldsymbol{x}) = \pi\left(\boldsymbol{\theta}|t_1 + \sum_{i=1}^{n} t_1(x_i), t_2 + \sum_{i=1}^{n} t_2(x_i), ..., t_{k+1} + n\right)$$

- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X|\mu \sim N(\mu, \sigma^2)$ com σ^2, conhecido.
- Seguindo o exemplo anterior, temos que:

$$\pi(\theta|\boldsymbol{x}) = \pi(\theta|t_1 + \sum_{i=1}^{n} t_i, t_2 + n) \text{ (ou seja)}$$

$$\theta|\boldsymbol{X} = \boldsymbol{x} \sim N\left(\frac{t_1 + \sum_{i=1}^{n} x_i}{t_2 + n}, \frac{\sigma^2}{t_2 + n}\right)$$

Cont.

- Por exemplo,

$$\mathcal{E}(\theta|\boldsymbol{X}) = \frac{t_1 + \sum_{i=1}^{n} X_i}{t_2 + n}(\text{estimador}); \mathcal{E}(\theta|\boldsymbol{X}) = \frac{t_1 + \sum_{i=1}^{n} x_i}{t_2 + n}(\text{estimativa})$$

- A priori de θ é $N(\eta, \tau^2)$, em que:

$$\eta = \frac{t_1}{t_2}; \tau^2 = \frac{\sigma^2}{t_2} \text{ ou } t_2 = \frac{\sigma^2}{\tau^2} \text{ ou } t_1 = \frac{\eta\sigma^2}{\tau^2}$$

- Dessa forma:

$$\begin{aligned}
\mathcal{E}(\theta|\boldsymbol{X}) &= \frac{\frac{n\sigma^2}{\tau^2} + \sum_{i=1}^{n} x_i}{\frac{\sigma^2}{\tau^2} + n} = \frac{(\sigma^2/\tau^2)\eta}{\frac{\sigma^2}{\tau^2} + n} + \frac{n\overline{x}}{\frac{\sigma^2}{\tau^2} + n} \\
&= \alpha\eta + \beta\overline{x}, \alpha + \beta = 1
\end{aligned}$$

Cont.

- Exemplo: Sejam $X_1, ..., X_n$ aa de $X|\sigma^2 \sim N(\theta, \sigma^2)$, θ conhecido. Assim:

$$f_{\boldsymbol{X}}(\boldsymbol{x}|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\}$$

- Sejam $\psi = \sigma^2$ e $s = \sum_{i=1}^{n}(x_i - \theta)^2$. Assim

$$f(\boldsymbol{x}|\psi) \propto \exp\left\{-\frac{s}{2\psi}\right\}\psi^{-n/2}$$

- Vamos considerar que a priori de ψ é dada por:

$$\pi(\psi) \propto \psi^{-k/2} \exp\left\{-\frac{s_0}{2\psi}\right\}$$

em que k e s_0 são conhecidos.

Cont.

- Então se $\lambda = \frac{s_0}{\psi} \rightarrow \psi = \frac{s_0}{\lambda}$. Logo $\frac{d\psi}{d\lambda} = -\frac{s_0}{\lambda^2}$.
- Portanto,

$$\begin{aligned}\pi(\lambda) &\propto \left(\frac{s_0}{\lambda}\right)^{-k/2} e^{-\lambda/2}\frac{s_0}{\lambda^2} \\ &\propto \lambda^{k/2-2}e^{-\lambda/2} = \lambda^{(k-4)/2}e^{-\lambda/2}\end{aligned}$$

- Seja agora, $X \sim \text{gama}(\alpha, 1/\beta)$, $f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$.
- Se $\alpha = k/2$ e $\beta = 1/2$, então

$$f_x(x) = \frac{1}{2^k\Gamma(k/2)}x^{k/2-1}e^{-x/2}, X \sim \chi^2_k$$

Cont.

- Então, se $\nu = k - 2$, temos que:

$$\pi(\lambda) \propto \lambda^{k/2-1} e^{-\lambda/2} = \lambda^{\nu/2-1} e^{-\lambda/2}, \lambda \sim \chi^2_\nu$$

- Note que $\psi = \frac{s_0}{\lambda} \to \psi \sim \frac{s_0}{\chi^2_\nu}$.
- Assim,

$$\pi(\lambda) = f_\lambda(\lambda) \frac{d\lambda}{d\psi} \propto \psi^{-(\nu/2-1)} e^{-s_0/\psi}$$

Cont.

- Ou seja

$$\pi(\psi) = c\psi^{-(\nu/2-1)}e^{-s_0/\psi}, c = \int_0^{\infty} \psi^{-(\nu/2-1)}e^{-s_0/\psi}d\psi$$

- Portanto, a distribuição a posteriori de ψ é dada por

$$\begin{aligned} \pi(\psi|\boldsymbol{x}) &\propto f(\boldsymbol{x}|\psi)\pi(\psi) \propto \psi^{-n/2}e^{-s/2\psi}\psi^{-(\nu/2-1)}e^{-s_0/2\psi} \\ &= \psi^{-(n+\nu)/2-1}e^{-(\psi/2)(s+s_0)} \end{aligned}$$

- Dessa forma: $\psi|\boldsymbol{X} = \boldsymbol{x} \sim \frac{s+s_0}{\chi^2_{n+\nu}}$

Cont.

- Logo,

$$\begin{aligned}\mathcal{E}(\psi|\boldsymbol{X}) &= \frac{s_0 + s}{n+\nu-2} = \frac{s_0 + \sum_{i=1}^{n}(x_i-\theta)^2}{n+\nu-2} \\ &= \frac{s_0}{n+\nu-2} + \frac{n}{n+\nu-2}\widetilde{\sigma}^2\end{aligned}$$

- em que $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2$
- Exercício: No exemplo anterior, usando o teorema visto, obtenha a distribuição a priori de σ^2 que seja conjugada com respeito à $N(\theta,\sigma^2)$

Introdução à Teoria da Decisão

- Seja $\boldsymbol{X} = (X_1, ..., X_n)'$ um vetor aleatório com fdp $f_{\boldsymbol{X}}(.|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^k$.
- Os elementos básicos da Teoria da Decisão são:
  1. Θ : espaço paramétrico (estado da natureza).
  2. $\mathcal{A}$: é o conjunto das possíveis ações que podem ser tomadas (exemplos):
     - $\mathcal{A}$: {é o conjunto dos estimadores pontuais}.
     - $\mathcal{A}$: { é o conjunto das funções de teste}.
  3. Uma função de decisão definida por $d : \mathcal{X} \rightarrow \mathcal{A}$, $(\boldsymbol{x} \rightarrow d(\boldsymbol{x}))$ que é formulada para se obter informações sobre $\boldsymbol{\theta}$. Notação: $\mathcal{D}$: {classe de todas as funções de decisão}.
  4. Para avaliar a escolha de uma função de decisão $(d \in \mathcal{D})$, vamos considerar o conceito de função perda.

Função de perda

- A função de perda (as vezes chamada de função perda) é definida por:

$$\begin{aligned} L(d,\theta) &= L(d(\boldsymbol{x}),\theta) \\ (\mathcal{D},\Theta) &\to \mathcal{R}^{+} \end{aligned}$$

  satisfazendo as seguintes propriedades:

  a) $L(\theta, d) \geq 0, \forall d \in \mathcal{D}, \theta \in \Theta$.
  a) $L(\theta, d) = 0, \text{ se } d = \theta$.

- Note que considerarmos $d = d(\boldsymbol{X})$ então $L(d(\boldsymbol{X}), \theta)$ é uma va.

Exemplos de função de perda

- $L_1 = (d - \theta)^2$ (perda quadrática).
- $L_2 = |d - \theta|$ (perda modular ou perda absoluta).
- $L_3 = \begin{cases} 1, & \text{se } |d - \theta| > \epsilon, \epsilon > 0 \\ 0, & \text{se } |d - \theta| \leq \epsilon \end{cases}$ (perda (0-1)).
- $L_4 = \begin{cases} A, & \text{se } |d - \theta| > \epsilon, \epsilon > 0 \\ 0, & \text{se } |d - \theta| \leq \epsilon \end{cases}$ (perda (0 - A)).
- $L_5 = \rho(\theta)|d - \theta|^r, \rho(\theta) \geq 0, \forall\theta, r > 0$ (perda generalizada).

Perda esperada (risco frequentista)

- Para uma dada função de perda, o risco frequentista é definido como sendo o seu valor esperado em relação à distribuição de $\boldsymbol{X}|\theta$, ou seja

$$R(d,\theta) = \mathcal{E}_{\boldsymbol{X}|\theta}[L(d,\theta)] = \begin{cases} \sum_{\mathcal{X}} L(d(\boldsymbol{X}),\theta) f(\boldsymbol{x}|\theta), & \text{se } \boldsymbol{X} \text{ for discreto} \\ \int_{\mathcal{X}} L(d(\boldsymbol{X}),\theta) f(\boldsymbol{x}|\theta) d\boldsymbol{x}, & \text{se } \boldsymbol{X} \text{ for contínuo} \end{cases}$$

- Exemplo [Erro quadrático médio (frequentista)]: $\mathcal{E}_{\boldsymbol{X}|\theta}[(d(\boldsymbol{X}) - \theta)^2]$.
- Em geral, na inferência frequentista, busca-se o estimador que minimiza o risco acima definido (restrito à uma determinada subclasse de estimadores).

Perda esperada (risco frequentista)

- Obs: Se $L(\theta, d) = (\theta - d)^2$, então

$$
\begin{aligned}
R(\theta, d) &= \mathcal{E}_{\boldsymbol{X}|\theta}\left[(\theta - d)^2\right] = \mathcal{EQM}_{\theta}(d) \\
&= \mathcal{V}_{\theta}(d) + B_{\theta}^2(d)
\end{aligned}
$$

- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X \sim N(\mu, \sigma^2)$ e considere a classe de estimadores de σ^2 dada por $d_b(\boldsymbol{X}) \equiv \widehat{S}_b = bS^2$, em que $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ e b é constante.
- Temos que: $\mathcal{E}(S^2) = \sigma^2$, $\mathcal{V}(S^2) = \frac{2\sigma^4}{n-1}$. Seja

$$
\begin{aligned}
R(\theta, \widehat{S}_b) &= R(\sigma^2, \widehat{S}_b) = \mathcal{V}_{\sigma^2}(\widehat{S}_b) + \left[\mathcal{E}_{\sigma^2}(\widehat{S}_b - \sigma^2)^2\right] \\
&= \frac{2b^2\sigma^4}{n-1} + (b\sigma^2 - \sigma^2)^2 = \frac{2b^2\sigma^4}{n-1} + (b-1)^2\sigma^4 = g(b)
\end{aligned}
$$

Perda esperada (risco frequentista)

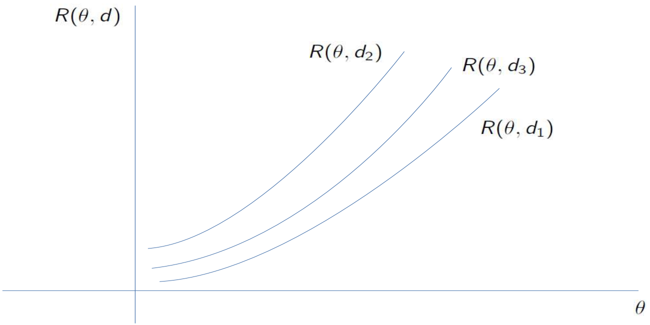
- Portanto, o estimador da classe $\widehat{S}_b$ que tem menor risco (considerando perda quadrática) é dado por:

$$\widehat{S}_b = \frac{n-1}{n+1} S^2 = \frac{1}{n+1} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2$$

- Obs: considere as decisões:

$$\begin{aligned} d_1(\boldsymbol{X}) &= \frac{n-1}{n+1} S^2 (R(\theta, d_1)), \\ d_2(\boldsymbol{X}) &= S^2 (R(\theta, d_2)), \\ d_3(\boldsymbol{X}) &= \widehat{\sigma}^2 (R(\theta, d_3)) \end{aligned}$$

Compar. de riscos (n=5, veja também Casella & Berger (2006))



Perda esperada (risco frequentista)

- Considere agora a função perda (Exercício: mostre que é uma função perda):

$$L(\theta, d) = \frac{d}{\sigma^2} - 1 - \ln\left(\frac{d}{\sigma^2}\right)$$

- Então

$$\begin{aligned} R(\sigma^2, \widehat{S}_b) &= \mathcal{E}_{\boldsymbol{X}|\sigma^2}\left(\frac{\widehat{S}_b}{\sigma^2} - 1 - \ln\frac{\widehat{S}_b}{\sigma^2}\right) \\ &= \frac{b}{\sigma^2}\mathcal{E}_{\boldsymbol{X}|\sigma^2}(S^2) - 1 - \ln(b) + \mathcal{E}\left(-\ln\frac{S^2}{\sigma^2}\right) \\ &\rightarrow g(b) = b - 1 - \ln(b) + \mathcal{E}\left(-\ln\frac{S^2}{\sigma^2}\right) \\ &\rightarrow \frac{\partial g(b)}{\partial b} = 1 - \frac{1}{b} = 0 \leftrightarrow b = 1 \end{aligned}$$

Perda esperada (risco frequentista)

- Def: Dizemos que um procedimento d_1 é melhor do que um procedimento d_2 se

$$R(\theta, d_1) \leq R(d_2, \theta), \forall \theta \in \Theta \text{ e}$$

$$R(\theta, d_1) < R(\theta, d_2), \text{ para pelo menos um } \theta.$$

- Se d_1 e d_2 são tais que: $R(d_1, \theta) = R(d_2, \theta), \forall \theta \in \Theta$ dizemos que d_1 e d_2 são equivalentes.

Risco de Bayes

- O risco de Bayes é definido como a perda média do risco frequentista, à priori, em relação à θ, ou seja

$$r(\pi, d) = \int_{\Theta} R(d, \theta)\pi(\theta)d\theta$$

- Em princípio, tal medida considera (explicitamente) apenas a priori.
- Como introduzir a distribuição a posteriori no risco de Bayes?

Risco de Bayes (cont.)

- Note que

$$
\begin{aligned}
r(\pi, d) &= \int_{\Theta} R(\theta, d)\pi(\theta)d\theta \\
&= \int_{\Theta} \left[\int_{\mathcal{X}} L(\theta, d(\boldsymbol{x}))f(\boldsymbol{x}|\theta)d\boldsymbol{x}\right] \pi(\theta)d\theta \\
&= \int_{\Theta} \left[\int_{\mathcal{X}} L(\theta, d(\boldsymbol{x}))\frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{f(\boldsymbol{x})} f(\boldsymbol{x})d\boldsymbol{x}\right] d\theta \\
&= \int_{\boldsymbol{X}} \underbrace{\left[\int_{\Theta} L(\theta, d(\boldsymbol{X}))\pi(\theta|\boldsymbol{x})d\theta\right]}_{R_B(\pi,d)} f(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\boldsymbol{X}} R_B(\pi, d)f(\boldsymbol{x})d\boldsymbol{x}
\end{aligned}
$$

Risco à posteriori (ou integrado)

- Em geral, $R_B(\pi, d) = \int_{\Theta} L(\theta, d(\boldsymbol{x}))\pi(\theta|\boldsymbol{x})d\theta = \mathcal{E}_{\theta|\boldsymbol{x}}\left(L(\theta, d(\boldsymbol{x}))\right)$ é chamado de risco integrado ou à posteriori de Bayes.
- O estimador Bayesiano (d^{π}) que minimiza $R_B(\delta, p)$ é chamado de estimador de Bayes. Ou seja,

$$r(\pi, d^{\pi}) = \inf_{d \in \mathcal{D}} r(\pi, d) \quad (2)$$

- Encontrar o estimador de Bayes significa minimizar $\mathcal{E}_{\theta|\boldsymbol{X}}(L(\theta, d))$.

Risco à posteriori

- Obs: Segundo a definição acima, pode ocorrer que:
  - Não exista d^{π} em $\mathcal{D}$ que satisfaça (2).
  - Pode existir mais de um estimador de Bayes.
- A priori não precisa ser uma distribuição de probabilidade, tampouco ser própria ($\int_{\Theta} \pi(\theta) d\theta = \infty$).

Estimadores Bayesianos associados à funções de perda específicas

- O estimador EAP é o estimador de Bayes para a função de perda $L_1 = (d - \theta)^2$.
- O estimador MeAP (mediana a posteriori) é o estimador de Bayes para a função de perda $L_2 = |d - \theta|$, em que A é uma constante.
- O estimador MAP (moda a posteriori) é o estimador de Bayes para a função de perda $L_4 = \begin{cases} A, & \text{se } |d - \theta| > \epsilon, \epsilon > 0 \\ 0, & \text{se } |d - \theta| \leq \epsilon \end{cases}$

Demonstração para a função de perda $L_1 = (\delta - \theta)^2$

- Queremos minimizar $g(d) = \int_{\Theta} (d - \theta)^2 \pi(\theta|\boldsymbol{x}) d\theta$ em relação à d. Por outro lado, note que:

$$\begin{aligned} g(d) &= d^2 \int_{\Theta} p(\theta|\boldsymbol{x}) d\theta - 2d \int_{\Theta} \theta p(\theta|\boldsymbol{x}) d\theta + \int_{\Theta} \theta^2 p(\theta|\boldsymbol{x}) d\theta \\ &= d^2 - 2d\mathcal{E}(\theta|\boldsymbol{X}) + \mathcal{E}(\theta^2|\boldsymbol{X}) \end{aligned}$$

- Derivando a expressão acima com relação à d, igualando a expressão resultante à 0 e resolvendo-a em relação à d, temos que

$$\widehat{d} = \widehat{\theta}_{Bayes} = \mathcal{E}(\theta|\boldsymbol{X}).$$

- Obs: Se $\tau(\theta)$ é uma função de θ com distribuição a priori $\pi(.)$, então, sob perda quadrática, o estimador de Bayes é

$$d^{\pi}(\boldsymbol{X}) = \mathcal{E}(\tau(\theta)|\boldsymbol{X} = \boldsymbol{x}).$$

Exemplo

- Seja $X_1, ..., X_n$ uma aa de $X|\theta \sim \text{Poisson}(\theta)$, com $\theta \sim \text{gamma}(a, 1/b)$ (família conjugada).
- Temos que

$$
\begin{aligned}
\pi(\theta|\boldsymbol{x}) &\propto f(\theta|\boldsymbol{x})\pi(\theta) \\
&= e^{-n\theta}\theta^{n\overline{x}}\theta^{a-1}e^{-b\theta} \\
&= \theta^{\sum_{i=1}^{n} x_i + a - 1}e^{-(n+b)\theta}\mathbb{1}_{(0,\infty)}(\theta)
\end{aligned}
$$

- Assim, $\theta|\boldsymbol{X} = \boldsymbol{x} \sim \text{gama}(\sum_{i=1}^{n} x_i + a, 1/(n+b))$. Logo,

$$
\mathcal{E}(\theta|\boldsymbol{X}) = \frac{n\overline{X}}{b+n} + \frac{a}{b+n}
$$

Exemplo

- Qual seria o estimador de Bayes, de $\tau(\theta) = \theta^2$, sob perda quadrática? Temos que

$$
\begin{aligned}
\mathcal{E}(\theta^2|\boldsymbol{X}) &= \mathcal{V}(\theta|\boldsymbol{X}) + E^2(\theta|\boldsymbol{X}) \\
&= \frac{\sum_{i=1}^{n} X_i + a}{(n+b)^2} + \left[\frac{n\overline{X}}{b+n} + \frac{a}{b+n}\right]^2 \\
&= \frac{\sum_{i=1}^{n} X_i + a}{(n+b)^2}\left(1 + \sum_{i=1}^{n} X_i + a\right)
\end{aligned}
$$

Exemplo

- Função de Risco frequentista. Temos que : $d^{\pi}(\boldsymbol{X}) = \frac{\sum_{i=1}^{n} X_i + a}{(b+n)}$ (sob perda quadrática). Assim

$$
\begin{aligned}
R(\theta, d^{\pi}) &= \mathcal{E}\left[\left(\frac{\sum_{i=1}^{n} X_i + a}{b+n} - \theta\right)^2\right] \\
&= \mathcal{V}\left(\frac{\sum_{i=1}^{n} X_i + a}{b+n}\right) + \left[\mathcal{E}\left(\frac{\sum_{i=1}^{n} X_i + a}{b+n}\right) - \theta\right]^2 \\
&= \frac{n\theta}{(n+b)^2} + \left(\frac{a - b\theta}{b+n}\right)^2 = \frac{n\theta + a^2 - 2ab + b^2 - \theta^2}{(n+b)^2} \\
&= \frac{1}{(n+b)^2}\left[a^2 + b^2\theta^2 + (n - 2ab)\theta\right]
\end{aligned}
$$

Exemplo

- Cont.:

$$\begin{aligned} R(\theta, d^{\pi}) &= \mathcal{EQM}_{\theta}(d^{\pi}) = \frac{1}{(n+b)^2/n}\left[\frac{a^2}{n} + \frac{b^2\theta^2}{n} + \frac{(1-2ab)\theta}{n}\right] \\ &= \frac{1}{\left(\frac{b}{\sqrt{n}} + \sqrt{n}\right)^2}\left[\frac{a^2}{n} + \frac{b^2\theta^2}{n} + \frac{(1-2ab)\theta}{n}\right] \end{aligned}$$

- Logo, $\lim_{n\to\infty} R(\theta, d^{\pi}) = 0$.
- Assim, $d^{\pi}(\boldsymbol{X})$ é um estimador consistente.

Exemplo

- Risco de Bayes: temos que:

$$
\begin{aligned}
r(\pi, d^{\pi}) &= \int_{\Theta} R(\theta, d^{\pi})\pi(\theta)d\theta = \mathcal{E}_{\theta}\left(R(\theta, d^{\pi})\right) \\
&= \frac{1}{(b+n)^2}\left(a^2 + b^2\mathcal{E}(\theta^2) + (n - 2ab)\mathcal{E}(\theta)\right) \\
&= \frac{1}{(b+n)^2}\left(a^2 + b^2\left(\frac{a}{b^2} + \frac{a^2}{b^2}\right) + (n - 2ab)\frac{a}{b}\right) \\
&= \frac{a}{(b+n)^2}\left(1 + \frac{n}{b}\right) = \frac{a}{b(b+n)}
\end{aligned}
$$

Teorema

- Seja $\boldsymbol{X}$ um vea com fdp $f_{\boldsymbol{X}}().|\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \Theta \subseteq \mathcal{R}^k$ e $\boldsymbol{T}$ uma estatística suficiente para $\boldsymbol{\theta}$. Se $\pi(.)$ é a priori para $\boldsymbol{\theta}$, o estimador de Bayes de $\boldsymbol{\theta}$ é função de $\boldsymbol{t}$, isto é:

$$d^{\pi} = G(\boldsymbol{T}(\boldsymbol{X}))$$

- Exercício: demonstre que a posteriori também depende de $\boldsymbol{T}$.
- Exemplo: Seja $X_1, .., X_n$ uma aa de $X \sim N(\theta, \sigma^2)$ e $\theta \sim N(\mu, \tau^2)$, com σ^2, μ, τ^2 conhecidos. Considerando perda quadrática, temos que:
$d^{\pi}(\boldsymbol{X}) = \mathcal{E}(\theta|\boldsymbol{X}) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \overline{X} + \mu$

Procedimento Minimax

- Dizemos que d^* é uma regra (estimador) minimax se

$$\sup_{\theta \in \Theta} R(\theta, d^*) = \inf_{d \in \mathcal{D}} \sup_{\theta \in \Theta} R(\theta, d) \text{ ou}$$

$$\sup_{\theta \in \Theta} R(\theta, d^*) \leq \sup_{\theta \in \Theta} R(\theta, d), \forall d \in \mathcal{D}$$

- Teorema: Suponhamos que d^π é um estimador de Bayes e

$$R(\theta, d^\pi) \leq r(\pi, d^\pi), \forall \theta \in \Theta.$$

  Então d^π é uma regra (estimador) minimax.

- Teorema: Seja π uma priori para θ e d^π o estimador de Bayes tal que $R(\theta, d^\pi)$ não depende de θ. Então d^π é um estimador minimax.

Procedimento Minimax

- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X \sim$ Bernoulli(θ), $\theta \sim$ beta(a, b) (a,b) conhecidos e considere a perda quadrática. Obtenha o estimador minimax.
- Temos que (estimador de Bayes):

$$d^{\pi} = \frac{n\overline{X} + a}{n + a + b}$$

Procedimento Minimax

- Neste caso, temos que

$$
\begin{aligned}
R(d^{\pi},\theta) &= \mathcal{E}_{\boldsymbol{X}|\theta}\left[\left(\frac{n\overline{X}+a}{n+a+b}-\theta\right)^2\right] \\
&= \mathcal{V}_{\boldsymbol{X}|\theta}\left(\frac{n\overline{X}+a}{n+a+b}\right)+\left[\mathcal{E}\left(\frac{n\overline{X}+a}{n+a+b}-\theta\right)\right]^2 \\
&= \frac{1}{(n+a+b)^2}\left[A\theta^2-B\theta+a^2\right]
\end{aligned}
$$

Procedimento Minimax

- em que $A = ((a+b)^2 - n)$ e $B = (2a(a+b) - n)$. Assim, d^{π} é minimax $\Leftrightarrow$ A = B = 0, ou seja, $\Leftrightarrow a = b = \frac{\sqrt{n}}{2}$.
- Portanto,

$$R(\theta, d^{\pi}) = \frac{1}{4(1+\sqrt{n})^2} \text{ e}$$

$$d^{\pi} = \frac{n\overline{X} + \sqrt{n}/2}{n + \sqrt{n}}$$

Conjuntos (intervalos) de credibilidade

- Def: um conjunto de credibilidade de nível $\gamma = 1 - \alpha$ é um subconjunto C de $\Theta, (C \subseteq \Theta)$, tal que:

$$P(\theta \in C|\boldsymbol{x}) \geq 1 - \alpha$$

em que

$$P(\theta \in C|\boldsymbol{x}) = \begin{cases} \sum_{\theta \in C} \pi(\theta|\boldsymbol{x}), & \text{se } \theta \text{ for discreto} \\ \int_C \pi(\theta|\boldsymbol{x}) d\theta, & \text{se } \theta \text{ for contínuo} \end{cases}$$

- Seja $\Theta \subseteq \mathcal{R}^k$, então um "boa" região de credibilidade é um subconjunto C de Θ que tem o menor "tamanho" (comprimento, área, volume). Por exemplo, se $\Theta \subseteq \mathcal{R}$ e $C = [L, S]$, então C é um "bom" intervalo de credibilidade se tiver menor comprimento possível.

Conjuntos (intervalos) de credibilidade

- Se $\Theta \subseteq \mathcal{R}$, então podemos ter as seguintes situações:
  - i) $C = [L, S]$.
  - ii) $C = [a, S]$.
  - iii) $C = [L, b]$.

  em que a e b são, respectivamente, o limite inferior e superior do espaço paramétrico.
- Usualmente, no caso acima, calculamos

$$P(t_1(\boldsymbol{x}) \leq \theta \leq t_2(\boldsymbol{x})|\boldsymbol{x}) \geq \gamma$$

- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X \sim$ Poisson(θ) e $\theta \sim$ gama$(a, 1/b)$. Então $\theta|\boldsymbol{x} \sim$ gama$(a + \sum_{i=1}^{n} x_i, 1/(n + b))$ (exercício).

Conjuntos (intervalos) de credibilidade

- Portanto, $2(n+b)\theta|\boldsymbol{x} \sim \chi^2_{2\left(a+\sum_{i=1}^n x_i\right)}$.
- Desse modo temos que:

$$P\left(q_1 \leq 2(n+b)b\theta \leq q_2\right) = 1-\alpha = \gamma$$
$$P\left(\frac{q_1}{2(n+b)} \leq \theta \leq \frac{q_2}{2(n+b)}\right) = 1-\alpha = \gamma$$

- Portanto, $IC_B[\theta,\gamma] = \left[\frac{q_1}{2(n+b)}, \frac{q_2}{2(n+b)}\right]$, em que $P(Y < q_1) = \frac{\alpha}{2}$, $P(Y > q_2) = \frac{\alpha}{2}$, $Y \sim \chi^2_{2\left(a+\sum_{i=1}^n x_i\right)}$.

Testes de hipótese

- Considere as hipóteses $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1 = \Theta_0^c$. Então, do ponto de vista clássico (frequentista), a mensuração da qualidade é feita em termos dos erros do tipo I e II e a decisão é tomada de acordo com as priporiedades da região crítica (embora também se possa avaliar a qualidade dos testes bayesianos através de critérios frequentistas).
- Do ponto de vista Bayesiano, temos que decidir entre H_0 e H_1 considerando as informações a priori de $\theta(\pi(.))$ e a distribuição a posteriori $\theta(\pi(.|\boldsymbol{x}))$.
- Dada a distribuição a posteriori, podemos calcular, por exemplo:

$$P(H_0|\boldsymbol{x}) = P(\theta \in \Theta_0|\boldsymbol{x}) = P(H_0 \text{ é verdadeira}|\boldsymbol{x})$$
$$P(H_1|\boldsymbol{x}) = P(\theta \in \Theta_1|\boldsymbol{x}) = P(H_1 \text{ é verdadeira}|\boldsymbol{x})$$

Testes de hipótese

- Podemos, assim, rejeitar H_0 se $\frac{P(H_1|\boldsymbol{x})}{P(H_0|\boldsymbol{x})} \geq c$, $c \in (0, \infty)$, em que c deve ser escolhido de forma conveniente.
- Em termos dos conceitos de testes de hipóteses frequentistas a RC é da forma

$$RC = \{\boldsymbol{x} \in \mathcal{X} : P(\theta \in \Theta_0^c|\boldsymbol{x}) \geq c\}$$

- Consequentemente, $RC^c = \{\boldsymbol{x} \in \mathcal{X} : P(\theta \in \Theta_0^c|\boldsymbol{x}) < c\}$
- Essencialmente, a abordagem acima é útil para testar hipóteses do tipo $\theta \leq (<)\theta_0$, $\theta \geq (>)\theta_0$, $\theta \in (\theta_0, \theta_1)$, $\theta \leq (<)\theta_0$ ou $\theta \geq (>)\theta_1$.

Testes de hipótese

- Para outras formas de testar hipóteses (do ponto de vista bayesiano) e para hipóteses precisas ($\theta = \theta_0$) veja aqui.
- Exemplo: Seja $X_1, ..., X_n$ uma aa de $X \sim N(\theta, \sigma^2)$ e $\theta \sim N(\mu, \tau^2)$, com μ, τ^2, σ^2 conhecidos. Considere as hipóteses: $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.
- Neste caso, temos que: $\theta | \boldsymbol{X} = \boldsymbol{x} \sim N((1-\eta)\overline{x} + \eta\mu, \eta\tau^2)$, $\eta = \frac{\sigma^2}{\eta\tau^2 + \tau^2}$.

Testes de hipótese

- Por outro lado, não se rejeita H_0 se

$$
\begin{aligned}
P(\theta \in \Theta_0|\boldsymbol{x}) \geq c &= P(\theta \leq \theta_0|\boldsymbol{x}) \geq c \\
&= \Phi\left(\frac{\theta_0 - (1-\eta)\overline{x} - \eta\mu}{\sqrt{n}\tau}\right) \geq c \\
&= \frac{\theta_0 - (1-\eta)\overline{x} - \eta\mu}{\sqrt{n}\tau} \geq \Phi^{-1}(c) \\
&= -\overline{x} \geq \frac{-\theta_0 + \eta\mu}{1-\eta} + \Phi^{-1}(c) \\
&\rightarrow \overline{x} \leq \frac{\theta_0 - \eta\mu}{1-\eta} + \Phi^{-1}(c)
\end{aligned}
$$

- Assim, a RC é dado por:

$$
RC = \left\{\boldsymbol{x} \in \mathcal{R}^n : \overline{x} > \frac{\theta_0 - \eta\mu}{1-\eta} + \Phi^{-1}(c)\right\}
$$