AGA 0505 - Análise de Dados em Astronomia

8. Comparação de Modelos

Laerte Sodré Jr.

1o. semestre, 2024

aula de hoje:

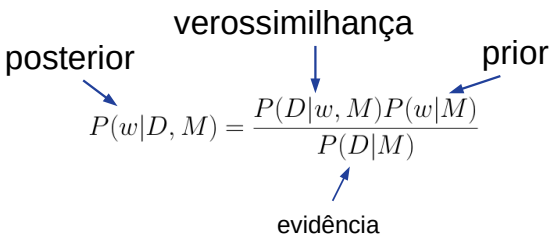
1. comparação bayesiana de modelos
2. a estória de Mr. A e Mr. B
3. as novas teorias de Mr. A e Mr. B
4. as moedas de Mr. A e Mr. B
5. métodos aproximados: BIC e AIC
6. comparação de métodos frequentistas e bayesianos

We balance probabilities and choose the most likely. It is more than possible; it is probable.

Sherlock Holmes em Silver Blaze, Arthur Conan Doyle

relembrando:

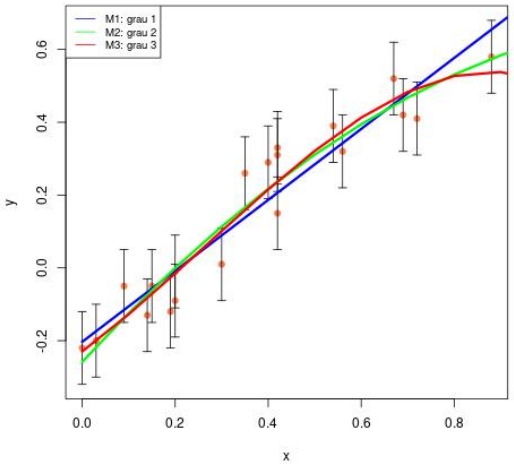
inferência de parâmetros w de um modelo M com dados D



comparação bayesiana de modelos

- suponha que temos um conjunto de dados D que queremos modelar ou com o modelo A, que tem parâmetros w_A, ou com o modelo B, que tem parâmetros w_B
  $>>>$ que modelo preferir?

- ex.: qual é o melhor modelo para os dados da figura: uma reta, uma parábola, um polinômio de terceiro grau?



comparação bayesiana de modelos

- suponha que temos um conjunto de dados D que queremos modelar ou com o modelo A, que tem parâmetros w_A, ou com o modelo B, que tem parâmetros w_B
  $>>>$ que modelo preferir?
- inferência dos parâmetros de um modelo M

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

evidência:

$$P(D|M) = \int P(D|w,M)P(w|M)dw$$

- probabilidade de um modelo M

$$P(M|D) = P(D|M)P(M)/P(D)$$

note que $P(D|M)$ é a evidência na inferência de parâmetros!

- comparação dos modelos A e B:

$$P(A|D) = P(D|A)P(A)/P(D)$$

$$P(B|D) = P(D|B)P(B)/P(D)$$

comparação bayesiana de modelos

- comparação dos modelos A e B:

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

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

logo

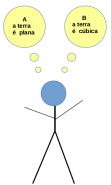
$$\frac{P(A|D)}{P(B|D)} = \frac{P(D|A)}{P(D|B)}\frac{P(A)}{P(B)} =$$

$$= \frac{\text{evidencia de A}}{\text{evidencia de B}} \times \frac{\text{prior de A}}{\text{prior de B}}$$

- se não temos preferência por A ou por B, $P(A) = P(B)$, e a razão entre os posteriores dos modelos fica igual à razão entre suas evidências:

$$\frac{P(A|D)}{P(B|D)} = \frac{P(D|A)}{P(D|B)} = \frac{\text{evidencia de A}}{\text{evidencia de B}}$$

- mas se temos, devemos usar a informação!



comparação bayesiana de modelos

- comparação dos modelos A e B:

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

- *odds*: razão de probabilidades (chances)
  - odds dos posteriores

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

- odds das evidências: fator de Bayes

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

- odds dos priores:

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

- comparação de modelos:
  odds dos posteriores =
  fator de Bayes x odds dos priores

comparação bayesiana de modelos

- comparação de modelos: *odds dos posteriores = fator de Bayes x odds dos priores*
- escala de Jeffreys (modificada por Trotta) para avaliar a diferença entre modelos:

$\lvert \ln B \rvert$	odds relativos B	probabilidade do modelo favorecido*	confiança na diferença
< 1.0	$< 3:1$	< 0.750	muito pouca
1 a 2.5	3 a $12:1$	0.923	fraca
2.5 a 5.0	12 a $150:1$	0.993	moderada
> 5.0	$> 150:1$	> 0.993	forte

* supondo $P(M_1) = P(M_2) = 1/2$ e $P(M_1|D) + P(M_2|D) = 1$

a estória de Mr. A e Mr. B

- a estória de Mr. A e de Mr. B (Gull 1989; Sivia & Skilling sec. 4.1)
- Mr. A tem uma teoria, Mr. B também tem uma teoria, mas com um parâmetro (w) a mais
- que teoria devemos preferir com base nos dados D?
- note que a teoria de Mr. B deve ajustar melhor os dados (por ter um parâmetro a mais), mas também é mais complexa (pela mesma razão)
- como decidir?
- Teorema de Bayes:

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

note que w não aparece nessa expressão!

- podemos incluir w na análise via marginalização:

$$P(D|B) = \int P(D, w|B) đw = \int P(D|w, B) P(w|B) đw$$

a estória de Mr. A e Mr. B

- evidência do modelo de Mr. B:

$$P(D|B) = \int P(D, w|B)dw = \int P(D|w, B)P(w|B)dw$$

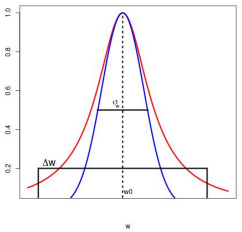
- Mr. B adota um prior uniforme para w:

$$P(w|B) = \begin{cases} \frac{1}{w_{max} - w_{min}} = \frac{1}{\Delta w} & \text{se } w_{min} < w < w_{max} \\ 0 & \text{nos outros casos} \end{cases}$$

- vamos supor que a verossimilhança tem um valor w_0 que provê o melhor ajuste às medidas (máxima verossimilhança)

- *aproximação de Laplace*: aproximamos a verossimilhança por uma gaussiana de média w_0 e largura σ_w:

$$P(D|w, B) \simeq P(D|w_0, B) \exp\left[-\frac{(w - w_0)^2}{2\sigma_w^2}\right]$$



a estória de Mr. A e Mr. B

- evidência do modelo de Mr. B:

$$P(D|B) = \int P(D, w|B)dw = \int P(D|w, B)P(w|B)dw$$

- prior de Mr. B :

$$P(w|B) = \frac{1}{\Delta w}$$

- verossimilhança:

$$P(D|w, B) \simeq P(D|w_0, B) \exp\left[-\frac{(w - w_0)^2}{2\sigma_w^2}\right]$$

- então, supondo $\Delta w >> \sigma_w$,

$$P(D|B) \approx \frac{P(D|w_0, B)\sigma_w\sqrt{2\pi}}{\Delta w}$$

e a razão dos posteriores então fica:

$$\frac{P(A|D)}{P(B|D)} = \frac{P(A)}{P(B)} \times \frac{P(D|A)}{P(D|w_0, B)} \times \frac{\Delta w}{\sigma_w\sqrt{2\pi}}$$

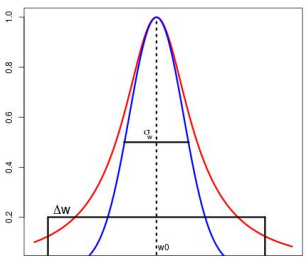
= odds dos priores x fator de bayes x fator de Occam

a estória de Mr. A e Mr. B

- razão dos posteriores de A e B:

$$\frac{P(A|D)}{P(B|D)} = \frac{P(A)}{P(B)} \times \frac{P(D|A)}{P(D|w_0, B)} \times \frac{\Delta w}{\sigma_w \sqrt{2\pi}}$$

= odds dos priores x fator de bayes x fator de Occam



- o primeiro termo contém nossas preferências a priori relativas sobre as teorias (para não se ter viés, pode ser o caso de supor que ele vale 1)
- o segundo termo é a razão entre as melhores previsões de cada modelo para os dados disponíveis; é de se esperar que a verossimilhança de B seja maior que a de A, devido ao parâmetro adicional
- o terceiro termo é o *fator de Occam*: penaliza o modelo B por seu parâmetro adicional

a navalha de Occam

- a navalha de Occam (sec.XIV):

  princípio da simplicidade

  Frustra fit per plura quod potest fieri per pauciora

  tradução livre:
  é bobagem fazer com mais o que se pode fazer com menos

- o fator de Occam penaliza a complexidade

as novas teorias de Mr. A e Mr. B

- Mr. A tem uma teoria: o parâmetro w é fixo e igual a zero
- Mr. B tem outra teoria: o parâmetro w obedece a uma distribuição gaussiana de média 0 e desvio padrão Σ
- fazemos uma medida e concluímos que w é descrito por uma gaussiana de média μ e variância σ
- qual teoria devemos preferir?

- modelo probabilístico:
  - prior de w no modelo A:

$$P(w|A) = \delta(0)$$

  - prior de w no modelo B:

$$P(w|B) = \frac{1}{\sqrt{2\pi}\Sigma} \exp\left[-\frac{w^2}{2\Sigma^2}\right]$$

  - verossimilhança dos dados:

$$P(\mu, \sigma|w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(w-\mu)^2}{2\sigma^2}\right]$$

as teorias de Mr. A e Mr. B

- evidência de A:

$$P(D|A) = \int P(\mu, \sigma|w)P(w|A)dw =$$

$$= P(\mu, \sigma|w = 0) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\mu^2}{2\sigma^2} \right]$$

- evidência de B:

$$P(D|B) = \int P(\mu, \sigma|w)P(w|B)dw =$$

$$= \frac{1}{2\pi\sigma\Sigma} \int \exp\left[-\frac{(w-\mu)^2}{2\sigma^2} - \frac{w^2}{2\Sigma^2} \right] dw =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{1+\Sigma^2/\sigma^2}} \exp\left[-\frac{\lambda^2}{2}\left(1 - \frac{1}{1+\sigma^2/\Sigma^2}\right)\right]$$

onde

$\lambda = \mu/\sigma$

- fator de Bayes:

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

$$= \sqrt{1+\Sigma^2/\sigma^2} \exp\left[-\frac{\lambda^2}{2(1+\sigma^2/\Sigma^2)} \right]$$

B_{AB} depende de dois parâmetros: Σ/σ e λ

- $\lambda = \mu/\sigma$ mede o parâmetro adicional em termos de sigmas: λ grande favorece Mr. B
- Σ^2/σ^2 mede a importância do prior em relação aos dados- é um fator de Occam: quanto maior este termo, pior para Mr. B

as novas teorias de Mr. A e Mr. B

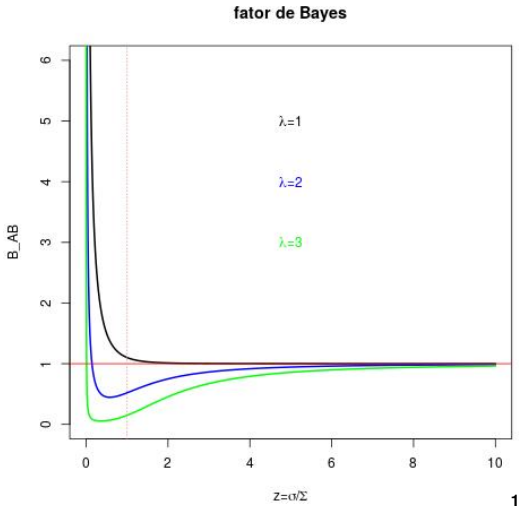
- fator de Bayes:

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

$$= \sqrt{1 + \Sigma^2/\sigma^2}\, \exp\left[- \frac{\lambda^2}{2(1 + \sigma^2/\Sigma^2)} \right]$$

- $\lambda = \mu/\sigma$ mede o parâmetro adicional em termos de sigmas: λ grande favorece B

- σ/Σ mede a importância dos dados em relação ao prior: quanto maior melhor para B



as moedas de Mr. A e Mr. B

- exemplo da wikipedia:

  https://en.wikipedia.org/wiki/Bayes_factor#Example

- jogamos uma moeda n =200 vezes e obtemos x =115 caras
- modelo A: probabilidade de sair cara: $w = 0.5$
- modelo B: probabilidade de sair cara: w é livre, com prior uniforme entre 0 e 1 ($P(w|B) = 1$)

- modelo A:

$$P(D|A) = \text{Binomial}(n, x, w = 0.5) = \binom{n}{x} w^x (1-w)^{n-x} = \binom{n}{x} 0.5^n$$

- modelo B:

$$P(D|B) = \int P(D, w|B) dw = \int P(D|w, B) P(w|B) dw =$$

$$\int_0^1 \binom{n}{x} w^x (1-w)^{n-x} dw = \binom{n}{x} \text{Beta}(x+1, n-x+1)$$

função beta:

$$\text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

as moedas de Mr. A e Mr. B

- modelo A:

$$P(D|A) = \binom{n}{x} 0.5^n$$

- modelo B:

$$P(D|B) = \binom{n}{x} \text{Beta}(x+1, n-x+1)$$

- jogamos uma moeda n =200 vezes e obtemos x =115 caras

- vamos supor que os dois modelos têm a mesma probabilidade a priori: $P(A) = P(B)$

- logo,

$$B_{AB} = \frac{P(A|D)}{P(B|D)} = \frac{\binom{n}{x} 0.5^n}{\binom{n}{x} \text{Beta}(x+1, n-x+1)} = \frac{0.00595}{0.00497} \approx 1.197$$

o modelo A é ligeiramente melhor, mas não dá para decidir...

- análise frequentista:
modelo A seria rejeitado com $\alpha = 0.05$

métodos aproximados: BIC

- a evidência é difícil de calcular- envolve integração no espaço de parâmetros:

$$P(D|M) = \int P(D|w, M)P(w|M)dw$$

e é comum usar-se aproximações:

- BIC: Bayesian Information Criterion
- AIC: Akaike Information Criterion

BIC: Bayesian Information Criterion:

- para um modelo com máxima verossimilhança $\hat{L}$, k parâmetros livres e n observações (dados):

$$BIC = \ln(n) \times k - 2 \ln \hat{L}$$

- o primeiro termo penaliza a *complexidade do modelo* (número de parâmetros) e o segundo a *qualidade do ajuste*: modelos com menor BIC devem ser preferidos

métodos aproximados: AIC

- **AIC: Akaike Information Criterion**
  - baseado em teoria da informação: se dois modelos A e B descrevem os mesmos dados, o AIC mede a diferença de informação entre eles
  - para um modelo com máxima verossimilhança $\hat{L}$ e k parâmetros livres:

$$AIC = 2k - 2\ln\hat{L}$$

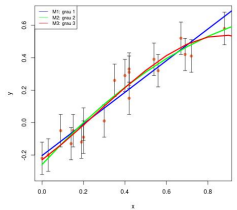
  - modelos com menor AIC devem ser preferidos

  - compare AIC com BIC:

$$BIC = \ln(n) \times k - 2\ln\hat{L}$$

  para $n \geq 8, \quad k\ln n > 2k$

- seleção de modelos polinomiais



grau	log(E)	BIC	AIC	CV
1	7.45	-38.03	-40.02	0.137
2	7.09	-36.90	-39.88	0.120
3	7.09	-34.37	-38.35	0.254
sel	M1	M1	M1	M2

comparação de modelos: as moedas de Mr. A e Mr. B

- critérios:

$$AIC = 2k - 2\ln\hat{L}$$

$$BIC = \ln(N)k - 2\ln\hat{L}$$

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

- **as moedas de Mr. A e Mr. B:**
  - $B_{AB} = 1.197 \longrightarrow$ A
    (A ligeiramente melhor que B, mas não dá para decidir)
  - $AIC_{AB} = 53.905 \longrightarrow$ B
  - $BIC_{AB} = 50.606 \longrightarrow$ B
  - p-value = 0.014 (H0: $w = 0.5$)
    - nível de confiança 0.05 $\longrightarrow$ B
    - nível de confiança 0.01 $\longrightarrow$ A

Referências

- Data Analysis: a Bayesian Tutorial, Sivia & Skilling, 2006
- Bayesian Methods in Cosmology, R. Trotta, arXiv:1701.01467 (2017)
- https://www.r-bloggers.com/aic-bic-vs-crossvalidation/
- http://jakevdp.github.io/blog/2015/08/07/frequentism-and-bayesianism-5-model-selection/