UNIVERSIDADE ABERTA
www.uab.pt

Ministério da Ciência, Tecnologia e Ensino Superior

p-Fólio

U.C. 21073
Introdução à probabilidade e estatística Bayesianas

1 de Fevereiro de 2016

- INSTRUÇÕES -

- O p-fólio é composto por 5 questões (de onde serão escolhidas 3), contém 5 páginas e termina com a palavra **FIM**. Contém ainda uma página de formulário, em anexo.
- Verifique o seu exemplar e, caso encontre alguma anomalia, dirija-se ao professor vigilante nos primeiros 15 minutos da mesma, pois qualquer reclamação sobre defeito(s) de formatação e/ou de impressão que dificultem a leitura não será aceite depois deste período.
- O estudante deverá responder à prova na folha de ponto e preencher o cabeçalho e todos os espaços reservados à sua identificação, com letra legível. Não serão classificadas respostas apresentada em folhas de rascunho.Após a prova, o enunciado pode ficar na posse do estudante.
- Verifique no momento da entrega da(s) folha(s) de ponto se todas as páginas estão rubricadas pelo vigilante. Caso necessite de mais do que uma folha de ponto, deverá numerá-las no canto superior direito.
- Utilize uma letra legível e não use uma caneta de outra cor que não seja o preto ou o azul - as respostas a lápis não serão consideradas.
- É permitido o uso de máquina de calcular.
- Tenha em atenção que o exame tem a duração máxima de **90 minutos**.

Critérios de avaliação e cotação:

- É importante ter presente que só serão cotadas as questões devidamente justificadas. Se apresentar apenas um valor numérico como resposta sem qualquer justificação, mesmo que esteja correcto, terá cotação zero. Acautele a justeza dos conteúdos da resposta ao enunciado da questão.
- A apresentação de definições e resultados teóricos, desde que sejam relevantes para a resposta, serão alvo de classificação.
- As questões terão as cotações seguintes: cada questão vale 4 valores.

Por favor preencha os seus dados

Nome:

Nº de Estudante

B.I.:

Justifique todas as afirmações e apresente os cálculos realizados para as obter

Grupo I

Responda a uma (e apenas uma) das seguintes duas questões, à sua escolha.

Problema 1. *Diga, justificando, se o seguinte conjunto de atribuições de probabilidades é internamente consistente:*

$$\begin{aligned} p(A \Rightarrow B|C) &= 0.6 \\ p(A|C) &= 0.5 \\ p(AB|C) &= 0.2 \end{aligned}$$

Solução: Temos

$$\begin{aligned} p(A \Rightarrow B|C) &= p(\overline{A} + B|C) \\ &= p(\overline{A}|C) + p(B|C) - p(\overline{A}B|C) \\ &= p(\overline{A}|C) + p(B|C) - p(\overline{A}|BC)p(B|C) \\ &= p(\overline{A}|C) + p(B|C)\left(1 - p(\overline{A}|BC)\right) \\ &= p(\overline{A}|C) + p(B|C)p(A|BC) \\ &= p(\overline{A}|C) + p(AB|C) \\ &= 1 - p(A|C) + (AB|C) = 1 - 0.5 + 0.2 = 0.7 \neq 0.6. \end{aligned}$$

pelo que a primeira condição é incompatível com as outras duas.

Problema 2. *Um concurso de televisão consiste no seguinte: São apresentadas três portas fechadas ao concorrente, que tem que escolher uma. Por detrás de uma, e apenas uma, encontra-se o prémio. Suponhamos que o concorrente escolheu a porta número 3. Após essa escolha, e antes da porta ser aberta, o apresentador (que sabe onde está o prémio), abre a porta número 1 e mostra ao concorrente que o prémio não estava lá. Restam portanto apenas a porta 3, que o concorrente escolheu, e a porta 2. O apresentador diz ao concorrente que, se quiser, ele ainda pode mudar a sua escolha da porta 3 para a porta 2. Pergunta: Será que vale a pena mudar de escolha? Quais são as probabilidades so prémio estar atrás de cada uma das portas, antes e depois do apresentador ter aberto a porta 1?*

Solução:

Consideremos as proposições seguintes:
$X_i =$"o prémio está atrás da porta i"(para $i \in \{1,2,3\}$)
$E_i =$"o concorrente escolheu a porta i"(para $i \in \{1,2,3\}$)
$Y_i =$"o apresentador abriu a porta i"(para $i \in \{1,2,3\}$)

Sabemos do enunciado que os X_i são mutuamente exclusivos e a sua disjunção é verdadeira (o prémio tem que estar atrás de uma e uma só porta). Então $p(X_1)+p(X_2)+p(X_3) = 1$, e $p(X_iX_j) = \delta_{ij}$. É-nos também dito que $p(E_3) = 1$ e que $p(Y_i) = 1$ (sendo os outros $p(E_j) = 0$).

Então

$$p(X_3|E_3Y_1) = \frac{p(Y_1|X_3E_3)p(X_3|E_3)}{p(Y_1|E_3)} = \frac{(1/2)\cdot(1/3)}{p(Y_1|E_3)}$$

Porque é que $p(Y_1|X_3E_3) = 1/2$? Porque se o prémio está na porta 3, o apresentador pode escolher livremente entre abrir a porta 1 ou a porta 2, pois ambas estão vazias. Como não temos informação em prol de uma ou outra, atribuimos probabilidades iguais a essa escolha.

$$p(X_2|E_3Y_1) = \frac{p(Y_1|X_2E_3)p(X_2|E_3)}{p(Y_1|E_3)} = \frac{1\cdot(1/3)}{p(Y_1|E_3)}$$

Porque é que $p(Y_1|X_2E_3) = 1$? Porque se o concorrente escolheu a porta 3 e o prémio está na porta 2 então o apresentador não tem outra escolha que não abrir a porta 1, senão estragava o jogo! (note que o enunciado diz que o apresentador sabe onde está o prémio)

Mas então $p(X_2|E_3Y_1) = 2p(X_3|E_3Y_1) > p(X_3|E_3Y_1)$. Portanto vale a pena mudar de porta!

Na verdade, podemos dizer mais do que isso. Como $p(X_1|E_3Y_1) = 0$ (foi essa a porta que o apresentador mostrou estar vazia) e como $\sum_i p(X_i|E_3Y_1) = 1$, temos que $p(X_3|E_3Y_1) = 1/3$ e $p(X_2|E_3Y_1) = 2/3$.

Grupo II

Responda a duas (e apenas duas) das seguintes três questões, à sua escolha.

Problema 3. *Considere o seguinte modelo de Pareto, em que $x_m > 0$ é uma constante real conhecida e $\lambda > 0$ é um parâmetro real desconhecido:*

$$p(x|\lambda, x_m) = \begin{cases} \dfrac{\lambda x_m^\lambda}{x^{\lambda+1}}, \ \textit{para } x \geq x_m \\ 0, \ \textit{para } x < x_m \end{cases}$$

Mostre que a família Gama(α, β) é família conjugada de priors para o modelo de Pareto com x_m conhecido, e calcule a regra de actualização dos parâmetros α, β.

Nota: Recorde que

$$Gama(\lambda|\alpha, \beta) \propto \lambda^{\alpha-1} e^{-\beta\lambda}.$$

Solução:

Seja $\underline{x}$ o vector dos dados observados.

$$\begin{aligned}
P(\lambda|\underline{x}) \propto p(\underline{x}|\lambda)p(\lambda|\alpha,\beta) &= \frac{\lambda^n x_m^{n\lambda}}{\prod_i x_i^{\lambda+1}} \cdot \lambda^{\alpha-1} e^{-\beta\lambda} \\
&= \frac{\lambda^{n+\alpha-1}}{\prod_i x_i^{\lambda+1}} x_m^{n\lambda} e^{-\beta\lambda} \\
&= \frac{\lambda^{n+\alpha-1}}{e^{\log(\prod_i x_i^{\lambda+1})}} e^{log(x_m^{n\lambda})} e^{-\beta\lambda} \\
&= \frac{\lambda^{n+\alpha-1}}{e^{(\lambda+1)\sum_i \log(x_i)}} e^{n\lambda\log(x_m)} e^{-\beta\lambda} \\
&= \lambda^{n+\alpha-1} exp\left((\lambda+1)\sum_i \log(x_i) + n\lambda\log(x_m) - \beta\lambda\right) \\
&= \lambda^{n+\alpha-1} exp\left(-\lambda\left(\sum_i \log(x_i) - n\log(x_m) + \beta\right)\right) \exp\left(\sum_i \log(x_i)\right) \\
&\propto \lambda^{n+\alpha-1} exp\left(-\lambda\left(\sum_i \log(x_i) - n\log(x_m) + \beta\right)\right) \\
&\propto \lambda^{n+\alpha-1} exp\left(-\lambda\left(\sum_i \log(\frac{x_i}{x_m}) + \beta\right)\right)
\end{aligned}$$

Sendo que esta expressão final é de novo o núcleo de uma Beta, demonstrámos o que se pretendia, e verificámos que a regra de actualização dos parâmetros é

$$(\alpha, \beta) \mapsto \left(\alpha + n, \beta + \sum_i \log\left(\frac{x_i}{x_m}\right)\right)$$

Problema 4. *O número de novos autores que em cada ano fiscal conseguem fazer um contrato com uma certa editora é uma variável aleatória que segue uma distribuição Poisson de média θ cuja distribuição a priori é uma Gamma$(\alpha = 15, \beta = 5)$. Sabendo-se que nos últimos 10 anos a editora contratou um total de 44 novos autores determine:*

a) A distribuição a posteriori de θ.

b) A estimação de máxima verosimilhança de θ.

c) O valor de θ para o qual é atingido o máximo da distribuição à posteriori, e o valor médio a posteriori de θ.

Solução:

A distribuição a priori $p(\theta)$ é uma Gama de hiperparâmetros $(\alpha, \beta) = (15, 5)$. A família Gama é conjugada natural do modelo amostral de Poisson, pelo que a distribuição posterior é ainda uma Gama, obtida pela regra de update

$$(\alpha, \beta) \mapsto (\alpha + \sum_{i=1}^{n} x_i, \beta + n),$$

sendo que neste caso

$$(15, 5) \mapsto (15 + 44, 5 + 10) = (59, 15).$$

b) Temos que

$$\begin{aligned} \frac{\partial}{\partial\theta} log L(\theta|\underline{x}) &= \frac{\partial}{\partial\theta} log \left(\frac{\theta^{\sum x_i} e^{-n\theta}}{\prod x_i!} \right) = \frac{\partial}{\partial\theta} \left(\Sigma\underline{x} log\theta - n\theta - log(\underline{x}!) \right) \\ &= \frac{\Sigma\underline{x}}{\theta} - n \end{aligned}$$

e portanto a estimativa de máxima verosimilhança, resultante de igualar a zero esta expressão, é

$$\hat{\theta} = \frac{\Sigma\underline{x}}{n} = 44/10 = 4,4$$

Podemos comparar este valor com o que se obtém calculando o máximo da distribuição a posteriori. Obtemos

$$\frac{\partial}{\partial\theta} log Gama(59, 15)(\theta) = \frac{\partial}{\partial\theta}(26 log\theta - 7\theta) = 0 \implies \theta = 26/7 = 3.71...$$

O valor médio de uma gama é $\alpha/\beta = 59/15$.

Problema 5. *Um sensor óptico mede a distância θ entre dois satélites, com um erro que é Gaussiano, dado por uma distribuição normal de média nula e desvio padrão igual a $\sigma = 3 \times 10^{-3}$ metros. Dito de outra forma, para um objecto a uma distância real de θ metros, a leitura no sensor será de x metros com uma probabilidade dada por $p(x|\theta) = N(x|\mu = \theta, \sigma^2 = 9 \times 10^{-6})$. Suponhamos já antes tínhamos tentado determinar a distância ao mesmo alvo utilizando um aparelho com uma margem de erro maior, também gaussiana mas com desvio padrão $b = 10^{-2}$ metros. Esse aparelho inicial tinha dado uma leitura de $a = 105.11 metros$. Tomando para prior de θ a distribuição $N(\theta|a, b^2)$, e sabendo que o novo aparelho dá uma medição de 105.099 metros, determine o posterior de θ, e uma região de credibilidade a 95% para θ.*

Solução: O posterior de θ de um modelo normal $N(x|\theta, \sigma^2)$ com σ conhecido e com prior normal $p(\theta) = N(a, b^2)$, após observação de um valor de x, é ainda uma normal $N(a', b'^2)$ de parâmetros

$$a' = \frac{\frac{1}{b^2}a + \frac{1}{\sigma^2}x}{\frac{1}{b^2} + \frac{1}{\sigma^2}}, \quad b'^2 = \frac{1}{\frac{1}{b^2} + \frac{n}{\sigma^2}}$$

substituimos $x = 105.099$, $a = 105.11$, $b^2 = 10^{-4}$, $\sigma^2 = 9 \times 10^{-6}$, e obtemos

$$a' = 105.1, b'^2 = 8.26 \times 10^{-6}$$

pelo que a distribuição posterior é $p(\theta|x) = N(a' = 105.1, b'^2 = 8.26 \times 10^{-6})$.

Como a distribuição é normal, basta ler numa tabela o intervalo de credibilidade.

Como $b' = \sqrt{b'^2} = 8.26 \times 10^{-6} \approx 2,87 \times 10^{-3}$, uma região de credibilidade posterior para θ a 95% será portanto o intervalo $105.1 \pm 2b' = [105.094, 105.106]$ metros.

FIM

Formulário

Distribuição Exponencial:

$$f(x|\beta) = \frac{1}{\beta}e^{-x/\beta},\ E(f) = \beta, V(f) = \beta^2$$

Distribuição de Poisson

$$f(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad E(f) = V(f) = \lambda$$

Função Gama:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, \Gamma(z \in \mathbb{N}) = (z-1)!$$

Distribuição Gama:

$$Gama(\lambda|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \lambda > 0, \quad , E(Gama) = \frac{\alpha}{\beta}, V(Gama) = \frac{\alpha}{\beta^2}$$

Função Beta:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt, Re(x), R(y) > 0$$

Distribuição Beta:

$$Beta(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

$$E(Beta(\alpha,\beta)) = \tfrac{\alpha}{\alpha+\beta},\ \ Var(Beta(\alpha,\beta)) = \tfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$