21073 - Introdução às probabilidades e estatística bayesianas

Ano lectivo 2013/14

Docente: António Araújo

Acção Formativa 1

1.

Considere os seguintes axiomas:

A0: Se A é logicamente equivalente a B então $p(A|C) = p(B|C)$ para qualquer C, e se o valor lógico de $(A|B)$ é 1 ou 0, então a probabilidade $p(A|B)$ é igual ao valor lógico de $(A|B)$. (regra de compatibilidade com a lógica)

A1: $p(\overline{A}|X) + p(A|X) = 1$ (regra da soma)

A2: $p(AB|C) = p(A|BC)p(B|C)$ (regra do produto)

Considere a expressão

A3: $p(A + B|C) = p(A|C) + p(B|C) - p(AB|C)$

a) Demonstre $A3$ a partir de $A0, A1, A2$.

b) Demonstre $A1$ a partir de $A0, A2, A3$.

Resolução do exercício:

a)

$$\begin{aligned} p(A + B|C) &= 1 - p(\overline{A + B}|C) = 1 - p(\overline{A} \cdot \overline{B}|C) \\ &= 1 - p(\overline{A}|C)p(\overline{B}|\overline{A}C) \\ &= 1 - p(\overline{A}|C)(1 - p(B|\overline{A}C)) = p(A|C) + p(\overline{A}B|C) \\ &= p(A|C) + p(B|C)p(\overline{A}|BC) = p(A|C) + p(B|C)(1 - p(A|BC)) \\ &= p(A|C) + p(B|C) - p(A|BC)p(B|C) = p(A|C) + p(B|C) - p(AB|C) \end{aligned}$$

Nota:No segundo passo usámos a equivalência lógica

$$\overline{A + B} = \overline{A} \cdot \overline{B}$$

b)

$$P(A + \overline{A}|B) = P(A|B) + P(\overline{A}|B) - P(A\overline{A}|B)$$

Mas pela regra de consistência com a lógica, como $A + \overline{A} = 1$ (para qualquer condição) então $P(A + \overline{A}|B) = 1$, e como $A\overline{A} = 0$ (para qualquer condição), então $P(A\overline{A}|B) = 0$. Então fica

$$P(\overline{A}|B) = 1 - P(A|B)$$

2. Obtenha uma expressão para $p(A \Rightarrow B|C)$ em termos de $p(\overline{A}|C)$ e $p(AB|C)$.

Solução:

$$\begin{aligned} p(A \Rightarrow B|C) &= p(\overline{A} + AB|C) = p(\overline{A}|C) + p(AB|C) - p(\overline{A}AB|C) \\ &= p(\overline{A}|C) + p(AB|C) \end{aligned}$$

3. Suponha que A e B são proposições equivalentes sabendo C; isto é, que $p(A \Leftrightarrow B|C) = 1$. Suponha que $p(AB|C) = 0.2$. Calcule $p(A + B|C)$. (dica: recorde qual é a definição da função lógica "$\Leftrightarrow$"em termos das funções lógicas "$+$"e "$\cdot$")

Solução: Notamos que $A \Leftrightarrow B \equiv AB + \overline{A}\ \overline{B}$. Então

$$\begin{aligned} p(A \Leftrightarrow B|C) &= p(AB + \overline{A}\ \overline{B}|C) = p(AB|C) + p(\overline{A}\ \overline{B}|C) - p(AB\overline{A}\ \overline{B}|C) \\ &= p(AB|C) + p(\overline{A}\ \overline{B}|C) = p(AB|C) + p(\overline{A+B}|C) \\ &= 1 + p(AB|C) - p(A+B|C) \end{aligned}$$

Então, como por hipótese, $p(A \Leftrightarrow B|C) = 1$, temos

$$p(A + B|C) = p(AB|C) = 0.2.$$

4.

Considere a seguinte proposta de silogismo fraco:

$$p(A|A \Rightarrow B, B) \geq p(A|A \Rightarrow B)$$

Descreva o seu significado "por palavras". Será verdadeiro? Demonstre que sim ou que não.

Solução:

O silogismo fraco em causa é

A implica B
B é verdade

Então, A é mais provável

Ou, mais por extenso: "Se sabemos que 'A implica B' é verdade e se sabemos que B é verdade então A torna-se mais provável do que se apenas sabemos que 'A implica B' é verdade".

Vamos mostrar que este silogismo é verdadeiro:

$$p(A|B, A \Rightarrow B) = \frac{p(B|A, A \Rightarrow B)p(A|A \Rightarrow B)}{p(B|A \Rightarrow B)}$$
$$= \frac{1 \times p(A|A \Rightarrow B)}{p(B|A \Rightarrow B)} \geq p(A|A \Rightarrow B)$$

Nota: o segundo passo utiliza o modus ponens, isto é, se sabemos que A e que A implica B, sabemos que B, e portanto pela regra de consistência com a lógica, sabemos que a probabilidade de B é 1.

5. Suponha que existe uma doença rara - a colite calamitosa do cólon coriáceo - que afecta uma proporção muito pequena da população (digamos uma em cada dez mil pessoas). Uma empresa farmacêutica, a YDOPM (Your Desiase is Our Profit Margin inc), desenvolveu um teste caseiro para a referida doença. Nos testes realizados sobre portadores da doença a empresa verificou que o rácio de falsos negativos era de 2 por cento e nos testes realizados sobre não-portadores verificou que o rácio de falsos positivos era de 1 por cento. Suponhamos que o Hipólito Hipocondríaco resolveu fazer o teste, apesar de não ter motivos especiais para crer que possua a doença. Suponhamos que o teste deu positivo.
Qual é a probabilidade do Hipólito ter de facto a doença? Discuta o resultado e refira o que isto implica para métodos de diagnóstico de doenças raras.
Pista: Suponha que a doença se chamava "sindroma irritativo do colo do útero"e que o resto do enunciado se mantinha constante. Qual era a probabilidade nesse caso?

Solução:

Vamos considerar as seguintes proposições
Pos ="Teste resultou positivo"
$Neg = \overline{Pos}$ ="Teste resultou negativo"
$Doente$ ="O sujeito em causa possui a doença"
$\overline{Doente}$ ="O sujeito em causa não possui a doença"

Sabemos do enunciado que a probabilidade de falsos negativos é $p(Neg|Doente) = 0.02$. Consideremos um indivíduo que tem a doença. O teste só pode dar positivo ou negativo, e, se der negativo, será necessariamente falso negativo. Então a probabilidade de o teste detectar a doença em alguém que de facto a possui é $p(Pos|Doente) = 1 - p(Neg|Doente) = 1 - 0.02 = 0.98$. Sabemos ainda que a probabilidade de falsos positivos é $p(Pos|\overline{Doente}) = 0.01$. O teste parece à primeira vista bastante eficaz. Mas qual é a probabilidade do Hipólito ter de facto a doença? Não é nehnum dos valores referidos. É $p(Doente|Pos)$, e para determinar esse valor temos que recorrer ao teorema de Bayes.

$$\begin{aligned} p(Doente|Pos) &= \frac{p(Pos|Doente)p(Doente)}{p(Pos)} \\ &= \frac{p(Pos|Doente)p(Doente)}{p(Pos \text{ e } Doente) + p(Pos \text{ e } \overline{Doente})} \\ &= \frac{p(Pos|Doente)p(Doente)}{p(Pos|Doente)p(Doente) + p(Pos|\overline{Doente})p(\overline{Doente})} \end{aligned}$$

O que é $p(Doente)$? É a probabilidade a priori do sujeito ter a doença. Mas isso não é mais que a incidência da doença na população em geral, já que o Hipólito não tinha motivos especiais para achar que era portador. Então $p(Doente) = 1/10000 = 0.0001$. Então

$$p(Doente|Pos) = \frac{0.98 \cdot 0.0001}{0.98 \cdot 0.0001 + 0.01 \cdot 0.9999} \approx 0.01$$

Ou seja, apesar do teste dar positivo, a probabilidade do Hipólito ter a doença é da ordem de um ponto percentual! O teste é praticamente inútil! E porquê? Note-se que o quociente que calculámos é dominado pelo facto da incidência da doença na população ser tão baixa. Quando o teste dá positivo, estamos a lidar com dois acontecimentos improváveis: um deles é o teste ter falhado, mas o outro é o Hipólito, escolhido ao acaso de entre a população, ter de facto a doença. A questão é que este segundo acontecimento é muitíssimo mais improvável que o primeiro, e, no caso de um teste dar positivo sobre um individuo escolhido ao acaso, é muito mais provável que se tenha dado uma falha do teste (admitidamente improvável) do que de facto termos apanhado um dos pouquíssimos individuos que possuem a doença.
Na verdade podemos ver o problema assim: Se o teste deu positivo, então ou o Hipólito está de facto doente (hipótese "D"), ou o teste é um falso positivo (hipótese "FP"). Ou seja, $p(D) + p(FP) = 1$. Mas a probabilidade de um falso positivo (0.01) é 100 vezes maior que a probabilidade a priori de o Hipólito estar doente (0.0001). Então $1 = p(D) + p(FP) = p(D)(1 + 100)$. Então $p(D) = 1/101 \approx 0.01$

Este problema ilustra a necessidade de uma grande precisão em qualquer teste que incida sobre doenças raras.

Nota: a "pista"é um caso extremo do problema. Obviamente o Hipólito não pode ter uma doença uterina porque não tem útero. Assim sendo a probabilidade do teste estar errado é igual a 1, independentemente da precisão do mesmo ser muito alta, já que o Hipólito pertence a uma (sub)população onde a incidência da doença é exactamente zero.

Nota 2: Um dia uma aluna apontou a propósito deste exercício que "o"Hipólito pode ser um hermafrodita! Tem razão, e nesse caso teríamos que contar com a proporção de hermafroditas na população! Outro aluno apontou que não sabemos se o Hipólito é um homem ou uma mulher. Isso não é exacto, pois, a priori, sabemos que o número de mulheres chamadas

"Hipólito"é muito pequeno na população. Mas de novo, tem razão no sentido que teríamos que levar em conta esse número para sermos exactos. Em qualquer dos casos, parece-me que ambas as probabilidades seriam pequenas face à probabilidade (de 1 por cento) do teste ter simplesmente dado um resultado errado. Ambos os cometários são interessantes porque mostram que as respostas que obtemos estão sempre condicionadas, não apenas às meras probabilidades prévias, mas ainda mais à própria fase de formulação dos problemas - é crucial ter imaginação (e bom senso) para construir o espaço de hipóteses correcto, e isso é algo que o formalismo das probabilidades (Bayesiano ou clássico) não pode fazer por nós. Isso pode no entanto mostrar-nos como é falaciosa a preocupação com a suposta "subjectividade"decorrente de postular probabilidades prévias, quando há tantas mais coisas que têm necessariamente que ser postuladas.

6.Sejam X, Y, Z proposições. Diz-se que X é independente de Y sabendo (ou "condicionado a") Z se $p(X|YZ) = p(X|Z)$. Mostre que

a) X é independente de Y sabendo Z se e só se $p(XY|Z) = p(X|Z)(Y|Z)$.

b) Se X é independente de Y sabendo Z então Y é independente de X sabendo Z.

c) Se X é independente de Y sabendo Z então não-X também é independente de Y sabendo Z.

Solução:

a) Supondo independência, temos $p(XY|Z) = p(X|YZ)p(Y|Z) = p(X|Z)p(Y|Z)$. Recíprocamente, assumindo que $p(XY|Z) = p(X|Z)p(Y|Z)$, vem que $p(X|YZ) = p(XY|Z)/p(Y|Z) = p(X|Z)p(Y|Z)/p(Y|Z) = p(X|Z)$.

b) Feita a alínea a) torna-se um mero corolário da simetria do produto.

c)

$$\begin{aligned} p(\overline{X}Y|Z) &= p(\overline{X}|YZ)p(Y|Z) = (1 - p(X|YZ))p(Y|Z) \\ &= p(Y|Z) - p(X|YZ)p(Y|Z) \overset{(\text{hip.})}{=} p(Y|Z) - p(X|Z)p(Y|Z) \\ &= p(Y|Z)(1 - p(X|Z)) = p(Y|Z)p(\overline{X}|Z) \end{aligned}$$

7. Suponha que quer estimar a proporção θ de fumadores numa dada população. Suponha que recolheu uma amostra de vinte indivíduos e que registou sete fumadores nessa amostra. Calcule a distribuição posterior de θ quando

a) inicialmente atribui a mesma credibilidade a todos os valores possíveis da proporção θ.

b) inicialmente supõe que $p(\theta)$ é descrita por uma distribuição $Beta(3, 12)$.

c) Na continuação da alínea b), calcule a aproximação Normal da distribuição posterior. Use ainda essa aproximação para obter um intervalo de credibilidade a 95 por cento para o valor de θ (recorde que um intervalo do tipo $[\mu \pm 1.96\sigma]$ contém 95 por cento da probabilidade de uma distribuição Normal).

Solução:

a) Para um prior $Beta(a, b)$ e um modelo Binomial temos

$$p(\theta|x, I) \propto p(x|\theta)p(\theta|I) \propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} = \theta^{a+x-1}(1-\theta)^{b+n-x-1}$$

que é o núcleo de uma $Beta(a + x, b + n - x)$.

Ora um prior constante é o mesmo que um prior $Beta(a, b)$ com $a = b = 1$. Então o posterior é (com $x = 7, n = 20, n - x = 13$)

$$p(\theta|x, I) = Beta(1 + x, 1 + n - x) = Beta(8, 14) \propto \theta^7(1-\theta)^{13}.$$

b) Basta repetir a alínea anterior, agora com $a = 3, b = 12$. Então o posterior é (com $x = 7, n - x = 13$)

$$p(\theta|x, I) = Beta(3 + x, 12 + n - x) = Beta(10, 25) \propto \theta^9(1-\theta)^{24}.$$

c) A aproximação normal de uma $Beta(a, b)$ é uma normal $N(\mu, \sigma^2)$ com $\mu = \dfrac{a-1}{a+b-2}$ e $\sigma^2 = \dfrac{\mu(1-\mu)}{a+b-2}$. No caso presente, como $a = 10, b = 25$, temos $\mu = 2.7 \times 10^{-1}$, $\sigma^2 = 6.0 \times 10^{-3}$, portanto $\sigma = 7.7 \times 10^{-2}$. Então o intervalo pretendido é $[\mu \pm 1.96\sigma] = [0.12, 0.42]$.

FIM