Primeira Prova - Gabarito
Fundamentos Estatísticos para Ciência dos Dados

06/04/2016

1. Num conjunto de documentos, apenas 1% deles são relevantes para uma certa busca de um usuário. Divida os documentos em R (relevantes) e NR (não-relevantes). Um algoritmo de recuperação de informação retorna alguns documentos de cada vez. Se o documento for do tipo R, ele tem probabilidade 0.20 de ser retornado. Se for do tipo NR, ele tem probabilidade 0.05 de ser retornado. Dado que um documento foi retornado, qual a probabilidade de ele seja relevante?

   Solução: Basta usar probabilidades condicionais (ou a regra de Bayes). Denotando o evento de que um documento foi retornado por Ret, temos:

$$\mathbb{P}(R|Ret) = \frac{\mathbb{P}(R, Ret)}{\mathbb{P}(Ret)} = \frac{\mathbb{P}(Ret|R)\mathbb{P}(R)}{\mathbb{P}(Ret)}$$

   O numerador é fácil a partir da informação fornecida: $\mathbb{P}(Ret|R)\mathbb{P}(R) = 0.20 \times 0.01 = 0.002$. Quanto ao denominador, o evento $[Ret] = [Ret \cap R] \cup [Ret \cap NR]$ e estes dois eventos são disjuntos. Portanto, a probabilidade de Ret é a soma das probabilidades desses dois eventos:

$$\begin{aligned}\mathbb{P}(Ret) &= \mathbb{P}(Ret, R) + \mathbb{P}(Ret, NR) \\ &= \mathbb{P}(Ret|R)\mathbb{P}(R) + \mathbb{P}(Ret|NR)\mathbb{P}(NR) \\ &= 0.20 \times 0.01 + 0.05 \times (1 - 0.01) = 0.0515\end{aligned}$$

   Portanto, $\mathbb{P}(R|Ret) = 0.002/0.0515 = 0.039$.

---

2. Um fato empírico recorrente tem sido a verificação de que uma fração minúscula dos maiores jobs são responsáveis por metade da carga total de um sistema. Por exemplo, é comum que os 1% maiores jobs sejam respondam por metade da carga. Seja X o tamnho aleatório de um job. Uma densidade de probabilidade para X sempre considerada é a seguinte:

$$f(x) = \begin{cases} 2/x^3, & \text{se } x > 1 \\ 0, & \text{caso contrário} \end{cases}$$

   Qual o intervalo de valores possíveis de X? Obtenha $\mathbb{E}(X)$ e $\mathbb{P}(X > 100)$.

   Solução: O suporte da distribuição de X éo intervalo $(1, \infty)$. Temos

$$\mathbb{E}(X) = \int_1^\infty x \frac{2}{x^3} dx = 2 \int_1^\infty x^{-2} dx = 2 \left(-x^{-2+1} \Big|_1^\infty \right) = 2 \left(\frac{-1}{\infty} - \frac{-1}{1} \right) = 2$$

   e

$$\mathbb{P}(X > 100) = \int_{100}^\infty \frac{2}{x^3} dx = 2 \left(-\frac{1}{2x^2} \Big|_{100}^\infty \right) = \left(\frac{-1}{\infty^2} - \frac{-1}{100^2} \right) = 0.0001$$

---

3. Tabelas hash são incríveis mas os problemas começam quando tentamos armazenar mais de um item no mesmo slot. A eficiência de todos os algoritmos de hash dependem de quantas vezes isso acontece. A seguinte situação é uma caricatura relevante para o cálculo desta eficiência. Existem três itens para serem alocados e 10 posições possíveis para isto. Cada item é alocado de forma

independente dos demais de forma que pode haver colisão, quando mais de um item é alocado a uma mesma posição. Cada uma das 10 posições possui a mesma probabilidade de ser escolhida.

Seja X o número aleatório de posições *distintas* escolhidas pelos três itens. Encontre a distribuição de probabilidades de X apresentando as duas listas, de valores possíveis e de probabilidades associadas.

Solução: Vamos estabelecer um espaço amostral Ω simples de associar probabilidades e definir X como uma função avaliada em cada elemento $\omega \in \Omega$.

Assuma que as 10 posições estejam numeradas de 0 a 9 e que os três itens estão rotulados como 1, 2 e 3. Seja

$$\Omega = \left\{(i, j, k) \in \{0, 1, \ldots, 9\}^3\right\}$$

o conjunto de todas as atribuições possíveis de posições aos três itens. Por exemplo, $\omega = (5, 5, 5)$ significa que os 3 itens foram alocados ao mesmo slot 5 enquanto $\omega = (3, 9, 3)$ significa que o item 1 foi alocado ao slot 3, o item 2 foi para o slot 9 e o item 3 também foi alocado ao slot 3.

Como cada item escolhe seu slot independentemente e com igual probabilidade, temos $\mathbb{P}(\omega) = 1/10^3$ para todo $\omega \in \Omega$. Assim, a probabilidade de qualquer evento $B \subset \Omega$ é dada por $\mathbb{P}(B) = n/10^3$ onde n é o número de elementos em B.

Mas nosso interesse não é neste espaço amostral mas num resumo desses resultados, no número $X(\omega)$ de slots distintos de ω. Assim, $X((5, 5, 5)) = 1$ enquanto que $X((3, 9, 2)) = 3$.

O suporte de X é o conjunto discreto e finito frmado por $\{1, 2, 3\}$. Queremos encontrar as três probabilidades, uma para cada um dos 3 valores possíveis. Devemos escrever $[X = x]$ como um evento B em Ω e obter sua probabilidade.

O evento $B = [X = 1]$ corresponde a escolher a mesma posição para os três itens:

$$B = [X = 1] = \{(1, 1, 1),\ (2, 2, 2), \ldots, (9, 9, 9)\}$$

e portanto

$$\mathbb{P}(X = 1) = 10\ \frac{1}{10^3} = 0.01$$

Ao invés de calcular a probabilidade do evento $[X = 2]$, vamos calcular o outro valor extremo de X.

$$B = [X = 3] = \{(i, j, k) \in \Omega \text{ tais que } i \neq j \neq k \neq i\}$$

Este conjunto possui $10 \times 9 \times 8$ elementos e portanto

$$\mathbb{P}(X = 3) = \frac{10 \times 9 \times 8}{10^3} = 0.72$$

Resta obter $\mathbb{P}(X = 2)$. O evento $[X = 2]$ é um pesadelo, com muitas possibilidades a considerar. Será mais fácil obter a probabilidade usando que

$$\mathbb{P}(X = 2) = 1 - \mathbb{P}(X = 1) - \mathbb{P}(X = 3) = 1 - 0.01 - 0.72 = 0.27$$

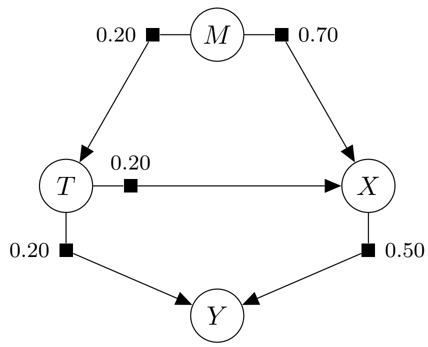
Se você quiser trilhar o caminho espinhoso, vamos lá. Precisamos encontrar quantas configurações $\omega \in \Omega$ correspondem ao evento $[X = 2]$. Podemos raciocinar em etapas de seguinte forma:

- Primeiro escolha dois slots distintos. Existem $\binom{10}{2} = 10 \times 9/2 = 45$ possibilidades. Por exemplo, $\{3, 5\}$ é uma dessas 45 possibilidades.
- Um dos slots escolhidos deverá ser escolhido para ser duplicado e existem duas possibilidades. Por exemplo, no caso acima, teríamos 355 e 335. Assim, temos $2 \times 45 = 90$ possibilidades.
- Com cada dessas possbilidades (como 355), temos 3 itens para receber o slot que não se repete (neste caso, 3).
- Assim, temos ao todo 3×90 configurações em B.

Portanto, $\mathbb{P}(X = 2) = \mathbb{P}(B) = 3 \times 90/10^3 = 0.27$, a mesma resposta que encontramos pelo método do complementar.

---

4. Um programa possui quatro módulos: M, X, T e Y. Eles são executados de acordo com o grafo abaixo e com um input fornecido ao módulo M. Isto é, a entrada ocorre apenas pelo módulo M. O programa pode ser interrompido em qualquer módulo, incluindo o módulo de entrada. O módulo M pode chamar T com probabilidade 0.20 ou X com probabilidade 0.70. O programa pode ser interrompido no módulo M com probabilidade $1 - 0.20 - 0.70 = 0.10$. Estando sendo executado o módulo T, o módulo X pode ser chamado com probabilidade 0.20 ou chamar Y com probabilidade 0.20 ou ser interrompido com probabilidade $1 - 0.20 - 0.20 = 0.60$. Etc.



Usando estas probabilidades, obtenha a probabilidade de que o módulo Y seja executado dado que T foi executado. Calcule também a probabilidade de que o módulo T tenha sido executado dado que X foi executado.

Solução: Dado que o program está no módulo T, ele pode atingir Y apenas através de dois caminhos disjuntos: $T \to Y$ diretamente ou $T \to X \to Y$. Assim,

$$\mathbb{P}(Y|T) = \mathbb{P}(\text{ caminho } TY|\text{está em } T) + \mathbb{P}(\text{ caminho } TXY|\text{está em } T) = 0.20 + 0.20 \times 0.50 = 0.30$$

A segunda probabilidade é mais sutil. A probabilidade inversa é muito fácil pois existe apenas uma maneira de chegar a X dado que você está em T e esta probabiliade é lida diretamente da figura: $\mathbb{P}(T \to X|T) = 0.20$. Isto sugere inverter a probabilidade condicional desejada.

$$\mathbb{P}(T|X) = \frac{\mathbb{P}(X \text{ e } T)}{\mathbb{P}(X)} = \frac{\mathbb{P}(X|T)\mathbb{P}(\text{ passar por } T)}{\mathbb{P}(\text{ passar por } X)} = \frac{0.20 \times 0.20}{\mathbb{P}(\text{ passar por } X)}$$

Para o denominador, vamos considerar os caminhos disjuntos para chegar a X a partir de M: temos $M \to T \to X$ e $M \to X$. Portanto,

$$\mathbb{P}(\text{ passar por } X) = \mathbb{P}(M \to T \to X) + \mathbb{P}(M \to X) = 0.20 \times 0.20 + 0.70 = 0.74.$$

e então $\mathbb{P}(T|X) = (0.20 \times 0.20)/0.74 = 0.054$. Assim, se o programa está em X, é pequena a chance de que ele tenha chegado ali a partir de T.

---

5. Um vetor aleatório contínuo (X, Y) assume valores no quadrado $[0, 1] \times [0, 1]$ com uma densidade de probabilidade dada por

$$f(x, y) = f_{X,Y}(x, y) = \begin{cases} c(x + y), & \text{se } (x, y) \in [0, 1] \times [0, 1] \\ 0, & \text{caso contrário} \end{cases}$$

- Encontre a constante de normalização c.
- Considere quatro pequenos quadradinhos, todos de área 0.1^2, em $[0, 1] \times [0, 1]$. Eles são:
  - $A_1 = [0, 0.1] \times [0, 0.1]$.

- $A_2 = [0.9, 1] \times [0, 0.1]$.
- $A_3 = [0, 0.1] \times [0.9, 1]$.
- $A_4 = [0.9, 1] \times [0.9, 1]$.

Responda: Qual deles possui a seguintes regiões possui maior probabilidade de ocorrência $\mathbb{P}((X,Y) \in A_k)$? E a menor probabilidade? Observe que não precisa fazer nenhum cálculo exato, basta dar alguma justificativa para sua resposta.

- Obtenha a densidade marginal $f_X(x)$ da v.a. X.
- Obtenha a densidade condicional de Y dado que $X = 1/2$. Isto é, obtenha $f_{X|Y}(y|1/2)$.

Solução: Devemos ter

$$
\begin{aligned}
1 &= \int_0^1 \int_0^1 f(x,y)dxdy = \int_0^1 \int_0^1 c(x+y)dxdy \\
&= c\left(\int_0^1 \int_0^1 xdxdy + \int_0^1 \int_0^1 ydxdy\right) \\
&= c\left(\int_0^1 xdx + \int_0^1 ydy\right) \\
&= c\left(\frac{x^2}{2}\Big|_0^1 + \frac{y^2}{2}\Big|_0^1\right) \\
&= c(1/2 + 1/2) = c
\end{aligned}
$$

Assim, a constante de normalização é simplesmente $c = 1$.

Os quatro quadradinhos $A_1, \ldots, A_4$ possuem a mesma área. Se um deles tiver os valores $f(x,y)$ da altura da densidade de probabilidade claramente maior que os demais, ele será a região mais provável dentre os 4 considerados pois o volume sob a densidade é a probabilidade $\mathbb{P}((X,Y) \in A_k)$ da região A_k. A função $f(x,y)$ cresce tanto com x quant com y. Assim, o quadradinho com mais probabilidade é aqele mais próximo do canto $(1,1)$. Isto é, $\mathbb{P}((X,Y) \in (0.9,1)^2)$ é máxima. A menor probabilidade é a do quadradinho perto da origem $(0,0)$. Isto é, $\mathbb{P}((X,Y) \in (0.0,0.1)^2)$ é mínima.

A densidade marginal $f_X(x)$ da v.a. X é obtida por integração simples:

$$
\begin{aligned}
f_X(x) &= \int_0^1 f(x,y)dy = \int_0^1 (x+y)dy \\
&= \int_0^1 xdy + \int_0^1 ydy \\
&= x\int_0^1 dy + \frac{y^2}{2}\Big|_0^1 = x + 1/2.
\end{aligned}
$$

A densidade condicional de Y dado que $X = 1/2$:

$$
\begin{aligned}
f_{X|Y}(y|1/2) &= \frac{f(x,1/2)}{f_X(1/2)} \\
&= \frac{1/2 + y}{1/2 + 1/2} \\
&= y + 1/2
\end{aligned}
$$

---

6. Sejam X e Y os números de falhas em dois computadores de um lab numa dada semana. A distribuição conjunta dessas duas v.a.'s está na tabela abaixo.

   - Calcule a probabilidade de que haja pelo menos uma falha no lab na semana.

$P(x,y)$		x = 0	1	≥ 2
y	0	0.52	0.20	0.04
	1	0.14	0.02	0.01
	≥ 2	0.06	0.01	0.00

- Calcule $\mathbb{P}(X \geq 2)$ e $\mathbb{P}(Y \geq 2)$.
- Usando sua resposta acima, diga se as duas v.a.'s são independentes justificando sua resposta.

Solução: Pode-se calcular a probabilidade de pelo menos uma falha no lab na semana de duas maneiras, uma longa e uma rápida. A maneira longa enumera todas as possibilidades *disjuntas* de haver pelo menos uma falha. De fato, o evento $[X \geq 1 \text{ ou } Y \geq 1]$ é a união de 8 eventos disjuntos:

$$[X = 0, Y = 1],\ [X = 0, Y \geq 2],\ [X = 1, Y = 0],\ [X = 1, Y = 1],$$

$$[X = 1, Y \geq 2],\ [X \geq 2, Y = 0],\ [X \geq 2, Y = 1],\ [X \geq 2, Y \geq 2].$$

Cada um desses oito eventos está associado a uma das células da tabela de probabilidade conjunta. Assim, a probabilidade do evento $[X \geq 1 \text{ ou } Y \geq 1]$ é a soma das oito probabilidades dessa tabela:

$$\begin{aligned} \mathbb{P}(X \geq 1 \text{ ou } Y \geq 1) &= \mathbb{P}(X = 0, Y = 1) + \mathbb{P}(X = 0, Y \geq 2) + \ldots + \mathbb{P}(X \geq 2, Y \geq 2) \\ &= 0.14 + 0.06 + \ldots + 0.00 = 0.48. \end{aligned}$$

A maneira mais rápida é perceber que só não somamos uma célula da tabela. Portanto, á mais rápido obter a resposta por subtração da probabilidade total que é igual a 1. Realmente, o evento $[X \geq 1 \text{ ou } Y \geq 1]$ é o complementar do evento $[X = 0, Y = 0]$ e assim

$$\mathbb{P}(X \geq 1 \text{ ou } Y \geq 1) = 1 - \mathbb{P}(X = 0, Y = 0) = 1 - 0.52 = 0.48$$

As probabilidades $\mathbb{P}(X \geq 2)$ e $\mathbb{P}(Y \geq 2)$ são encontradas somando-se as probabilidades marginais na coluna $[X \geq 2]$ e na linha $[Y \geq 2]$:

$$\mathbb{P}(X \geq 2) = 0.04 + 0.01 + 0.00 = 0.05$$

e

$$\mathbb{P}(Y \geq 2) = 0.06 + 0.01 + 0.00 = 0.07$$

Duas v.a.'s discretas são independentes se, e somente se, $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$para todo par x, y de valores possíveis. Caso não sejam iguais para um único par (x, y), as v.a.'s não são independentes. Tomando $(x, y) = (\geq 2, \geq 2)$, os valores calculados e o valor de $\mathbb{P}(X \geq 2, Y \geq 2) = 0$ vindo da tabela, vemos que:

$$0 = \mathbb{P}(X \geq 2, Y \geq 2) \neq 0.05 \times 0.07 = \mathbb{P}(X \geq 2) \times \mathbb{P}(Y \geq 2)$$

e portanto X e Y não são independentes.