Primeira Prova
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?

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:

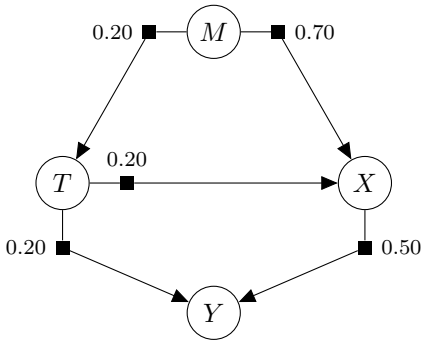
$$f(x) = \begin{cases} 2/x^3, & \text{se } x > 1 \\ 0, & \text{caso contrário} \end{cases}$$

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

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.

4. Um programa possui quatro módulos: M, X, T e Y. Eles são executados de acordo com o grafo abaixo e com o input fornecido. O programa pode ser interrompido em qualquer módulo. Assim, 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 seja executado dado que X foi executado.

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, batsa dar alguma justificativa para ua resposta.
   - Obtenha a densidade marginal $f_X(x)$ dav.a. X.
   - Obtenha a densidade condicional de Y dado que $X = 1/2$. Isto é, obtenha $f_{X|Y}(y|1/2)$.

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.

$P(x,y)$		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 a probabilidade de que haja pelo menos uma falha no lab na semana.
   - 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.