Primeira Prova – Solução
Fundamentos Estatísticos para Ciência dos Dados

19/04/2017

1. Uma imagem contém um objeto de formato e localização desconhecidos. Represente a região da imagem por um quadrado $[0,1] \times [0,1]$ (região contínua). A área do objeto na imagem é um valor desconhecido α. Imagine que um ponto aleatório (X, Y) tem uma distribuição uniforme sobre a imagem. Não se observa diretamente onde (X, Y) recai na imagem mas é possível saber se ele atingiu o objeto ou não. Seja Z a variável indicadora desse evento: $Z = 1$ se o ponto (X, Y) cai dentro da região delimitada pelo objeto e $Z = 0$, caso contrário. Mostre que $\mathbb{E}(Z) = \alpha$. Como podemos obter um valor aproximado para α a partir de n pontos lançados independentemente sobre a imagem, todos seguindo a distribuição uniforme?

   Resposta: (a) 2 pts; (b) 1 pt.

   (a) Mostre que $\mathbb{E}(Z) = \alpha$:

   Temos que a área total da imagem é $(1-0)(1-0) = 1$. A área do objeto é α. Portanto, a área da imagem externa ao objeto é $1 - \alpha$.

   Assim: $\mathbb{P}(Z = 1) = \alpha/1$ e $\mathbb{P}(Z = 0) = 1 - \mathbb{P}(Z = 1) = 1 - \alpha$

   Logo, considerando a distribuição uniforme, temos: $\mathbb{E}(Z) = 0(1-\alpha) + 1(\alpha) = \alpha$.

   (b) Como podemos obter um valor aproximado para α a partir de n pontos lançados independentemente sobre a imagem, todos seguindo a distribuição uniforme?

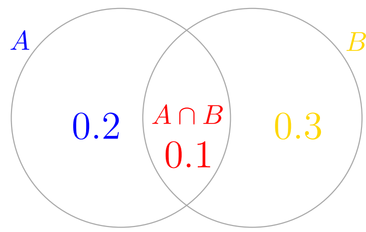
   Basta lançar um número suficientemente grande de pontos e verificar a proporção de pontos que caem dentro do objeto.

2. Os tipos de sangue que existem são A, B, AB e O. Os tipos A e O são os mais comuns, abrangendo 87% da população brasileira. O tipo B responde por 10%. O tipo AB está presente em apenas 3% da população brasileira. Uma amostra de sangue é colhida na avaliação de saúde de recrutas do exército brasileiro.

   $\mathbb{P}(AB) = 0.03 \Rightarrow \mathbb{P}(\text{Não } AB) = 1 - 0.03 = 0.97$

   - Ache a probabilidade de que o primeiro recruta a ter o tipo AB seja o décimo da fila.
     Resposta:(1 pt) $\mathbb{P}(\text{Primeiro recruta ser o décimo da fila}) = (0.97)^9(0.03)^1$
   - Ache a probabilidade de que exatamente um recruta entre os 10 primeiros seja do tipo AB.
     Resposta:(1 pt) $\mathbb{P}(\text{Exatamente um } AB \text{ em 10 recrutas}) = \binom{10}{1}(0.97)^9(0.03)^1$
   - Identifique a distribuição de X, o número de pessoas entre os SETE primeiros que NÃO possuem tipo AB. **Resposta:(0,5 pt)** Binomial: Bin(7, 0.97)
     Quanto é $\mathbb{E}(X)$? **Resposta: (0,5 pt)** $\mathbb{E}(X) = \sum_{x_i} x_i \mathbb{P}(X = x_i) = 7(0.97) = 6.79$

3. Um animal infectado pela doença da vaca louca pode exibir dois sintomas: A (comportamento agressivo) ou B (dificuldade de manter-se em pé). Estima-se que entre o gado infectado, 20% exiba o sintoma A apenas, 30% exiba apenas o sintoma B e 10% exiba os dois sintomas, com o restante sendo assintomático (não exibindo nenhum sintoma). De uma população infectada, um animal é escolhido ao acaso.



- Qual a probabilidade de ele não exiba nenhum sintoma?
  Resposta: (0,5 pt) $\mathbb{P}(\text{Não exibir sintoma}) = 1 - (\mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(A \cap B)) = 1 - (0.2 + 0.3 + 0.1) = 0.4$

- Qual a probabilidade de que ele exiba pelo menos um dos sintomas?
  Resposta: (0,5 pt) $\mathbb{P}(\text{Exibir pelo menos 1 sintoma}) = 1 - \mathbb{P}(\text{Não exibir sintoma}) = 1 - 0.4 = 0.6$

- Qual a probabilidade de ele tenha os dois sintomas dado que ele tem o sintoma A?
  Resposta: (1 pt)
  $\mathbb{P}(\text{Exibir os dois sintomas dado que tem o sintoma A}) = \mathbb{P}(A \cap B | A) =$
  $= \frac{\mathbb{P}((A\cap B)\cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)} = \frac{0.1}{0.3}$

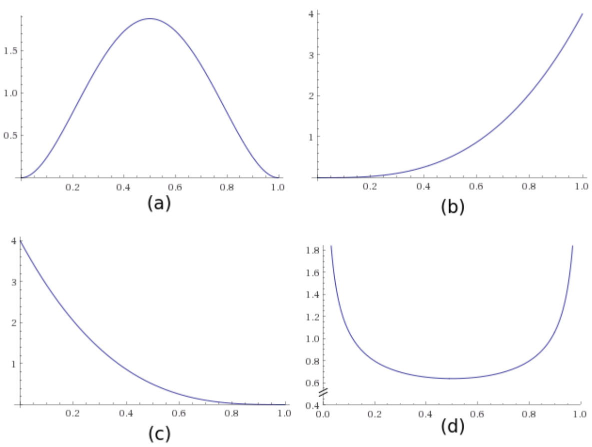
- Qual a probabilidade de ele tenha os dois sintomas dado que ele tem pelo menos um dos sintomas?
  Resposta: (1 pt)
  $\mathbb{P}(\text{Exibir os dois sintomas dado que tem pelo menos um dos sintomas}) = \mathbb{P}(A \cap B | A \cup B) =$
  $= \frac{\mathbb{P}((A\cap B)\cap(A\cup B))}{\mathbb{P}(A\cup B)} = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A\cup B)} = \frac{0.1}{0.6}$

4. A carga numa rede é representada como uma v.a. contínua X com suporte $\mathcal{S} = (0, 1)$, sendo que $X \approx 0$ significa carga quase nula e $X \approx 1$ carga máxima. Existem quatro possíveis densidades de probabilidade para esta v.a.:

   (A) $f(x) = 30x^2(1-x)^2$
   (B) $f(x) = 4x^3$
   (C) $f(x) = 4(1-x)^3$
   (D) $f(x) = 1/(\pi\sqrt{x(1-x)})$

   onde $x \in \mathcal{S}$.

- Mapeie a densidade correspondente para cada um dos seguintes cenários:
  - Tipicamente, a carga na rede é baixa. **Resposta: (0,25 pt)** (c)
  - Tipicamente, a carga na rede ou é baixa ou é alta. **Resposta: (0,25 pt)** (d)
  - Tipicamente, a carga na rede é alta. **Resposta: (0,25 pt)** (b)
  - Tipicamente, a carga na rede fica em torno de 0.5. **Resposta: (0,25 pt)** (a)



- Para a densidade (B), obtenha $\mathbb{P}(X > 0.2)$

  Resposta: (0,5 pt) $\mathbb{P}(X > 0.2) = 1 - \mathbb{P}(X \leq 0.2) = 1 - \int_0^{0.2} 4x^3 dx = 1 - 0.0016 = 0.9984$

- Para a densidade (B), obtenha $\mathbb{E}(X)$

  Resposta: (0,5 pt) $\mathbb{E}(X) = \int_0^1 xf(x) = \int_0^1 x4x^3 dx = 4/5 = 0.8$

- Sem fazer nenhuma conta adicional, diga qual das densidades tem o maior desvio-padrão. Justifique sua resposta em poucas palavras.

  Resposta: (1 pt) A densidade com maior desvio padrão é (d), já que a carga na rede é a unica variando entre alto e baixo.

5. Usando a definição de probabilidade condicional de um evento A dado um evento B, mostre que

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\ \mathbb{P}(B|C)\ \mathbb{P}(C)$$

Resposta: (3 pts)

Temos que: $\mathbb{P}(A|B \cap C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)}$ e $\mathbb{P}(B|C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}$.

Portanto: $\mathbb{P}(A|B \cap C)\ \mathbb{P}(B|C)\ \mathbb{P}(C) = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)}\ \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}\ \mathbb{P}(C)$.

Simplificando, temos: $\mathbb{P}(A|B \cap C)\ \mathbb{P}(B|C)\ \mathbb{P}(C) = \mathbb{P}(A \cap B \cap C)$.

6. Um modelo probabilístico para uma hash table: nós alocamos ao acaso n itens em uma de m posições, com cada posição sendo escolhida completamente ao acaso e os itens sendo alocados de forma independente uns dos outros. Qual a probabilidade de que a urna 3 esteja vazia?

Resposta:

Considerando a escolha de alocação dos itens de maneira aleatória, temos que a $\mathbb{P}$(Alocar 1 item na posição 1) = $\mathbb{P}$(Alocar 1 item na posição 2) = ... = $\mathbb{P}$(Alocar 1 item na posição m) = $1/m$. Assim, a $\mathbb{P}$(Alocar 1 item na posição 3) = $1/m$ e, portanto, a $\mathbb{P}$(NÃO alocar 1 item na posição 3) $= 1 - 1/m = (m-1)/m$. **(1,5 pts)**

Temos que esse evento é repetido n vezes, já que queremos alocar n itens na tabela de m posições. Portanto, seguindo uma distribuição binomial, queremos saber a probabilidade de que nenhum item seja alocado à posição 3 em n repetições desse evento.

Assim, $\mathbb{P}(\text{Alocar 0 itens na posição 3}) = \binom{n}{0}\left(\frac{m-1}{m}\right)^{n-0}\left(\frac{1}{m}\right)^{0} = \left(\frac{m-1}{m}\right)^{n}$. **(1,5 pts)**

7. Um programa consiste de dois módulos. O primeiro módulo contém um erro com probablidade 0.15. O segundo módulo é mais complexo. Ele tem uma probabilidade de 0.35 de conter um erro neste módulo, independentemente do primeiro que ocorrer no primeiro módulo. Um erro apenas no primeiro módulo, sem erros no segundo, implica em interromper o programa com probabilidade 0.5. Para erros apenas no segundo módulo, esta probabilidade é 0.80. Se existem erros nos dois módulos, o programa é interrompido com probabilidade 0.90. Suponha que o programa foi interrompido. Qual a probabilidade de que existam erros nos dois módulos?

Resposta: Partindo do enunciado da questão, temos que $\mathbb{P}(m_1 \text{ errar }) = 0.15$ e $\mathbb{P}(m_2 \text{ errar }) = 0.35$. Como os erros são independentes:

$\mathbb{P}(m_1 \text{ errar } \cap m_2 \text{ errar }) = \mathbb{P}(m_1 \text{ errar })\mathbb{P}(m_2 \text{ errar }) = 0.0525$ **(*)**.

Além disso, temos que:

$\mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar } \cap m_2 \text{ N-errar }) = 0.5$

$\mathbb{P}(\text{Interrromper} \mid m_1 \text{ N-errar } \cap m_2 \text{ errar }) = 0.8$

$\mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar } \cap m_2 \text{ errar }) = 0.9$ **(**)**.

Queremos saber:

$\mathbb{P}(m_1 \text{ errar } \cap m_2 \text{ errar } \mid \text{Interrromper}) = \frac{\mathbb{P}(\text{Interrromper } \cap \text{ m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar})}{\mathbb{P}(\text{Interromper})} = \ ?$ **(***)**.

Passo 1: Calcular $\mathbb{P}(\text{Interrromper } \cap \text{ m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar})$**. (1,5 pts)**

Em **(**)** temos:

$\mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar } \cap m_2 \text{ errar }) = 0.9 = \frac{\mathbb{P}(\text{Interrromper } \cap \text{ m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar})}{\mathbb{P}(\text{m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar})}$

$\Leftrightarrow \mathbb{P}(\text{Interrromper} \cap \text{m}_1 \text{ errar} \cap \text{m}_2 \text{ errar}) = \mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar} \cap m_2 \text{ errar })\mathbb{P}(\text{m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar})$

Portanto, por **(*)**: $\mathbb{P}(\text{Interrromper } \cap \text{ m}_1 \text{ errar } \cap \text{ m}_2 \text{ errar}) = 0.9 * 0.0525 = 0.04725$

Passo 2: Calcular $\mathbb{P}(\text{Interromper})$**. (1,5 pts)**

Suponha que a probalidade de que o sistema não seja interrompido caso os dois módulos estejam funcionando seja 0. Assim, o sistema é interrompido quando m_1 e m_2 erram, OU quando m_1 erra e m_2 não erra, OU quando m_1 não erra e m_2 erra. Portanto, temos:

$$\begin{aligned}\mathbb{P}(\text{Interrromper}) = &\ \mathbb{P}(\text{Interrromper } \cap\ m_1 \text{ errar } \cap m_2 \text{ errar }) + \\ &\ \mathbb{P}(\text{Interrromper } \cap\ m_1 \text{ N-errar } \cap m_2 \text{ errar }) + \\ &\ \mathbb{P}(\text{Interrromper } \cap\ m_1 \text{ errar } \cap m_2 \text{ N-errar })\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\text{Interrromper}) = &\ \mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar } \cap m_2 \text{ N-errar })\mathbb{P}(m_1 \text{ errar })\mathbb{P}(m_2 \text{ N-errar }) + \\ &\ \mathbb{P}(\text{Interrromper} \mid m_1 \text{ N-errar } \cap m_2 \text{ errar })\mathbb{P}(m_1 \text{ N-errar })\mathbb{P}(m_2 \text{ errar }) + \\ &\ \mathbb{P}(\text{Interrromper} \mid m_1 \text{ errar } \cap m_2 \text{ errar })\mathbb{P}(m_1 \text{ errar })\mathbb{P}(m_2 \text{ errar })\end{aligned}$$

$\mathbb{P}(\text{Interrromper}) = 0.5 * 0.15 * (1 - 0.35) + 0.8 * (1 - 0.15) * 0.35 + 0.9 * 0.15 * 0.35 = 0.334$

Voltando em **(***)**, então: $\mathbb{P}(m_1 \text{ errar } \cap m_2 \text{ errar } \mid \text{Interrromper}) = \frac{0.04725}{0.33395} = 0.1415$