Lista de Exercícios #0

Mineração de Dados - QS 2020

Instruções

1. A lista deve ser feita individualmente.
2. Desconsiderando as questões de implementação, **todas** as respostas devem ser manuscritas. O material entregue deve consistir da cópia digitalizada da folha, por exemplo, uma foto **legível** da folha de soluções.
3. Os exercícios de implementação podem ser feitos na linguagem de programação que quiser, desde que para executar o seu código não seja necessária a instalação de **nenhum** software com custo maior que zero, isso inclui o Windows :).
4. Qualquer tentativa de fraude identificada será punida de acordo com o código de honra.
5. Além das respostas de cada pergunta no formato textual, você deve escolher um dos exercícios marcados com $\star$ e enviar o link para *download* de um vídeo descrevendo a solução do exercício. O vídeo deve ter no máximo 5 minutos e no início deve aparecer o rosto e carteirinha do aluno. A falha no recebimento do vídeo fará a lista ser desconsiderada.

Exercício 1 $\star$ Probabilidade, Principio da inclusão-exclusão

Uma cidade possui 100.000 moradores e 3 jornais: A, B e C. A proporção de moradores que leem esses jornais são como seguem:

A: 10%	A e B: 8%	A e B e C: 1%
B: 30%	A e C: 2%	
C: 5%	B e C: 4%	

Existem, por exemplo, 8.000 pessoas que leem os jornais A e B. Responda as questões abaixo:

1. Qual o número de pessoas que lê apenas um jornal?
2. Quantas pessoas leem pelo menos dois jornais?
3. Se os jornais A e C são de bairro e B da mesorregião da cidade, quantas pessoas leem pelo menos um dos jornais locais e o jornal da mesorregião?
4. Quantas pessoas não leem nenhum jornal?
5. Quantas pessoas leem apenas um jornal local e o jornal da mesorregião?

Exercício 2 Probabilidade condicional

Considere uma companhia de seguros que classifica pessoas em três categorias: *risco baixo*, *risco médio* e *risco alto*. O histórico indica que a probabilidade de uma pessoa de risco baixo, médio e alto sofrerem um acidente em um dado ano é, respectivamente, 5%, 15% e 30%. Se 20% da população é classificado como risco baixo, 50% como risco médio e 30% como risco alto, qual a proporção de pessoas que sofrem acidentes em um dado ano? Se o segurado *CarefulPerson* não sofreu acidentes em 2019, qual a probabilidade de ele ser de risco baixo?

Exercício 3 Probabilidade, esperança, variável aleatória

Um amigo seu, que adora estratégias para ficar rico com jogos, bolou um novo esquema para ganhar dinheiro em jogos de roleta. O plano dele é sempre fazer apostas de R$1 no *vermelho*. Caso dê *vermelho* (chance de $\frac{18}{38}$), ele pega o R$1 de lucro e para de jogar. Caso ele perca a aposta, nas próximas duas rodada ele aposta R$1 no *vermelho* e então para de jogar, independente do resultado. Seja X os ganhos do jogador ao parar de jogar seguindo essa estratégia.

1. Determine os valores possíveis para X e $P(X > 0)$.
2. Teu amigo te convenceu? Essa é uma estratégia vencedora? Justifique sua resposta.
3. Encontre $\mathbb{E}[X]$, sendo $\mathbb{E}$ o operador de esperança.

Exercício 4 Álgebra Linear, Multiplicação de matrizes

Considere $\Delta = \mathbf{x}^T \mathbf{\Sigma}^{-1} \mathbf{x}$ em que $\mathbf{x} \in \mathbb{R}^2$ e $\mathbf{\Sigma} = \begin{bmatrix} 3 & 0 \\ 0 & 5 \end{bmatrix}$. Definindo $\Delta = 1$ encontramos a equação de uma forma geométrica. Qual é a equação e a forma encontrada?

Exercício 5 $\star$ Programação, Amostragem por rejeição

Implemente um programa que crie um histograma de amostras geradas da distribuição de probabilidade: $f(x) = 0,3\mathcal{N}(3,1) + 0,7\mathcal{N}(8,4)$, em que $\mathcal{N}(\mu, \sigma^2)$ é a distribuição Gaussiana com média μ e variância σ^2. Seu programa deve utilizar a abordagem de amostragem por rejeição seguindo a ideia abaixo. Esse tipo de técnica é utilizada para gerar amostras de uma distribuição de probabilidades da qual não sabemos/é complicado obter novas amostras. A ideia básica é gerar amostras de uma distribuição de probabilidades diferente (que sabemos amostrar) que "englobe" a distribuição desejada dada uma constante M, e utilizá-la para validar amostras. Para saber mais consulte a Seção 11.1.2 do livro *Pattern Recognition and Machine Learning*.

1. Crie em **b** 1.000 amostras da função de densidade de probabilidade $g(x) = \mathcal{N}(6, 10)$, ou seja, $b_i \sim \mathcal{N}(6, 10), \forall i \in \{1, \ldots, 1000\}$;
2. Seja $M = \max\left(\left\{\frac{f(b_i)}{g(b_i)}\right\}_{i=1}^{10^3}\right)$;
3. Gere um valor $u \sim Unif(0, 1)$, ou seja, um valor distribuído de acordo com uma distribuição uniforme entre 0 e 1;
4. Gere um valor $z \sim \mathcal{N}(6, 10)$;
5. Caso $u < \frac{f(z)}{Mg(z)}$, a amostra gerada (z) é aceita, caso contrário a amostra é rejeitada;

6. Atualize M como $\max\left(\left\{M, \frac{f(z)}{g(z)}\right\}\right)$;

7. Volte ao passo 3 para gerar uma nova amostra.

Para fins de ilustração, segue um exemplo de histograma gerado para esse problema. Gere quantas amostras forem suficientes para que o formato da distribuição fique bem definido, o exemplo abaixo foi obtido com 10.000 amostras (aceitas).

Extra: Você pode reparar que a taxa de rejeição para esse problema é relativamente alta, consegue ter uma ideia da razão? O que poderia ser feito para melhorar isso?