



7.9

**Atenção:** Leia atentamente as questões e escreva de forma clara e legível.

Nome: Mariano Dornelles de Freitas

Curso: Ciência da Computação Data: 05/09/2019

- 1) Na limpeza de um conjunto de dados ruidoso e cheio de valores faltantes. Cite 3 abordagens para complementar os valores faltantes (1.5). Para cada uma escreva um possível problema de usar essa abordagem. (1.5) <sup>0.7</sup> <sup>1.5</sup>

Regressão, média (tudo o que estiver, quanto mais, melhor),  
valor máximo e mínimo. <sup>2</sup> <sup>Não todos os pontos</sup>

Média: para valores contínuos, a média acaba se tornando um <sup>2</sup>  
valor discrepante em algumas posições (início de fim, por exemplo).

Min-max: similar à média, pode agravar o ruído em certas ocasiões. <sup>2</sup>

Regressão: eficaz para valores contínuos, mas não muito para ab-  
stúrios. <sup>?</sup>

- 2) Se formos extrair dados do Twitter a fim de minerar probabilidades de compra para um determinado produto. Qual tipo de dado estamos falando (quando a estrutura) e quais problemas podemos encontrar para tratar esses dados? (2)

Dado não-estruturado (tortos), <sup>2</sup> <sup>Se esse tipo de dados sempre require</sup>  
pré-processamento, pois contém bastante ruído, traçar a relação entre  
eles tende a ser uma tarefa difícil também.

- 3) Brevemente, descreva a relação entre KDD, mineração de dados, e análise de dados. (2)

KDD é um conjunto de ações que devem ser tomadas para extrair infor-  
mações úteis de dados. Mineração de dados é uma das etapas do  
KDD, onde é usado um algoritmo para filtrar as informações, embora o  
termo seja comumente usado para referir-se a KDD. Análise de  
dados é a (última etapa) que se faz na última etapa do KDD, com os  
dados processados, se faz uma análise a fim de obter informações.

4) Analise o seguinte conjunto e responda:

4.1) Caso nosso projeto de mineração tenha como objetivo encontrar correlação entre produtos e lucro. Quais informações podemos descartar? (1)

ID, Descrição

4.2) Considerando esse objetivo (4.1), devo alterar (adicionar ou editar) algo nesse conjunto?

Por que? (1)

Deve ser adicionada como lucro  
O valor de compra e venda poderiam se tornar uma informação  
só (lucro), para otimizar os dados e facilitar a visualização.  
Uma coluna de lucro total poderia ser útil (lucro x unidades  
vendidas), para uma informação mais absoluta dos lucros de ca-  
da produto. para a clareza

ID	Produto	Descricao	Valor compra	Valor venda	Unidades vendidas
1	Monitor LG 29p	Entradas 1,2,3. 29p. Full HD IPS 29UM68	250	290	29
...	...	...	...	...	...
999	Monitor Samsung 30p	Entr. 1,2,3,4. 30p. 4K	500	600	45

5) Em geral, dado um *dataset* X, qual a diferença entre a extração de um subconjunto, e a redução de dimensionalidade de X? (1)

Extração refere-se a atrelar este subconjunto a uma variável  
para que fique melhor de se trabalhar, X ainda possui o subconj.  
seria uma seleção, praticamente. Reduzir X é descartar uma de  
suas linhas ou colunas, ou juntá-las, se for o caso.  
pod ser em um caso, mas não para o geral