Universidade Federal de Alagoas

Instituto de Computação

Ciência de Dados

Professor: Bruno Pimentel

Aluno: Maxwell Esdra Acioli Silva

Data: 07/07/2020

Prova 1

1. Defina Ciência de Dados. (2 pontos)

Ciência de dados pode ser definida como uma metodologia utilizada para fazer o tratamento, análise e consequentemente gerar conhecimento a partir de uma base de dados de interesse. Esta área surgiu há cerca de trinta anos, porém teve um grande desenvolvimento a partir da evolução do poder de processamento e armazenamento dos computadores modernos.

Atualmente no mundo, diariamente são gerados enormes volumes de dados através de Big Data, estes são capazes de coletar, processar e armazenar dados de forma muito rápida e que possuem uma grande variedade tanto na sua composição como sua estrutura. Com essa massa enorme de dados que são criados, abre-se um espaço para utilização da ciência de dados para explorá-los, e inferir insights a partir da análise destes.

O processo de análise dos dados desde a sua coleta até a obtenção de algum conhecimento, consiste de algumas etapas fundamentais, são elas: seleção dos dados, pré-processamento, transformação, aplicação de modelos preditivos, validação do modelo e interpretação dos resultados. Por fim, o conhecimento final obtido pode ser utilizado no processo de tomada de decisão de algum negócio.

2. Quais desafios os cientistas de dados possuem ao trabalhar com *Big Data*? (2 pontos)

Big Data pode ser definido como um grande volume de dados que são gerados em uma alta velocidade e de forma variada. Um cientista de dados irá encontrar uma série de desafios ao trabalhar com este imenso volume de dados.

O primeiro desafio que podemos destacar é o fato da não estruturação dos dados trabalhados, sendo assim é necessário fazer uma seleção dos dados que realmente interessa, bem como fazer um pré-processamento dos dados para que estes possam ser aplicados a modelos preditivos e ser possível fazer alguma análise sobre os resultados obtidos.

Como a velocidade em que os dados são gerados é alta, muitas vezes há uma necessidade de obter informações a curto prazo pois estas podem impactar positivamente na indústria ou na sociedade. Logo, o segundo desafio do cientista é obter informações sobre os dados coletados em um tempo hábil, para que estas possam ser utilizadas em algum processo tomada de decisão ou processo de negócio.

3. De que forma a Estatística pode auxiliar Ciência de Dados? (2 pontos)

Estatística auxilia a ciência de dados através de fornecimento de métodos estatísticos que possibilitam fazer a análise e obter um melhor entendimento dos dados.

Ela também auxilia no processo de definição de modelos que ajudam a predizer eventos futuros a partir de amostras de dados de uma população, ou seja, obter o entendimento de uma população analisando apenas uma parte desta. Este tipo de informação pode ser obtida através do uso de inferência estatística, ou a partir da análise de propriedades fornecidas pela estatística descritiva, como, por exemplo, utilizar média ou mediana de uma amostra para analisar o comportamento de um conjunto de dados.

Outra forma em que a estatística pode auxiliar a ciência de dados é na validação e avaliação dos modelos preditivos utilizados no processo de análise de um conjunto de dados. Esta validação pode ser feita através do uso de teste de hipótese.

4. Mostre a importância de pré-processar os dados para a extração de informação. (1 ponto)

O pré-processamento é uma etapa fundamental no processo de extração de informação de um conjunto de dados. Na maioria das vezes os dados coletados são provenientes de big data, em muitos casos os dados não estão estruturados, e têm

uma enorme variedades de informações. Estas informações podem estar inconsistentes, ou seja, podem conter dados equivocados sobre o objeto de estudo, dados redundantes, valores faltantes ou inconsistências que comprometam a integridade dos dados. Todas as inconformidades citadas anteriormente podem comprometer a qualidade da mineração, bem como comprometer o processo de aplicação de algum algoritmo de aprendizagem de máquina, logo o resultado oriundo da modelagem aplicada pode não trazer resultados que representem de forma real os dados analisados, assim o conhecimento obtido pode não ser íntegro. Por estes motivos é muito importante a etapa de pré-processamento dos dados.

5. Indique a importância de utilizar métodos de avaliação de modelos no processo de extração de informação. (1 ponto)

Técnicas de avaliação de modelos podem ajudar a entender melhor e qualificar o modelo aplicado no processo de análise de dados a fim de obter alguma informação sobre os mesmos. Um dos grandes problemas encontrados na definição de modelos preditivos é a ocorrência de overfitting, ou seja, um ajuste excessivo do modelo. Este tipo de problema pode fazer com que o modelo aplicado tenha um bom desempenho apenas quando executado sobre a base de dados utilizada no treinamento, ou seja, quando o modelo é aplicado em um novos conjuntos de dados este tende a não apresentar bons resultados. Uma forma de evitar este tipo de problema é utilizar boas técnicas de avaliação do modelo, e assim obter através do uso de alguma métrica uma medida que generalize o quão bom o seu modelo está quando este for fazer a avaliação de novos dados. Existem muitas técnicas de avaliação de modelos, dentre estas podemos citar: hold-out, subamostragem aleatória, k-fold cross-validation, leave-one-out e bootstrap.

6. Com respeito à Análise de Agrupamento, cite um exemplo de aplicação onde é preferível utilizar agrupamento hierárquico. (1 ponto)

É preferível utilizar o método de agrupamento hierárquico quando é possível constatar que existe um conceito de hierarquia nos dados. Por exemplo, na análise de textos, podemos organizá-los de acordo com um tópico (política, esportes, etc.), neste exemplo podemos observar que é possível granular ainda mais os grupos, ou seja, no grupo de esportes podemos subdividi-los categorias de esportes como futebol, basquete, etc. Portanto, concluímos que é preferível a utilização de agrupamento hierárquico quando os dados em si já apresentam de alguma forma o conceito de hierarquia.

7. Comparando os métodos K-Vizinhos Mais Próximos e K-Means, mostre em quais aspectos eles têm semelhanças e diferenças. (1 ponto)

KNN e K-Means são dois métodos que no processo de análise fazem agrupamento dos dados. Existem algumas semelhanças e diferenças entre estes dois modelos, a seguir trataremos de discuti-las e ver onde eles se assemelham e diferem.

Ambos os modelos é que utilizam distância entre pontos para fazer o agrupamento de objetos que representam os dados do conjunto. Os dois modelos começam com a letra K, porém isto é uma mera coincidência, porque em cada modelo o K tem um propósito diferente, no K-Means o k faz referência ao número de clusters, já no KNN o k representa o número de vizinhos mais próximos.

A primeira diferença entre os dois é que o KNN pertence ao grupo de modelos preditivos supervisionados, já o K-Means ao modelos não-supervisionados. O K-Menas é considerado um algoritmo de clusterização e o KNN é um algoritmo de classificação. O KNN tem uma performance muito melhor quando os dados estão distribuídos em uma mesma escala, porém o K-Means não tem um bom desempenho sob estas circunstâncias.