MBA em Inteligência Artificial e Big Data
Curso 3: Administração de Dados Complexos em Larga Escala

Lista de exercícios

Exercício 1

Considerando o conteúdo visto na primeira videoaula da semana, **instale** o *Apache Spark* em sua máquina de trabalho, e **faça as configurações** necessárias para rodar o *PySpark* em um *Jupyter notebook*. Caso necessário, **busque**, **estude** e **siga** as instruções de um dos diversos tutoriais disponíveis na Web para tal tarefa. Como exemplo, indicam-se os seguintes tutorias:

- **Instalação no Windows:**
  - https://bigdata-madesimple.com/guide-to-install-spark-and-use-pyspark-from-jupyter-in-windows/
- **Instalação no Linux:**
  - https://www.sicara.ai/blog/2017-05-02-get-started-pyspark-jupyter-notebook-3-minutes
- **Instalação no Mac OS:**
  - https://medium.com/designed-by-data/instalando-apache-pyspark-para-funcionar-com-jupyter-notebook-no-macos-42f992c45842
  - https://www.lukaskawerau.com/local-pyspark-jupyter-mac/

Em seguida, **indique** as principais dificuldades encontradas durante o processo, e **descreva** como elas foram solucionadas.

Exercício 2

Considerando o conteúdo visto na segunda videoaula da semana, faça o download de um **novo conjunto de dados** a ser estudado: o *dataset letter*, disponível em:

- https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/letter.scale

Em seguida, **adapte os notebooks** criados em aula para o uso com este novo conjunto de dados.

Adicionalmente, **crie novos notebooks** para processar estes dados utilizando **outro algoritmo** de agrupamento; em específico, o algoritmo **Bisecting kMeans**. Para tanto, tome como base o exemplo de uso deste algoritmo, o qual é disponível em:

- https://spark.apache.org/docs/latest/ml-clustering.html#bisecting-k-means

Por fim, **identifique** e **reporte** o número de grupos ideal, i.e., o valor de k que maximiza a medida *silhouette*, e o correspondente valor de *silhouette* para cada um dos três algoritmos investigados, i.e., *Gaussian Mixture Model*, *kMeans* e *Bisecting kMeans*, neste novo dataset, considerando valores de k entre 2 e 50. Use repetições por meio dos comandos *for* ou *while* de maneira a automatizar o processo.

Exercício 3

Considerando o conteúdo visto na terceira videoaula da semana, **adapte os notebooks** criados em aula para identificar de maneira automatizada os melhores valores de parâmetros, i.e., os valores que maximizam a medida *accuracy*, considerando cada um dos algoritmos de classificação *Multilayer Perceptron* e *One-vs-Rest,* ao processar os dados de treinamento do *dataset iris*. Use repetições por meio dos comandos *for* ou *while* nesta tarefa.

Valores de parâmetros a serem considerados:

- *Multilayer Perceptron*:
  - Layers: [4, 3], [4, 5, 3], [4, 5, 4, 3] e [4, 10, 7, 5, 3]
  - maxIter: 10, 20, 30, 40 e 50
  - blockSize: 64, 128 e 256

- *One-vs-Rest:*
  - *maxIter: 10, 20, 30, 40 e 50*
  - *tol: 1E-5, 1E-6 e 1E-7*
  - *fitIntercept: True e False*

Reporte o melhor valor de *accuracy* identificado nos dados de treinamento e nos de teste com cada algoritmo, junto aos correspondentes valores de parâmetros utilizados.

Adicionalmente, **crie novos notebooks** para processar o *dataset iris* utilizando **outro algoritmo** de classificação; em específico, o algoritmo ***Logistic Regression***. Para tanto, tome como base o exemplo de uso deste algoritmo, o qual é disponível em:

- https://spark.apache.org/docs/latest/ml-classification-regression.html#multinomial-logistic-regression

Por fim, **identifique** os valores de parâmetros ideais para o uso deste novo algoritmo nos dados de treinamento do *dataset iris*, assim como foi feito para os outros algoritmos no primeiro passo deste exercício.

Valores de parâmetros a serem considerados:

- *Logistic Regression*:
  - maxIter: 10, 20, 30, 40 e 50
  - regParam: 0.2, 0.3, 0.4, 0.5, 0.6 e 0.7
  - elasticNetParam: 0.7, 0.75, 0.8 e 0.85

Reporte o melhor valor de *accuracy* identificado nos dados de treinamento e nos de teste com o *Logistic Regression*, junto aos correspondentes valores de parâmetros utilizados.

Exercício 4

Considerando o conteúdo visto nas primeiras videoaulas da semana, **descreva** com suas próprias palavras o que é a abordagem *NoSQL* de manipulação de dados em larga escala. Em seguida, **compare** os tradicionais sistemas relacionais com os demais sistemas *NoSQL*, indicando **um exemplo** de aplicação bem adequada a cada uma destas duas categorias.

Exercício 5

Considerando o conteúdo visto na penúltima videoaula da semana, **baixe** o *Amazon DynamoDB* em sua máquina de trabalho, e **rode** os arquivos `html` dados como exemplo, iniciando-se pela

criação da tabela `Movies`, e o carregamento dos dados a partir do arquivo `json` de exemplo. Em seguida, **estude** os arquivos `html` fornecidos; use-os então como base para **criar** dois novos arquivos `html`, como segue:

- `MoviesUpdateItem.htlm`: arquivo que efetua uma atualização nos dados do filme `The Big New Movie` de modo a adicionar valores de sua escolha aos atributos `info.rating`, `info.plot` e `info.actor`;
- `MoviesConditionalUpdateItem.htlm`: arquivo que efetua uma atualização condicional nos dados do filme `The Big New Movie` de modo remover o primeiro ator listado no atributo `info.actors`, se e somente se existirem ao menos quatro atores cadastrados neste atributo.

Em seguida, **indique** as principais dificuldades encontradas durante o processo, e **descreva** como elas foram solucionadas.

Exercício 6

Considerando o conteúdo visto na última videoaula da semana, **descreva** com suas próprias palavras como a programação funcional tem contribuído para o desenvolvimento de sistemas *NoSQL*, por meio de ferramentas como o *Apache Spark* e o *Apache Hadoop*. Em seguida, **liste** ao menos uma vantagem e uma desvantagem do uso de ferramentas como estas para o processamento de *Big Data*.