Programa de Pós-graduação em Modelagem Computacional e Sistemas

Primeira Aula Prática da Disciplina Mineração de Dados

Professor: Renato Dourado Maia

1 – Objetivo

Esta Aula Prática tem como objetivos: 1) promover uma **reflexão** sobre os conteúdos abordados nas Unidades I – Pré-processamento dos Dados e II – Descrição de Classes e Conceitos (Análise Descritiva); 2) Apresentar o UCI *Machine Learning Repository*; e 3) Instigar a utilização de ferramentas computacionais para análise de dados. É importante destacar que esta atividade deverá ser trabalhada utilizando *Python* (vocês podem utilizar tudo o que estiver incluído no *Python*(*x, y*)[1] e *Orange*[2] e também outras ferramentas, desde que esclareçam como ela pode ser obtida/instalada).

2 – Detalhamento

O repositório da UCI[3] mantém 332[4] bases de dados para a comunidade de aprendizagem de máquina. Todos as bases podem ser visualizadas por intermédio de uma interface de busca. Esclarecimentos podem ser obtidos na página *About*[5] e informações sobre a citação das bases em publicações científicas encontram-se nas políticas de citações (*Citation Policy*[6]). O repositório é de domínio público e aceita doações (ver *Donation Policy*[7]).

2.1 – Acesse o repositório e identifique as 6 bases de dados mais populares.

2.2 – Para cada uma das seis bases mais populares, apresente:

a) As características.
b) O número de instâncias.
c) A quantidade de atributos.
d) As características dos atributos.
e) A existência de valores ausentes

2.3 – Para a base de dados *Íris*:

a) Determine o domínio de cada um dos atributos.
b) Apresente um gráfico com a distribuição de cada uma das variáveis.
c) Analise os gráficos visualmente e verifique se há inconsistências e/ou *outliers*.
d) Calcule a correlação entre os pares atributos e indique o tipo.

1 http://python-xy.github.io/.
2 http://orange.biolab.si/.
3 http://mlr.cs.umass.edu/ml/ (acesso em 02/09/2015).
4 Esse número considera o estado do repositório em 02/09/2015.
5 http://mlr.cs.umass.edu/ml/about.html.
6 http://mlr.cs.umass.edu/ml/citation_policy.html.
7 http://mlr.cs.umass.edu/ml/donation_policy.html.

e) Assuma que uma correlação maior ou igual a 0,6 é alta e que, nesse caso, um dos atributos pode ser eliminado (eliminação de redundância). Há algum par de atributos que permita eliminação de redundância? Se sim, qual(is) atributo(s) você eliminaria?

2.4 – A base de dados *Adult* apresenta diversos valores ausentes. Para os 100 primeiros objetos dessa base, impute os valores ausentes utilizando:

a) Uma constante global.
b) Utilizando a média de cada atributo (no caso de atributos nominais, utilize o subconjunto dominante – moda).
c) Utilizando a média de todos os objetos da mesma classe. No caso de atributos nominais, utilize a moda de cada classe.

2.5 – Para a base de dados *Breast Cancer Wisconsin*, normalize os atributos utilizando os seguintes métodos:

a) Max-Min.
b) Escore-z.
c) Escalonamento decimal.

3 – Algumas Observações

3.1 – O trabalho deve ser feito **em equipes de, no máximo, dois acadêmicos**: a discussão dos problemas e das estratégias de solução com os colegas de outras equipes é **permitida e aconselhável**. Todavia, a concepção e a documentação das soluções deve ser feita **apenas pelos membros da equipe**.

3.3 – A **pesquisa** sobre os conceitos envolvidos no trabalho, bem como a **familiarização** com a ferramenta computacional a ser utilizada, faz parte da Aula Prática.

3.4 – Caso você(s) utilize(m) quaisquer **fontes externas** para elaborar as suas respostas, elas devem ser **citadas**: artigos ou livros, amigos ou colegas, informações encontradas na Internet, **qualquer coisa encontrada em qualquer lugar**! É melhor tentar solucionar os problemas, pois **solucionar problemas é um componente fundamental** para a nossa área de estudo. Não haverá penalidades no caso de utilização de ajuda externa, desde que devidamente citada, e desde que essa ajuda **não seja a cópia** do trabalho de um colega. **Utilizar o trabalho dos outros, como se fosse seu, é plágio, é desonestidade acadêmica.**

3.5 – Uma **documentação escrita** deve ser elaborada utilizando LaTeX, explicando os **objetivos** do trabalho, e apresentando **resultados** e **discussões**. Naturalmente, a documentação deverá apresentar uma **conclusão**, argumentando sobre as **dificuldades**, sobre o **aprendizado**, e sobre a **relevância** do prática no contexto do **curso** e da **disciplina**, e apontando eventuais **sugestões** para melhorias em futuras "edições".

3.7 – Quaisquer **suposições** feitas por você(s) (que não estiverem incluídas na descrição original das questões) **devem também ser bem documentadas**.

4 – Data de Entrega

O relatório da Primeira Aula Prática **deverá ser entregue até o dia 06/09/2015 às 23:59**, **por intermédio do *Moodle***: deverá ser enviado **um único arquivo compactado (7z, zip ou rar)**, contendo os arquivos da **documentação** (envie **todos os arquivos necessários** para a geração da documentação em

pdf, inclusive as **figuras**, sendo que, **caso seja utilizado o TexnicCenter**[8], deve ser enviado o **projeto**), (caso seja utilizado o **Overleaf**[9], o projeto deverá ser **compartilhado comigo** e, naturalmente, **apenas** o arquivo em formato pdf deverá ser enviado), bem como **todos os arquivos das implementações utilizadas**. O arquivo compactado **deve ter** o seguinte nome:

MD_API_Nome1_Nome2

(MD_API_JaspionJoseDaSilva_BlackKamenRaiderDaSilva, por exemplo, **sem espaços e/ou caracteres especiais**). **CASO O TAMANHO DO ARQUIVO SUPERE 1 MEGABYTE, TENTE ENVIAR MAIS DE UM ARQUIVO, COM CADA UM, NATURALMENTE, SENDO DE TAMANHO MENOR OU IGUAL A UM MEGABYTE.**

8 http://www.texniccenter.org/
9 https://www.overleaf.com/