UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
INSTITUTO METROPOLE DIGITAL

APRENDIZADO DE MÁQUINA
IMD0104 – 2018.2

O principal objetivo deste trabalho prático é validar os conceitos de Aprendizado de máquina supervisionado e não supervisionado abordados durante a disciplina.

1. Metodologia do Trabalho:

A base de dados deve ser uma escolha de cada grupo de alunos (duplas ou trios), podendo ser oriunda do próprio ambiente de trabalho, ou retirada de algum repositório aberto[1] na web, observando-se, no entanto, um número mínimo de 500 instâncias e 8 atributos.

Uma vez escolhida a base de dados, os grupos devem seguir as três etapas do processo de aprendizado de máquina, que são: pré-processamento, utilização dos métodos de aprendizado e pós-processamento. Com relação aos métodos, serão utilizados os algoritmos supervisionados (preditivos) no contexto de classificação e os não supervisionados (descritivos) no contexto de agrupamento ou *Clustering*.

2. Pré-processamento:

Nessa etapa, é importante organizar os dados de maneira que a base de dados possa ser utilizada de forma satisfatória pelos métodos de aprendizado de máquina escolhidos. Para tal, três ações se fazem necessárias:

1. Retirar os dados faltosos;
2. Suavizar os dados ruidosos ou aberrantes;
3. Colocar na mesma escala os atributos numéricos.

3. Métodos não Supervisionados:

Para os grupos que optarem por métodos descritivos (agrupamento) devido as características da base de dados, aconselha-se a utilização dos métodos abordados na disciplina (*k-Means* e Hierárquico). Além desses dois, outros métodos também poderão ser utilizados. Observando sempre o número mínimo de dois métodos.

Para alguns métodos, o valor de *k* deverá ser informado a priori. Quando esse for o caso, tal valor deverá variar entre 2 e $\boldsymbol{log_2(n)}$, onde *n* é o número de instâncias da base.

---

[1] https://archive.ics.uci.edu/ml/datasets.html

Caso o método *k-Means* seja escolhido entre os métodos descritivos, para cada valor de *k*, serão feitas 3 execuções variando-se o valor da *seed* (semente) em 37, 110 e 777. Após rodar os experimentos, salve cada um deles em arquivos individuais, especificando-se o nome do método, o valor de *k* e da semente (*s*). Por exemplo, para um experimento com *k-Means*, onde *k* = 3 e *seed* = 37, salve o arquivo com o seguinte nome: **kMeans_3k_s37**. Podendo a extensão do mesmo ser tanto arff ou csv. Note, que o valor máximo de *k* deve ser arredondado para cima. Por exemplo, para uma base com 500 instâncias, $k = log_2 (500) => 2,69 =>$ k = 3.

Caso o método Hierárquico seja escolhido entre os métodos descritivos, por ser determinístico, é necessário apenas uma execução para cada valor de *k*. Por exemplo, para um experimento com hierárquico, onde *k* = 3, salve o arquivo com o seguinte nome: **Hira_3k.**

4. Métodos Supervisionados:

Para os grupos que optarem por métodos preditivos, aconselha-se a utilização de pelo menos quatro métodos, dentre os seis abordados na disciplina, sendo eles: Redes Neurais (MultiLayerPerceptron), k-NN (IBk), Árvores de decisão (J48), Naive Bayes, Bagging e Boosting.

Cada método supervisionado deverá ser treinado e testado com as seguintes abordagens: *10-fold cross validation* e percentage *split* (70/30, 60/40 e 50/50). Dessa forma, para uma mesma base, a matriz resultante dos experimentos dos classificadores utilizados será: M[4,4], onde as linhas representarão as metodologias de treinamento e teste, e as colunas representarão os próprios classificadores escolhidos.

Caso o grupo esteja utilizando métodos descritivos para criar as classes da base de dados, este deverá utilizar as três melhores partições geradas pelos métodos de agrupamento e validadas pelos índices de validação. Nesse caso, a matriz resultante dos experimentos dos classificadores será: M[12,4], onde as linhas representarão as quatro metodologias de treinamento/teste utilizadas para cada uma das três partições.

5. Pós-processamento:

Durante essa fase, será necessário validar os resultados obtidos pelos métodos descritivos e preditivos. Para os descritivos (agrupamento), os resultados (as partições) deverão ser analisados através dos índices de validação (DB e Silhueta). Os resultados encontrados, para cada um dos índices, deverão ser expressos em forma de gráfico. O teste estatístico (Friedman) deverá ser aplicado para garantir a diferença estatística entre os resultados de cada índice.

Para os preditivos (classificação), a análise comparativa será feita através da acurácia (erro médio absoluto) de cada classificador. O teste estatístico (Friedman) deverá ser aplicado para garantir a diferença estatística entre os métodos preditivos.

5.1 Teste de Friedman:

É muito importante a formatação da matriz de resultados sobre a qual será aplicado o teste. Dessa forma, para os algoritmos descritivos, a matriz de resultados será composta por *n* linhas e quatro (4) colunas, onde as colunas representarão os algoritmos (1 hierárquico e 3 *k-Means*), e *n* representará os possíveis valores de *k*.

Já para algoritmos os preditivos, a matriz de resultados será composta por quatro linhas e quatro colunas, para os casos onde a base de dados já possui classes, ou por doze linhas e quatro colunas, nos casos onde a fase agrupamento foi utilizada para gerar as classes.

Anexo A

Modelo de relatório

1. Seção 1--- Introdução

Qual é o problema? Por que ele é relevante? Como será resolvido?

2. Seção 2 ---- Descrição do problema e Base de Dados

- Descrição da base de dados
- Pré-processamento

3. Seção 3 ---- Experimentos (Explicar a metodologia dos experimentos).

4. Seção 4 ---- Conclusão sobre os experimentos

5. ---- Referências bibliográficas