Universidade do Minho
Departamento de Informática

FE03

Curso: Mestrado Integrado em Informática – Engenharia do Conhecimento
U.C.: Descoberta do conhecimento

Folha de Exercícios FE03	
Docente	Cristiana Neto
Tema	Explorar o Weka
Turma	PL
Ano Letivo	2019-20 – 2º Semestre
Duração da aula	2 horas

1. Enunciado

[1] Qual ou quais as diferenças entre uma base de dados, um datawarehouse e um dataset?

[2] Quais são algumas das limitações do data mining e como podem ser ultrapassadas?

[3] Qual a diferença entre datawarehouse e data mart?

[4] Indique alguns constrangimentos éticos da utilização e aplicação do Data Mining.

[5] O que é a normalização de bases de dados e quais os impactos em sistemas OLTP e OLAP?

[6] Desenhe uma base de dados relacional com 3 tabelas. Garanta que cria o número de colunas e as colunas adequadas para estabelecer relações entre as tabelas.

[7] Desenhe uma tabela datawarehouse com algumas colunas que normalmente seriam normalizadas. Explique porque faz sentido desnormalizar nesta situação.

[8] Faça uma pesquisa online e encontre 3 sites que contenham informação que pode ser aplicada ao processo de Data Mining.

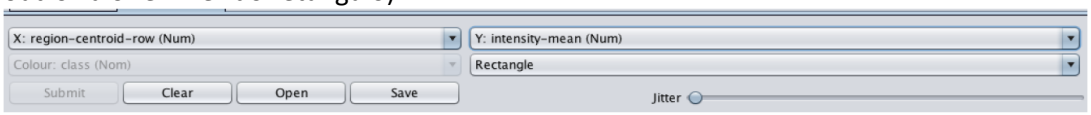
[9] Faça uma pequena pesquisa online e descubra um data set disponível para download. Descreva sucintamente o data set (conteúdo, propósito, tamanho, antiguidade).

No ecrã inicial do Weka abrir o "package manager" Tools -> Package Manager. Instalar o package "*UserClassifier 1.0.3*". Após este passo responder às seguintes questões seguintes.

[10] Abrir o Weka / Explorer e carregar o data set "*segment-challenge.arff*". No separador Classify definir conjunto de dados "*segmento-test.arff*" como conjunto de teste.

[a] Utilizar o trees -> UserClassifier;
Clicar em Start;

Selecionar o separador Data Visualizer; e selecionar as seguintes opções (pode ser utilizado outro valor em vez do retângulo):



Ir selecionando os grupos possíveis de definir.
Determinar o resultado da classificação.

[b] Comparar os resultados obtidos com este método de criação de árvore de decisão com os resultados do algoritmo J48.

[11] Abrir o Weka/Explorer e carregar o data set "*segment-challenge.arff*". Com este data set carregado responda às seguintes questões:

[a] Usar o algoritmo J48 como classificador; Usar o data set "*segment-test.arff*" como conjunto teste. Qual o valor da classificação?

[b] Usando a opção "*Use training set*" determine o valor da classificação. Porque não deve ser usada esta opção para determinar a qualidade e a aplicabilidade dos algoritmos aos dados?

[c] Escolha o J48 como classificador e vá alterando as percentagens de divisão ("*Percentage Split*") dos grupos de treino e de teste em: 10%, 20%, 40%, 60% e 80%. O que observa?

[d] Repetir a questão anterior usando 90%, 95%, 98% e 99%. O que acontece ao número de instâncias corretamente classificadas? E o que acontece à percentagem de instâncias corretamente classificadas? Explicar esta variação.

[e] Apesar de com uma percentagem de 98% para o treino e 2% para o teste dar uma classificação de 100% isto quer dizer que o modelo construído é o mais indicado para o problema apresentado?

[f] Com base nas experiências acima, qual considera a melhor estimativa da verdadeira precisão de J48 para este data set?

[13] Abrir o Weka/Explorer e carregar o data set "*diabetes.arff*". Com este data set carregado responda às seguintes questões:

[a] Selecionando "*Percentage Split*" a 80% quantas instâncias serão usadas para treino e quantas serão usadas para teste? (O Weka arredonda ao número inteiro mais próximo).

[b] Mudando o "*Random seed*" entre 1,2,3,4 e 5, mantendo o "*Percentage Split*" a 80% indique o valor mínimo e máximo de instâncias incorretamente classificadas.

[c] Qual a média da percentagem de instâncias corretamente classificadas?

[d] Se repetisse o exercício [13/b] com 10 "*random seed*" em vez de 5 qual seria o efeito na média?

COMPARAR COM "BASE LINE"

[14] Abrir o Weka/Explorer e carregar o data set "*iris.arff*". Com este data set carregado responda às seguintes questões:

[a] Este data set caracteriza 3 classes com 50 instâncias cada uma. Qual será a percentagem de acerto do algoritmo ZeroR quando aplicado ao training set?

[b] Qual é o resultado da classificação base line quando é usado o método "*Percentage Split*" em 66%?

[15] Abrir o Weka/Explorer e carregar o data set "*glass.arff*". Com este data set carregado responda às seguintes questões:

[a] Qual é a percentagem de acerto do algoritmo ZeroR com 66% de "*Percentage Split*"?

[b] Qual o valor usando o J48 e os restantes parâmetros por defeito?

[c] Qual a precisão (accuracy) do algoritmo NaiveBayes' usando os parâmetros por defeito?

[16] Abrir o Weka/Explorer e carregar o data set "*segment-challenge.arff*". Utilize o data set "*segment-test.arff*" para dataset de avaliação (teste). Com estes data sets carregados responda às seguintes questões:

[a] Qual a precisão do algoritmo ZeroR?

[b] Qual a precisão do algoritmo IBk's, com todos os parâmetros por defeito?

[c] Qual a precisão do algoritmo PART, com todos os parâmetros por defeito?