AP532 – PREPARAÇÃO DE DADOS PARA MINERAÇÃO DE DADOS

Primeira Lista de Exercícios

1. Sobre mineração de dados, responda as perguntas a seguir:

   a) Qual é a diferença entre tarefa e técnica de mineração de dados?

   b) Dê um exemplo para **dado**, **informação** e **conhecimento**.

   c) Qual é a diferença entre tarefas descritivas e preditivas?

2. Carregue o arquivo **cpu.csv** no Weka. Selecione um atributo do dataset, com exceção da classe, e depois elimine 10 instâncias, de forma aleatória, desse atributo. Para esse atributo:

   a) Preencha seus valores faltantes na tabela abaixo:

# instância	Valor Real	Média	kNN (k = 1)	Regressão	Redes Neurais

   b) Qual foi o método mais eficiente para preencher os valores faltantes? Por quê?

3. Suponha que você está analisando os seguintes dados para o atributo **idade**: 13, 15, 33, 16, 19, 20, 35, 21, 22, 22, 25, 70, 25, 35, 30, 33, 16, 35, 35, 20, 25, 36, 40, 45, 46, 52, 25, 74, 87,92.

   a) Use o método de particionamento baseado na distância (**Equi-width**) para eliminar possíveis ruídos dos dados.

   b) Use o método de particionamento baseado na freqüência (**Equi-depth**) para eliminar possíveis ruídos dos dados.

   c) Determine o sumário dos cinco números (min, max, mediana, Q1, Q3).

   d) Quais são os outliers nesses dados?

   e) Esboce o histograma de frequência.

4. Em que situação, na etapa de preparação de dados, você usaria a análise de correlação seguida da análise de regressão linear? Explique.

5. Sobre técnicas de amostragem, responda as perguntas a seguir:

   a) Aponte duas aplicações dessas técnicas na etapa da preparação de dados.

   b) Qual é a diferença entre os filtros de amostragem (**Resample**) para os métodos **supervisionado** e **não-supervisionado**?

   c) Que algoritmo de clusterização você usaria para realizar uma amostragem estratificada? Explique sua solução.

6. Clusterização (agrupamento) tem uma importância fundamental em projetos de mineração de dados. Responda as questões a seguir sobre clusterização:

   a) Por que, em geral, os atributos de um dataset são normalizados antes de uma clusterização?

   b) Considerando os métodos para normalização Min-Max e Z-score, qual deles é mais atrativo para clusterização? Por quê?

   c) Uma aplicação em que clusterização é a tarefa principal de mineração de dados.

   d) Uma aplicação em que clusterização é usada na fase de pré-processamento.

   e) Qual seria o melhor algoritmo de clusterização (**k-means** ou **DBScan**) para uma atividade de sintetização de dados, em que os dados serão publicados de forma agregada?

7. Discuta quais das atividades a seguir são tarefas de mineração de dados:

   a) Dividir os clientes de uma empresa de acordo com o seu sexo.

   b) Dividir os clientes de uma empresa de acordo com sua lucratividade.

   c) Cálculo do total de vendas de uma empresa.

   d) Ordenar um banco de dados de alunos baseado no número de identificação dos alunos.

   e) Prever o resultado de uma joga de um par de dados.

   f) Prever o futuro preço das ações de uma empresa usando registros históricos.

   g) Monitorar a taxa de batimentos cardíacos de um paciente procurando por anormalidades.

8. Suponha que você seja empregado como consultor de mineração de dados de uma empresa de vendas de produtos agropecuários. Descreva como a mineração de dados pode ajudar a empresa dando exemplos específicos de como as técnicas de clusterização, classificação e regras de associação podem ser aplicadas.

9. Classifique os seguintes atributos como binários, discretos ou contínuos. Classifique-os também como qualitativos (nominais ou ordinais) ou quantitativos (intervalares ou de faixa). Alguns casos podem ter mais de uma interpretação, então indique brevemente seu raciocínio se achar que possa haver alguma ambiguidade.

   Exemplo: Idade em anos: Discreta, quantitativa, de faixa.

   a) Horários em termos de AM ou PM.
   b) Brilho conforme medido pelo medidor de luz.
   c) Brilho conforme medido pelo julgamento das pessoas.
   d) Medalhas de bronze, prata e ouro, conforme dadas nas olimpíadas.
   e) Altura cima do nível do mar.
   f) Número de pacientes em um hospital.
   g) Números INBN para livros.
   h) Posto militar.
   i) Distância da Feagri ao centro do campus da Unicamp.
   j) Densidade de uma substância em gramas por centímetro cúbico.

10. Os métodos de particionamento de dados **Equi-width** e **Equi-depth** são utilizados para suavizar ruídos em atributos. Sobre esses métodos responda:

    a) No método **Equi-width**, se *A* e *B* são os valores mínimo e máximo de um atributo, a largura (**distância**) dos intervalos será: $W = (B-A)/N$. Explique porque os **outliers** podem dominar o número de intervalos.

    b) O método **Equi-depth** divide os valores de um atributo em *N* intervalos, cada um contendo aproximadamente o mesmo número de amostras. Embora esse método apresente bons resultados para atributos numéricos, explique porque lidar com atributos nominais pode ser complicado.

11. Um laboratorista quer usar a tarefa de associação para analisar os resultados de testes. Cada teste consiste de 50 questões, com quatro respostas em cada uma.

    a) Como você converteria esses dados para uma forma apropriada para análise de associação?

    b) Em especial, que tipos de atributos você teria ? Apresente uma sugestão.

12. Quais das seguintes quantidades provavelmente mostrarão mais autocorrelação temporal: a quantidade de chuva diária ou a temperatura diária? Por quê?

13. Os atributos a seguir são medidos para membros de uma manada de elefantes asiáticos: peso, altura, comprimento da presa, comprimento da tromba e área do ouvido. Baseado nessas medidas, que tipo de medida de semelhança você usaria para comparar ou agrupar esses elefantes?

14. Mostre que regras originadas de um mesmo conjunto frequente têm o mesmo **suporte**, mas diferentes valores para **confiança**.

15. Utilizando o conjunto de dados climáticos Clima_Taubate_SP.xls:

    a) Crie um atributo que represente a **amplitude de temperatura** (diferença entre Temperatura máxima e Temperatura Mínima.

    b) Crie um atributo que represente a temperatura média diária.

    c) Crie três atributos que representem, para cada dia, as precipitações um dia antes, dois dias antes e três dias antes, respectivamente.

    Exemplo:

Data	**Chuva**	**TMax**	**TMin**	**Chuva_1**	**Chuva_2**	**Chuva_3**
17/05/19XX	17	25	20	-	-	-
18/05/19XX	23	23	18	17	-	-
19/05/19XX	20	27	22	23	17	-
20/05/19XX	12	25	20	20	23	17
21/05/19XX	10	28	20	12	20	23

    d) Crie um atributo que represente a variação da temperatura média da véspera para o dia corrente.

    Exemplo:

Data	**Chuva**	**TMax**	**TMin**	**TMedia**	**DeltaTMedia_1**
17/05/19XX	17	25	20	22.5	-
18/05/19XX	23	23	18	20.5	-2.0
19/05/19XX	20	27	22	24.5	4.0
20/05/19XX	12	25	20	22.5	-2.0
21/05/19XX	10	28	20	24.0	1.5

    e) Crie um atributo binário (0/1) cujos valores representem 1 quando tiver chovido acima de 5 mm e 0, caso contrário.

        i) No Banco de Dados utilizado, quantos dias choveram acima de 5 mm ?

        ii) Faça o mesmo para 10 mm. Quantos dias choveram acima de 10 mm ?

f) Considerando o atributo do item (a) normalizado (Min = 0, Max = 1), qual a média e o desvio padrão ?

g) Considerando o atributo do item (b) normalizado Z-Score (Média = 0, Desvio Padrão = 1), qual o valor mínimo e o valor máximo ?

16. Para o arquivo trabalhado no item 1, considere a aba “Exec2”.

    a) Se a Coluna E da Planilha (Ocorrência de A) for o atributo Meta (desconsiderar a Coluna F), identifique um padrão de ocorrência.

       DICA: O padrão de ocorrência (S) está relacionado ao número de dias seguidos com alta Amplitude Térmica.

    b) Se a Coluna F da Planilha (Ocorrência de B) for o atributo Meta (desconsiderar a Coluna E), identifique um padrão de ocorrência.

       DICA: O padrão de ocorrência (S) está relacionado ao número de dias seguidos de chuva.

17. Procure encontrar algum padrão (considerando TMax, TMin e DeltaT) nos 5 dias anteriores aos dias em que tenha chovido mais do que 10 mm.

18. Calcule a entropia e o ganho de informação para cada atributo do conjunto de exemplos de treinamento abaixo.

ID Cliente	Sexo	Tipo de Carro	Tamanho da Camisa	Classe
1	M	Familiar	Pequeno	C0
2	M	Esportivo	Médio	C0
3	M	Esportivo	Médio	C0
4	M	Esportivo	Grande	C0
5	M	Esportivo	Extra Grande	C0
6	M	Esportivo	Grande	C0
7	F	Esportivo	Pequeno	C0
8	F	Esportivo	Pequeno	C0
9	F	Esportivo	Médio	C0
10	F	Luxo	Grande	C0
11	M	Familiar	Grande	C1
12	M	Familiar	Grande	C1
13	M	Familiar	Médio	C1
14	M	Luxo	Grande	C1
15	F	Luxo	Pequeno	C1
16	F	Luxo	Pequeno	C1
17	F	Luxo	Médio	C1
18	F	Luxo	Médio	C1
19	F	Luxo	Médio	C1
20	F	Luxo	Grande	C1

19. Qual é a diferença básica entre pré-poda e pós-poda (árvore de decisão)?

20. Dado o BD de transações, abaixo, calcule:

TID	Lista de Itens
10	{queijo, torrada, suco}
20	{café, pão, pizza}
30	(queijo, vinho, café}
40	{torrada, café, vinho, queijo}
50	{pizza, queijo, pão, vinho}

a) O suporte e a confiança da regra {pizza, pão } $\Rightarrow$ {queijo, vinho}

b) Se considerarmos o suporte $\geq$ 30%, qual é o conjunto de dois itens com maior freqüência?

c) Se considerarmos o suporte $\geq$ 30%, qual é o conjunto de três itens com maior freqüência?

21. Dados os pontos P = (-1, 3, -2); Q = (-4, 5, -2); R = (4, -1, 0) e S = (7, 0, 1), pede-se:

a) O centróide dos pontos P, Q, R, S.

b) As distâncias Euclideana e de Manhattan entre os pontos PQ, RS e QS.

22. Dado o conjunto de dados abaixo, use o coeficiente de Jaccard (para variáveis assimétricas) e identifique quais são os dois pacientes que apresentam a maior similaridade:

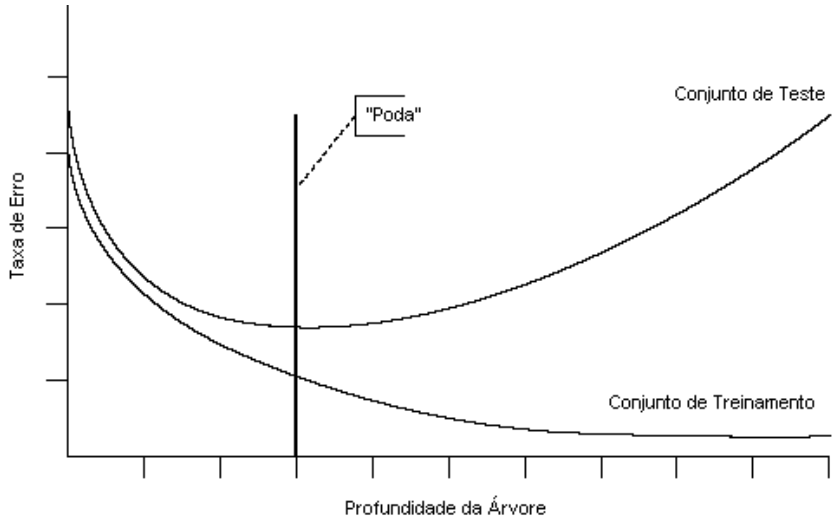
Paciente	Febre	Tosse	Teste1	Teste2	Teste3	Teste4
Pedro	P	N	P	N	N	N
Patrícia	N	P	P	N	P	P
Amélia	P	N	P	N	P	P
Gustavo	P	P	N	N	N	N

23. Faça uma comparação entre os métodos de clusterização: particionamento, hierárquico e baseado em densidade. Quais as vantagens e desvantagens de cada método?

24. Que algoritmo de clusterização você usaria para segmentar os clientes das Casas Bahia? Explique a sua decisão (imagine uma base de dados com mais de dez milhões de clientes).

25. Para a tarefa de Classificação, uma das técnicas mais utilizadas é a 'Árvore de Decisão'. Existem vários algoritmos possíveis de utilização para a construção de árvores de decisão e, entre eles, o C4.5, desenvolvido por Quinlan (1986).

Este algoritmo utiliza a ‘Entropia’ como medida para escolher o melhor atributo ‘split’ a cada fase, buscando encontrar conjuntos mais homogêneos com relação ao atributo definido como meta (Classe).



Um dos problemas que este algoritmo pode trazer é o ‘Overfitting’. Considere a figura acima e discuta sobre este problema, indicando o que é possível ser feito para evitá-lo.

26. Dê exemplo, no domínio agrícola, de uma aplicação de séries temporais curtas que pode ser transformada em um problema de clusterização e diga que algoritmo você usaria para essa aplicação.