Universidade Federal de São Carlos
Aprendizado de Máquina 1 - 01/2022
Prof. Murilo Naldi

Lista de exercícios 1 – Dados

Adaptados do livro Introduction data mining / c2006 - (Livros) PANG-NING, Tan; STEINBACH, Michael; KUMAR, Vipin. Introduction data mining. Boston: Pearson Education, c2006. 769 p. ISBN 0-321-32136-7

1. No exemplo inicial de ilustração do problema da Aula 02, a estatística diz “Os campos 2 e 3 são basicamente o mesmo”. A partir das três linhas apresentas, podemos afirmar que ela está certa? Porque?

2. 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) Ângulos conforme medidos em graus entre 0 e 360.
(e) Medalhas de bronze, prata e ouro conforme dadas nas Olimpíadas,
(t) Altura acima do nível do mar.
(g) Número de pacientes em um hospital.
(h) Números ISB para livros (Veja o formato na Web.).
(i) Habilidade de passar luz em termos dos seguintes valores: opaco, translúcido, transparente.
(j) Posto militar.
(k) Distância do centro do campus.
(1) Densidade de uma substância em gramas por centímetro cúbico.
(m) Número da mesa em um restaurante.

3. Você é abordado pelo diretor de marketing de uma empresa local, o qual acredita que imaginou uma forma segura de medir a satisfação do cliente. Ele explica seu esquema da seguinte maneira: “É tão simples que não consigo acreditar que ninguém tenha pensado nisso antes. Eu simplesmente registro o número de reclamações de clientes para cada produto. Eu li em um livro sobre mineração de dados que conta que atributos de reclamação e satisfação são proporcionais e, assim, minha medida de satisfação com o produto deve ser um atributo proporcional. Contudo, quando classifiquei os produtos com base na minha nova taxa de satisfação dos clientes e mostrei-os ao meu chefe, ele me disse que eu tinha deixado de lado o obvio e que minha medida não tinha valor. Ele ficou irritado porque nosso produto que mais

vendia tinha a pior taxa de satisfação, pois possuía a maior parte das reclamações.”
(a) Quem está certo, o diretor de marketing ou seu chefe? Se você respondeu “seu chefe”, o que você faria para consertar a medida de satisfação?
(b) O que você usaria como atributo de satisfação do produto?'

4. Alguns meses depois, você é abordado novamente pelo mesmo diretor de marketing do Exercicio 3. Desta vez, ele imaginou uma abordagem melhor para medir o quanto um cliente prefere um produto em comparação a outro produto semelhante. Ele explica: “Quando desenvolvemos novos produtos, geralmente criamos diversas varições e avaliamos qual os clientes preferem. Nosso procedimento padrão é dar a quem esta testando nosso produto todas as variações deste ao mesmo tempo e solicitar que classifiquem essas variações do produto em ordem de preferência. Entretanto, as pessoas que estão testando os produtos ficam muito indecisas, especialmente quando ha mais de dois produtos. Como consequência, os testes demoram demais. Eu sugeri que executássemos as comparações em pares e entao usássemos essas comparações para criar as classificadores. Assim, se tivermos três variações do produto, fazemos os consumidores compararem as variações 1 e 2, depois a 2 com a 3 e, finalmente, a 3 com a 1. Nosso tempo de teste com meu novo procedimento é um terço do que era com o procedimento antigo, mas os funcionários conduzindo os testes reclamam que não conseguem obter uma classificação consistente a partir dos resultados. Meu chefe quer as avaliações do produto mais recente, para ontem. Eu deveria mencionar também que foi ele a pessoa que imaginou a abordagem antiga de avaliação de produtos. Você pode me ajudar?”

(a) O diretor de marketing está em apuros? A abordagem dele funcionara para uma classificação ordinal das variações do produto em termos de preferência dos clientes? Explique.

(b) Existe alguma forma de consertar a abordagem do diretor de marketing? De maneira mais geral, o que você pode dizer sobre a tentativa de criação de uma escala de medição ordinal baseada em comparações de pares?

(c) Para o esquema original de avaliação de produtos, as classificações gerais de cada variação do produto são encontradas através do calculo da sua média entre todas as pessoas executando o teste. Comente se você acha que esta é uma abordagem razoável. Quais outras abordagens você poderia usar?

5. Você consegue imaginar uma situação na qual números únicos identificadores seriam úteis para previsões?

6. Um psicólogo quer usar a análise de associativo para analisar os resultados dos testes. O teste consiste de 100 questões com quatro respostas possíveis em cada uma.

(a) Como você converteria estes dados para a análise de associação?
(b) Em especial, que tipo de atributos você teria e quantos deles existem?

7. Quais das seguintes quantidades provavelmente mostrarão mais auto-correlação temporal: a quantidade de chuva diária ou a temperatura diárias? Porque?

8. Muitas ciências se baseiam na observação em vez de (ou além de) experimentos projetados. Compare problemas relacionados a qualidade de dados envolvidos na ciência observacional aos dos dados da ciência experimental e mineração de dados.

9. Faça a distinção entre ruído e externos (*outliers*). Assegure-se de considerar as seguintes questões:

(a) Os ruídos são interessantes ou desejáveis em alguma situação? E os externos?

(b) Os objetos de ruídos podem ser externos?

(c) Os objetos de ruídos são sempre externos?

(d) Os fatores externos são sempre objetos de ruído?

(e) Os ruídos podem transformar um valor típico em um incomum ou vice-versa?

10. Você recebe um conjunto de m objetos que está dividido em k classes, onde a classe de índice i possui tamanho m_i. Se o objetivo for a obtenção de uma amostra de tamanho $n < m$, qual a diferença entre os dois esquemas de amostragem a seguir? (Suponha amostragem com substituição.)

(a) Selecionamos aleatoriamente $n*m/m$ elementos de cada classe.
(b) Selecionamos aleatoriamente n elementos do conjunto de dados, sem considerar a classe a qual um objeto pertence.

11. Este exercício compara e diferencia algumas medidas de semelhança e distância.

(a) Para dados binários, a distância L1 corresponde a distância de Hamming, ou seja, o número de bits que são diferentes entre dois vetores binários. A semelhança de Jaccard é uma medida de semelhança entre dois vetores binários. Calcule a distancia de Hamming e a semelhança de Jaccard entre os dois vetores binários a seguir:

x =0101010001
y =0100011000

(b) Em qual abordagem a disténcia de Hamming ou de Jaccard é mais semelhante ao Coeficiente de Correspondência Simples e em qual é abordagem é mais semelhante a medida do Co-seno? Explique.

(c) Suponha que esteja comparando o quão semelhantes dois organismos de

diferentes espécies são em termos do número de genes que compartilham. Descreva qual medida, a de Hamming ou a de Jaccard, você acha que seria mais apropriada para comparar a constituição dos dois organismos. Explique (Suponha que cada animal seja representado na forma de um vetor binário, onde cada atributo seja 1 se um determinado gene esteja presente no organismo e 0 caso contrário).

(d) Se você quisesse comparar a constituição genética de dois organismos da mesma espécie, e.g., dois seres humanos, você usaria a distancia de Hamming, o coeficiente de Jaccard ou uma medida diferente de semelhança ou diferença? Explique. (Observe que dois seres humanos compartilham > 99,9% dos mesmos genes.)

12. Considere o problema de se encontrar os *k* vizinhos mais próximos de um objeto de dados. Um programador projeta o algoritmo a seguir para essa tarefa.

Algoritmo para encontrar os *k* vizinhos mais próximos.

1: para i = 1 até número de objetos de dados faça
2: encontre as distâncias do objeto de índice *i* até todos os outros objetos.
3: ordene estas distâncias em ordem decrescente. (Registre qual objeto está associado a cada distância.)
4: retome os objetos associados com as primeiras *k* distancias da lista ordenada
5: fim do para

(a) Descreva os potenciais problemas desse algoritmo se houver objetos duplicados no conjunto de dados.
(b) Como você resolveria esse problema?

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

14. Para os vetores a seguir, x e y, calcule as medidas indicadas.
(a) x=(1919131)a y =(2,2,2,2) co-seno, correlação, Euclidiana
(b) x =(0,1,0,1), y =(1,0,1,0) co-seno, correlação, Euclidiana, Jaccard
(c) x =(0,-1,0,1), y =(1,0,-1,0) co-seno, correlação, Euclidiana
(d) x =(1,1,0,1,0,1), y =(1,1,1,0,0,1) co-seno, correlação, Jaccard
(e) x =(2,-1,0,2,0,-3), y =(-1,1,-1,0,0,-1) co-seno, correlação

15. Dada um medida de distância no intervalo de valor [0,n] para um real qualquer, como converter esses valores para um intervalo [0,1] ?

16. A proximidade geralmente é definida entre um par de objetos.

(a) Defina duas formas através das quais você poderia estabelecer a proximidade entre um conjunto de objetos.
(b) Como você poderia definir a distância entre dois conjuntos de objetos em espaço Euclidiano?
(c) Como você poderia definir a proximidade entre dois conjuntos de conjuntos objetos de dados?

17. Considere o seguinte conjunto de dados sobre clientes de uma determinada empresa a seguir e responda:

ID	Consumo	Sexo	Frequência	Classe
1	10000	M	3	B
2	1200	M	8	C
3	300	F	7	D
4	2100	M	6	C
5	11000	M	Faltante	A
6	240	M	Faltante	D
7	320	F	8	D

(a) Sabendo que uma técnica de classificação será aplicada neste conjunto de dados e que esta técnica não trabalha com atributos faltantes, como você faria o pré-processamento dos dados?

(b) Caso o conjunto de dados fosse pequeno (apenas os 7 objetos que você vê) e você não pudesse descartar objetos, o que você faria?

(c) Se a técnica de classificação só trabalhar com atributos reais, o que você deve fazer?

(d) E se a técnica de classificação for sensível a diferentes intervalos de valores de atributos, qual seria a última coisa a ser feita antes da aplicação da técnica de mineração propriamente dita?

18. Explique como uma grande quantidade de atributos pode influenciar no processo de mineração de dados? E se for uma grande quantidade de objetos?

19. Para que serve e quais as principais diferenças entre as técnicas de seleção de atributos e redução de atributos? Em quais cenários cada uma é mais vantajosa?