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

Lista de exercícios 2 – Explorando 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. Obtenha um dos conjuntos de dados disponíveis no UCI Machine Learning Repository e aplique tantas técnicas diferentes de visualização descritas em aula quanto for possível. Utilize ferramentas de visualização para isso (Weka, R Project, etc.)

2. Identifique pelo menos duas vantagens e duas desvantagens do uso de cores para representar visualmente informações.

3. Qual são as questões relacionadas à organização que surgem quanto a gráficos tridimensionais?

4. Discuta as vantagens e desvantagens de usar amostras para reduzir o número de objetos de dados que precisam ser exibidos. A simples amostragem aleatória (sem substituição) seria uma boa abordagem para a amostragem? Por quê ou por que não?

5. Descreva como você criaria visualizações para exibir informações que descrevam os seguintes tipos de sistemas:

(a) Redes de computadores. Assegure-se de incluir tanto os aspectos estáticos, como a conectividade, quanto os dinâmicos, como o tráfego.

(b) Distribuição de determinadas espécies de plantas e animais pelo mundo para um determinado momento no tempo.

(c) O uso de recursos computacionais, como tempo de processador, memória principal e disco para um conjunto de programas de bancos de dados.

(d) A mudança na ocupação de trabalhadores em um determinado país nos últimos trinta anos. Suponha que você tenha informações anuais sobre cada pessoa que também incluam sexo e nível de escolaridade.

Assegure-se de abordar as seguintes questões:

- Representação. Como você mapeará objetos, atributos e relacionamentos para

elementos visuais?

- Organização. Existem considerações especiais que precisam ser levadas em conta no que diz respeito a como os elementos visuais são exibidos? Exemplos específicos poderiam ser a escolha de um ponto de vista, o uso de transparência ou a separação de determinados grupos de objetos.
- Seleção. Como você lidará com um número grande de atributos e objetos de dados?

6. Descreva uma vantagem e uma desvantagem de um gráfico de caule-e-folhas quanto a um histograma padrão.

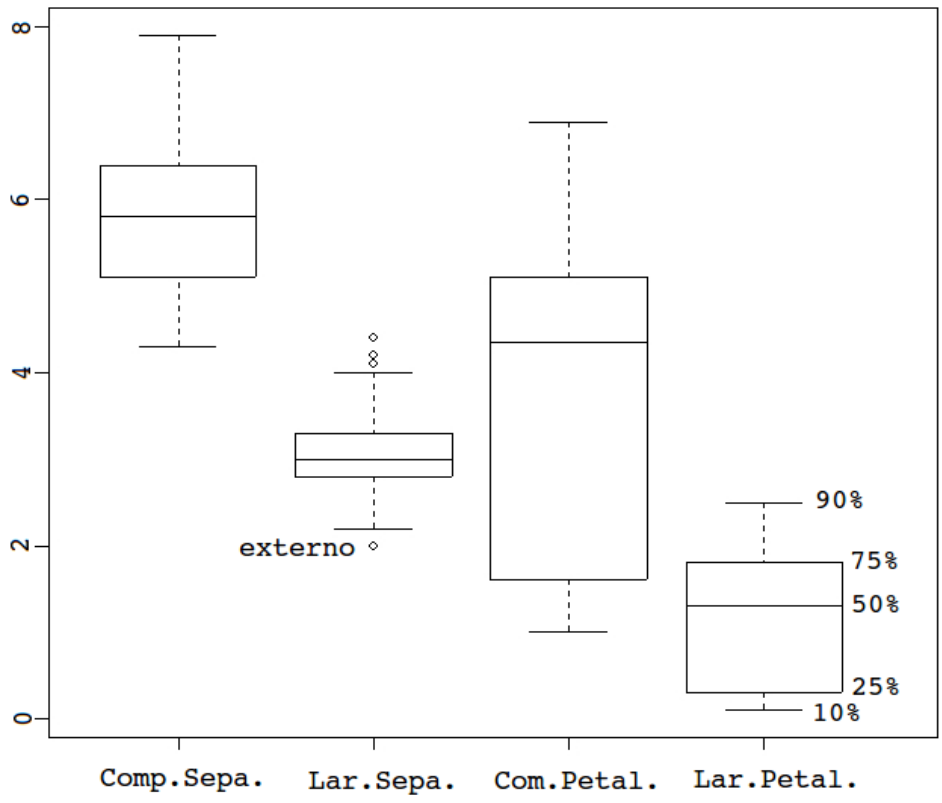
7. Considere o seguinte gráfico de caule-e-folhas construído a partir do atributo idade de um conjunto de pacientes :

```
3:22234556778999
4:1112345556
5:11112789
6:57
```

(a) Qual é a década em que a maioria dos pacientes nasceu?
(b) Qual a média de idades dos pacientes do conjunto de dados?

8. Como você abordaria o problema de um histograma depender do número e local dos *bins/buckets*?

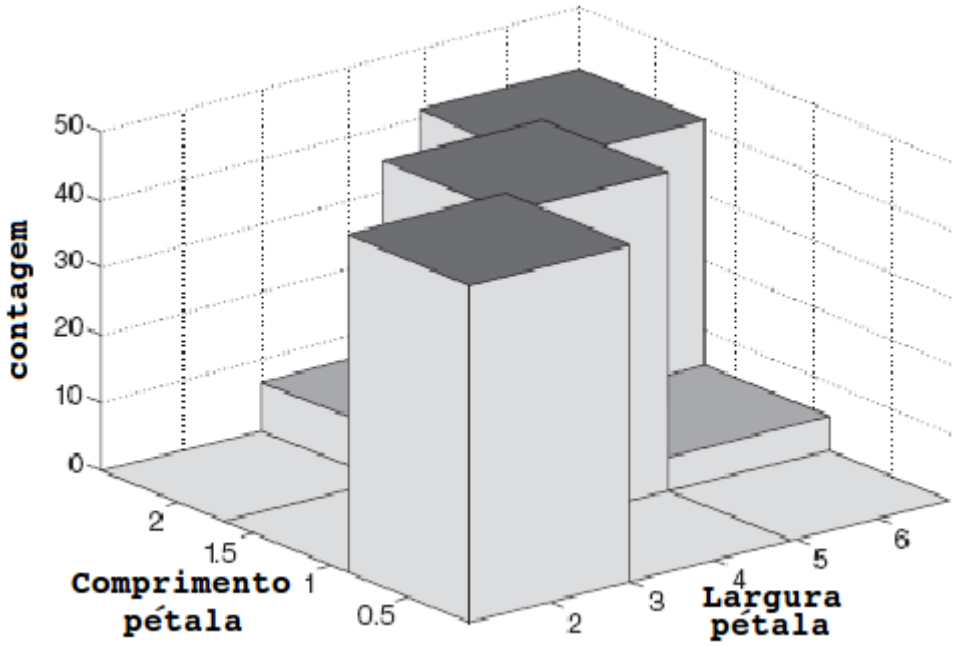
9. Descreva como um gráfico de caixa pode dar informações sobre se o valor de um atributo está distribuído simetricamente. O que você pode dizer sobre a simetria das distribuições dos atributos mostradas na figura a seguir?



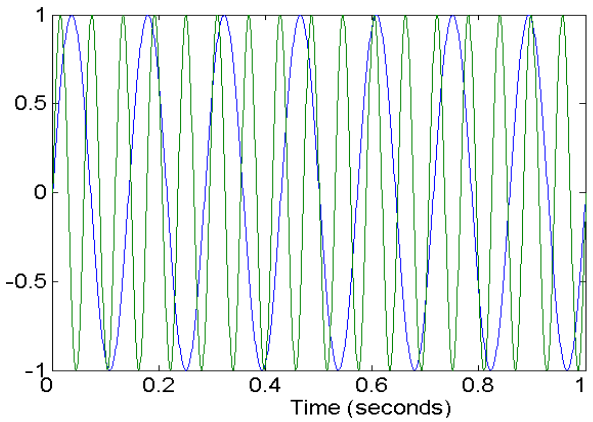
(a) Compare o comprimento da sépala, largura da sépala, comprimento da pétala e da largura da pétala.

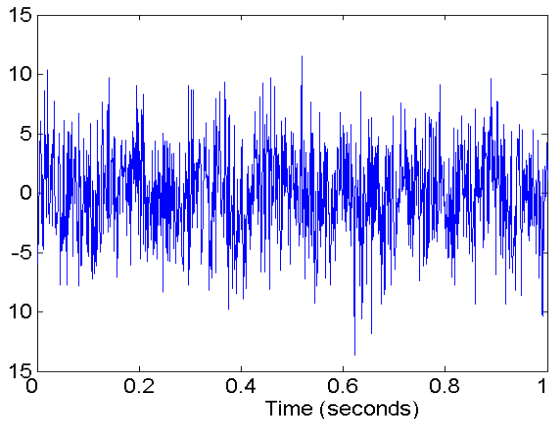
(b) Comente sobre o uso de um gráfico de caixa para explorar um conjun to de dados com quatro atributos: idade, peso, altura e renda.

10. Dê uma explicação possível sobre como a maioria dos valores do comprimento e largura da pétala ficam em baldes pela diagonal da Figura a seguir.

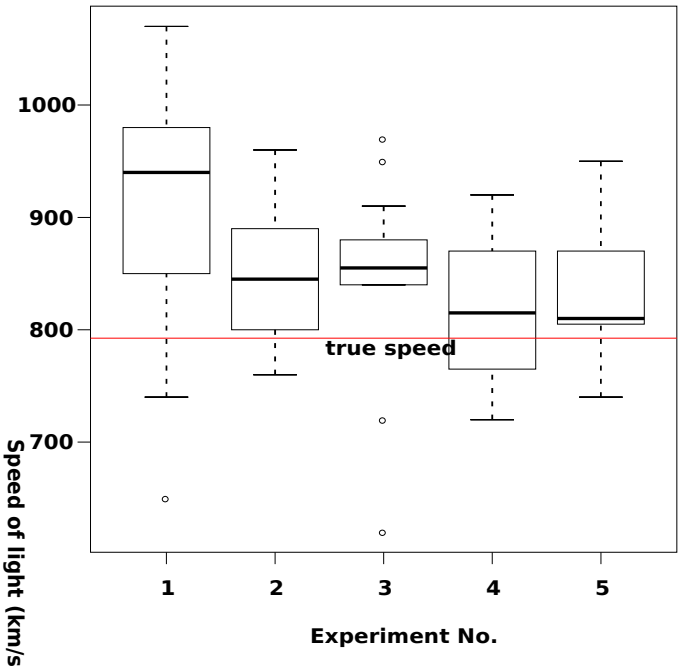


11. Gráficos simples de linhas, podem ser usadas para exibir eficientemente dados de dimensões altas. Por exemplo, na figura a seguir é fácil perceber que as frequências das duas séries de tempo são diferentes. Quais características de séries de tempo permite a visualização efetiva de dados com dimensão alta?





12. Considere o gráfico de caixas a seguir, que contém dados sobre 5 tipos de experimentos feitos por Michelson-Morley para mensurar a velocidade da luz e responda as perguntas a seguir:



(a) Sabendo que a verdadeira velocidade da luz é representada pela linha vermelha, qual tipo de experimento possui mais resultados abaixo da velocidade da luz conhecida? Justifique.

(b) Qual tipo de experimento mostrou maior variação de resultados e qual possui a menor variação? Justifique.

(c) Podemos afirmar que 75% dos experimentos de todos os tipos apresentaram velocidade acima da real? Justifique.

14. Descreva os tipos de situações que produzem cubos de dados de esparsos. Ilustre com exemplos.

15. Como você poderia estender a noção de análise de dados multidimensionais de modo que a variável alvo seja uma variável qualitativa? Em outras palavras, que tipos de estatísticas de resumo ou visualização de dados seriam interessantes?

16. Discuta as diferenças entre redução de dimensionalidade baseada em agregação e redução de dimensionalidade baseada em técnicas como PCA e SVD.