Lista de Exercícios de Data Mining 1
26 de Outubro de 2015

1. Numa determinada aplicação existe informação sobre as idades de um conjunto de 12 pessoas. Os seus valores são: `12, 30, 24, 10, 10, 23, 43, 67, 79, 34, 56, 51`.

   a) Qual é a mediana destas idades? Justifique.

   b) Qual é o valor da moda destas idades? Justifique.

   c) Indique como obteria em R a diferença entre o percentil 99% e o percentil 1% deste conjunto de idades.

   d) Qual seria o resultado da normalização e "standardization" destas idades?

2. Assuma que no conjunto acima, adicionamos duas novas pessoas com idades: 10 meses e 100 anos.

   a) Aplique a normalização e a "standardization" a este novo conjunto de dados.

   b) Há alguma preferência pela utilizaç ão dos dois métodos em cada um dos dois conjuntos de dados (original e com as duas novas pessoas adicionadas)?

3. Responda ás seguintes perguntas:

   a) Qual é o objetivo dos gráficos do tipo "boxplot" ("caixa de bigodes")?

   b) Para que servem as medidas de dispersão de dados: "range" e "interquartile range". Há alguma vantagem em usar um sobre o outro?

   c) Que outras medidas de dispersão de dados podemos utilizar?

4. A Figura 1 mostra um "scatterplot". Que informações consegue retirar deste gráfico?

5. A Figura 2 mostra um "parallel plot". Que informações consegue retirar deste gráfico?

6. Ao fazer visualização de dados pode ser importante ordenar valores de variáveis ou simplesmente mudar a ordem de disposição das mesmas. Dê um exemplo onde a ordenação torna-se importante para a melhor visualização dos dados.

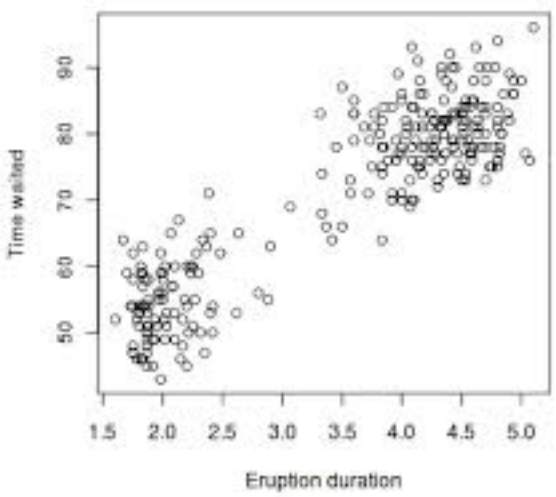

Figure 1: Scatterplot Example

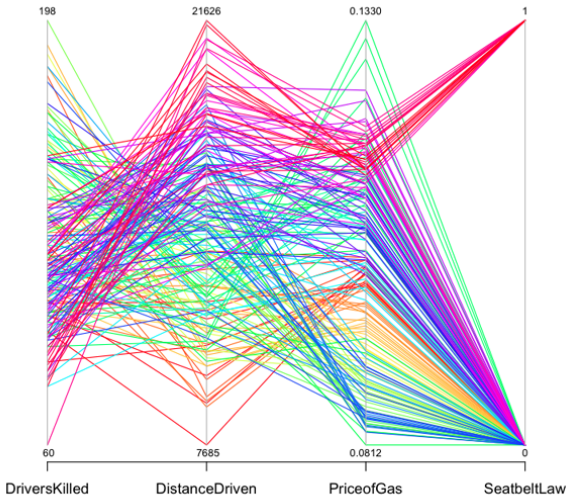

Figure 2: Parallel plot Example

7. A estratégia para encontrar os vizinhos mais próximos ("nearest neighbours") também pode ser utilizada para "imputar" valores de variáveis desconhecidos. Explique como.

8. Qual é a ideia básica da Análise de Componentes Principais (PCA – Principal Component Analsysis) e qual é a sua utilidade?

9. Explique sucintamente o algoritmo k-means.

10. Assuma que possui dados numa tabela em formato CSV (Comma-Separated Values). Quando utiliza a função `read.csv` do R quais são os tipos de dados armazenados internamente quando a variável é numérica com apenas dois valores? E no caso do software WEKA, qual é o tipo interno assumido?

11. Quais são os tipos básicos de dados existentes?

12. Explique a diferença entre a distância calculada utilizando "simple matching" e a distância de Jaccard. Em que situação estes dois tipos de métodos de cálculo de distância devem ser utilizados?

13. O que é um método de aprendizagem supervisionado?

14. Para que serve a análise de "clusters"?

15. Qual é a diferença entre a correlacão de Pearson e a regressão linear simples?

16. Considere a seguinte tabela de dados:

Inst/Var	V1	V2
I1	1.5	1.7
I2	2	1.9
I3	1.6	1.8
I4	1.2	1.5

    Dada uma nova observação (1.4,1.6), quais as duas observações da tabela mais próximas desta nova observação de acordo com a distância Euclideana.

17. Qual é a diferença entre o "clustering" hierárquico **aglomerativo** e o **divisivo**?

18. Explique porque em certas situações a utilização do "Error rate" como forma de avaliar um modelo de classificação pode não ser uma boa ideia?

19. Que estratégia(s) utilizaria para fazer o histograma de uma variável numérica com valores contínuos?

20. Qual é a finalidade de se fazer amostragem de dados?