List of Exercises: Data Mining 1
October 17th, 2018

1. In a given application, we have information about the ages of a set of 12 people. Their values are `12, 30, 24, 10, 10, 23, 43, 67, 79, 34, 56, 51`.

   a) What is the median of these ages? What is the meaning of the median?

   b) What is the mode of these ages? What is the meaning of the mode?

   c) How would you obtain the difference between the 99% percentile and the 10% of this set of ages, in R? What is the meaning of this metric?

   How would you obtain the 1% percentile for this data? What is the meaning of this measure?

   d) What are the results of normalizing and standardizing these data?

   e) Give an example of why normalizing this data.

   f) Give an example of why standardizing this data?

2. Suppose that we add two more age values to the set mentioned in item (1): 10 months and 100 years.

   a) Apply normalization and standardization to this new set of data.

   b) Should you give any preference to apply normalization or standardization to this new set of data?

3. Answer the following questions:

   a) What is the objective of boxplot graphs?

   b) What are the functions of the spread measures: "range" and "interquartile range". Is there any advantage of using one over the other?

   c) What other spread measures can we use to analyse data?

4. Figure 1 shows a "scatterplot". What information can you infer from this graph?

5. Figure 2 shows a "parallel plot". What kind of information can you infer from this graph?

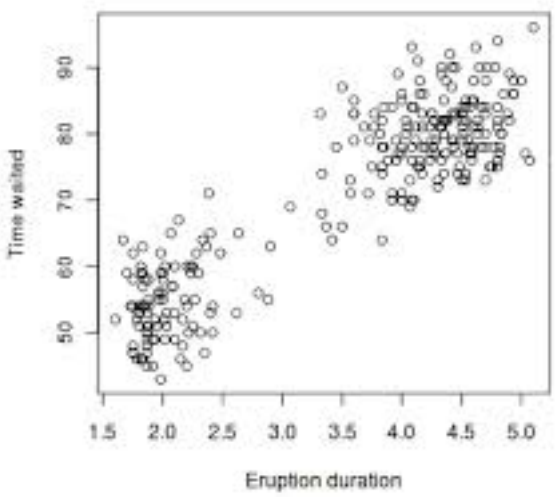

Figure 1: Scatterplot Example

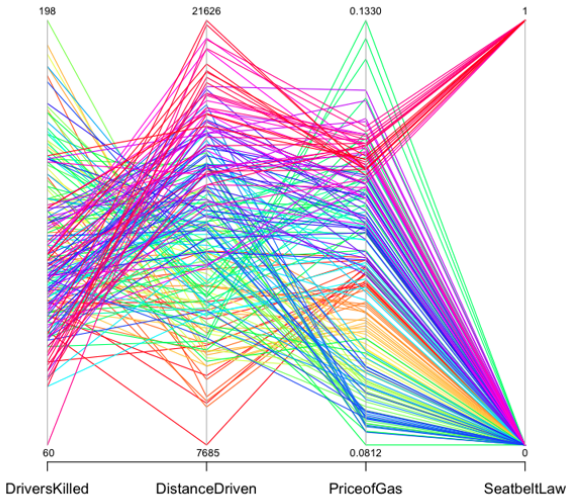

Figure 2: Parallel plot Example

6. When visualizing data, it may be important to reorder/rearrange variables or sort variable values. Give na example, where this order can yield a better visualization than the visualization of the original data.

7. The "nearest neighbours" strategy can also be used to "impute" values to unlnown variable values. Explain how you can use nearest neighbours to impute missing variable values.

8. What is the main idea behind PCA – Principal Component Analsysis and why is it useful?

9. Suppose you are given a CSV (Comma-Separated Values) table. When you read this table using the R function `read.csv`, what would be the internal stored type of a boolean variable in this table? How about when you read the same data in the WEKA software?

10. What types of variables can we find when performing data analysis?

11. Explain the difference between the distance calculated using "simple matching" and the "Jaccard" distance. In what situation, we apply one or the other?

12. What is the difference between Pearson correlation and simple linear regression?

13. Consider the following data table:

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

Given a new observation (1.4,1.6), which two observations in the table are nearer the new data point, using the Euclidean distance?

14. Explain why using the "Error rate" to evaluate a classification model may not be a good approach.

15. Which approach would you use to plot a histogram for a continuous numeric variable.

16. What is the objective of data sampling?

17. What is the objective of data sampling?

18. Consider the training set shown in Table 1 for a problem of binary classification.

Table 1: Table for question (18)

Instance	a_1	a_2	a_3	Class
1	T	T	2.0	+
2	T	T	5.0	+
3	T	F	7.0	−
4	F	T	3.0	+
5	F	T	8.0	−
6	T	F	2.0	−
7	T	F	9.0	−
8	F	F	6.0	+
9	F	F	4.0	−

a) What is the entropy of the class variable?

b) What is the of a_1 e a_2 related with the class variable?

c) For a_3, which is a numerical attribute, what is the best split point?

19. What is the importance of tree pruning?

20. Table 2 presents observations of objects. These observations describe the objects by their size, color and shape. Each object is classified as suitable (yes) or not (no) for a given task. Build two decision trees for this data. The first one should be built choosing the attribute names by alphabetical order. The second should be built by choosing the attributes in a more interesting way, using entropy. Discuss about the differences between the two trees.

Table 2: Table for question 20

ID	size	color	shape	class
1	medium	blue	brick	yes
2	small	red	sphere	yes
3	large	green	pillar	yes
4	large	green	sphere	yes
5	small	red	wedge	no
6	large	red	wedge	no
7	large	red	pillar	no