# Data Mining

## Week 4 Exercise Sheet

1. Classify the following attributes as binary, discrete or continuous. Further, classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

   (a) Age in years.

   (b) Time in terms of AM or PM.

   (c) Brightness as measured by a light meter.

   (d) Brightness as measured by people's judgements.

   (e) Angles as measured in degrees between 0 and 360.

   (f) Bronze, Silver and Gold medals as awarded at the Olympics.

   (g) Height above sea level.

   (h) Number of patients in a hospital.

   (i) ISBN numbers for books.

   (j) Ability to pass light in terms of the following values: opaque, translucent, transparent.

   (k) Military rank.

   (l) Distance from the centre of campus.

   (m) Density of a substance in grams per cubic centimetre.

2. (a) You work for a company which decides to keep track of the number of customer complaints for each product. As counts are ratio attributes, you use this to measure product satisfaction. However, your boss disagrees and believes your customer satisfaction measure is worthless. Who is right? Justify your answer fully. If you believe the boss is correct, what can you do to fix the measure of satisfaction?

   (b) A few months later you devise another approach to measure the extent to which a customer prefers one product over another. When developing new products, the usual strategy is to create several variations and then evaluate which one customers prefer. However, test subjects are often indecisive and testing takes a long time. Therefore, you suggest performing comparisons in pairs and then using these comparisons to get the rankings. Thus, if there are three product variations we have customers compare

variations 1 and 2, then 2 and 3, then 3 and 1. Testing time with this new procedure is a third of what it was for the old procedure, but employees conducting the tests complain they cannot come up with a consistent ranking from the results.

i. Will this approach work for generating an ordinal ranking of the product variations in terms of customer preference? Justify your answer.

ii. Is there a way to fix your approach? More generally, what can be said about trying to create an ordinal measurement scale based on pairwise comparisons?

iii. For the original product evaluation scheme, the overall rankings of each product variation are found by computing its average over all test subjects. Comment on whether you think that this is a reasonable approach. What other approaches can be taken?

3. Can you think of a situation in which identification numbers would be useful for prediction?

4. Which of the following quantities are likely to show more spatial autocorrelation: daily rainfall or daily temperature? Why?

5. Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

6. Distinguish between noise and outliers. Be sure to consider the following questions:

   (a) Is noise ever interesting or desirable? Same question for outliers.

   (b) Can noise objects be outliers?

   (c) Are noise objects always outliers?

   (d) Are outliers always noise objects?

   (e) Can noise make a typical value into an unusual one, or vice versa?

7. Perform both equal frequency and equal width binning on the following data:

   5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215

   . In each case, divide the data into 3 bins.

8. (a) Use smoothing by bin means (using a bin depth of 3) to smooth the following data set:

   $$[52, 15, 16, 40, 19, 33, 20, 22, 21, 22, 25, 35, 25, 25, 30, 20, 33, 70, 35, 25, 35, 36, 16, 45, 46, 13, 35].$$

   (b) How may you determine outliers in the data?

   (c) What other methods are there for data smoothing?