Data Mining HW 1

Matt Venter

Tan Chapter 1:

1. (Question1) Discuss whether or not each of the following activities is a data mining task.
   a. Dividing the customers of a company according to their gender.
      a. This **is not** a data .mining task as you can easily just search it using a query
   b. Dividing the customers of a company according to their profitability.
      a. This a **is not** a data mining task, this is already found data in the records and searching and dividing this attribute does not make it a data mining task.
   c. Computing total sales of a company.
      a. This a **is not** a data mining task, in fact this is something accountants do on the regular and is just a sum of sales.
   d. Sorting a student database based on student identification numbers.
      a. This a **is not** a data mining task, its just a database algorithm and can easily be done.
   e. Predicting the outcomes of tossing a (fair) pair of dice.
      a. This a **is not** a data mining task, because it is just a probability calculation without large amounts of data.
   f. Predicting the future stock price of a company using historical records.
      a. This **is** a datamining task, as it deals with prediction of the future based off of large data records from the past.
   g. Monitoring the heart rate of a patient for abnormalities
      a. This **is** a datamining task, and it is called anomaly detection
   h. Monitoring seismic waves for earthquake activities.
      a. This **is** a datamining task, and it is called anomaly detection
   i. Extracting the frequencies of a sound wave.

    a. This **is not** a data mining task; it seems as this is actually preparing data to be used for data mining though.

2. (Question 3) For each of the following data sets, explain whether or not data privacy is an important issue.
    a. Census data collected from 1900-1950.

       This data is old and most of it is inaccurate fir the current time, so it Is not an important issue.

    b. IP address and visit times of Web users who visit your Website.

       This data (especially IP address) can be used maliciously

    c. Images from Earth-orbiting satellites.

       This data should be fine but there is some stuff that is blurred out if need be .

    d. Names and addresses of people from the telephone book.

       This is already accessible by the public, so it is not an important issue.

    e. Names and email address collected form the Web.

       This is already accessible by the public, so it is not an important issue.

Tan Chapter 2:

1. (Question 2) Classify the following attributes as binary, discrete, or continuous as well as specific quantitative or qualitative?
    a. Time in terms of AM or PM.
        i. Binary, qualitative, nominal
    b. Brightness as measured by a light meter.
        i. Discrete, quantitative, ratio
    c. Brightness as measured by people’s judgments.
        i. Discrete, qualitative, ordinal
    d. Angles as measured in degrees between 0 and 360

            i. Discrete, quantitative, ratio

    e. Bronze, Silber, and Gold Metals as awarded in the Olympics
        i. Binary, qualitative, ordinal
    f. Height above sea level
        i. Discrete, quantitative, ratio (or interval if the regions use different measurement systems)
    g. Number of patients in a hospital.
        i. Discrete, quantitative, ratio
    h. ISBN numbers for books.
        i. Discrete, qualitative, nominal (unless order matters)
    i. Ability to pass light in terms of the following values: opaque, translucent, transparent.
        i. Discrete, qualitative, ordinal
    j. Military rank.
        i. Discrete, qualitative, ordinal
    k. Distance from the center of campus.
        i. Continuous, quantitative, ratio
    l. Density of a substance in grams per cubic centimeter
        i. Continuous, quantitative, ratio
    m. Coat check number.
        i. Discrete, qualitative, ordinal (unless not ordered with reason)

2. (Question 3)
    a. Who is right. The marketing director or his boss?
        i. The boss is correct because the total number of complaints for a product is skewed toward higher selling products a better measurement would be negative reviews as a ratio to how many products sold.
    b. What can you say about the attribute type of the original product satisfaction type?
        i. It is qualitative, quantitative, and ratio

3. (Question 7) Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall, or daily temperature?
    a. Daily temperature because most of the time it follows a pattern based on the time of year and other factors while daily rainfall can almost be random at times.

4. (Question 12) Distinguish between noise and outliers.
    a. Is noise or outliers interesting or desirable?
        i. Noise is not interesting or desirable but for some circumstance’s outliers are.
    b. Can noise objects be outliers?
        i. Yes, if the standard deviation is high enough

c. Are noise objects always outliers?
   i. No, most likely will be in the range of Std. Deviation as well

d. Are outliers always noise objects?
   i. No (same kind of as above)

e. Can noise make a typical value into an unusual one, or vice versa?
   i. Yes, it can do both.

ISLR 7e 3.7

1. (Question 1) Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.
   a. T null hypothesis is not present between sales and TV or sales and radio when it comes to newspapers though, the null hypothesis is very present. This basically means you get more sales when investing in advertising with tv or radio but when you invest in newspaper advertising you are likely to not see sales increases.

2. (Question 3) . Suppose we have a data set with five predictors, X1 = GPA, X2 = IQ,X3 = Gender (1 for Female and 0 for Male), X4 = Interaction between GPA and IQ, and X5 = Interaction between GPA and Gender. The response is starting salary after graduation (in thousands of dollars).Suppose we use least squares to fit the model, and get βˆ0 = 50, βˆ1 = 20, βˆ2 = 0.07, βˆ3 = 35, βˆ4 = 0.01, βˆ5 = −10.
   a. Which answer is correct?
      a. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough. (iii)
   b. Predict the salary of a female with IQ of 110 and a GPA of 4.0.
      a. $137,100
   c. True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.
      False there is not enough info given to find out if the is an interation effect.
3. (Question 4) I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. Y = β0 + β1X + β2X2 + β3X3 + ε.
   a. Suppose that the true relationship between X and Y is linear, i.e. Y = β0 + β1X + ε. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic

regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

a. Without all true info it is hard to tell but if the relationship is linear between X and Y then the RSS should be lower for the linear regression.