DATA MINING

CS422- Section 01

Illinois Institute of Technology

HOMEWORK – 1

Megha Tatti (CWID: A20427027)

Exercise 1:

Section 1.1: Chapter 1

Question 1: Discuss whether each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender**.**
Ans) No, it's not a data mining task. Here we just can use a simple database query to divide the customers according to their gender.

(b) Dividing the customers of a company according to their profitability.
Ans) No, this is not a data mining task. Using a formula, profitability can be calculated and that can later be used to divide the customers based on its value i.e. their profitability.

(c) Computing the total sales of a company.
Ans) No. This can be done using simple accounting calculation.

(d) Sorting a student database based on student identification numbers.
Ans) No, it's not a data mining task. Here we just can use a simple database query.

(e) Predicting the outcomes of tossing a (fair) pair of dice.
Ans) No, it's not a data mining task. This is clearly a probability calculation.

(f) Predicting the future stock price of a company using historical records.
Ans) Yes, it's a data mining task. Using historical records, the future stock price can be predicted by building a data model.

(g) Monitoring the heart rate of a patient for abnormalities.
Ans) Yes, it's a data mining task. A data model can be built to fetch the normal heart rate of a patient. This can be used to monitor for abnormalities.

(h) Monitoring seismic waves for earthquake activities.
Ans) Yes, it's a data mining task. We can build a data model of different types of seismic wave behavior and use this model to monitor seismic waves.

(i) Extracting the frequencies of a sound wave.
Ans) No. This is signal processing.

Question 3: For each of the following data sets, explain whether data privacy is an important issue.

(a) Census data collected from 1900–1950.
Ans) No.

(b) IP addresses and visit times of web users who visit your website.
Ans) Yes.

(c) Images from Earth-orbiting satellites.
Ans) No.

(d) Names and addresses of people from the telephone book.
Ans) No.

(e) Names and email addresses collected from the Web.
Ans) No.

Section 1.2 -Chapter 2

Question 2: Classify the following attributes as binary, discrete, or continuous. Also 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) Time in terms of AM or PM.
Ans) Binary, Qualitative, Ordinal.

(b) Brightness as measured by a light meter.
Ans) Continuous, Quantitative, Ratio.

(c) Brightness as measured by people's judgments.
Ans) Discrete, Qualitative, Ordinal.

(d) Angles as measured in degrees between 0◦ and 360◦.
Ans) Continuous, Quantitative, Ratio.

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.
Ans) Discrete, Qualitative, Ordinal.

(f) Height above sea level.
Ans) Continuous, Quantitative, Interval/Ratio (depends on where the measure is started).

(g) Number of patients in a hospital.
Ans) Discrete, Quantitative, Ratio.

(h) ISBN numbers for books. (Look up the format on the Web.)
Ans) Discrete, Qualitative, Nominal.

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent.
Ans) Discrete, Qualitative, Ordinal.

(j) Military rank.
Ans) Discrete, Qualitative, Ordinal.

(k) Distance from the center of campus.
Ans) Continuous, Quantitative, Interval/Ratio (Interval as the distance can be calculated from center to the starting point. Ratio as length can be either in km or miles.)

(l) Density of a substance in grams per cubic centimeter.
Ans) Discrete, Quantitative, Ratio.

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.)
Ans) Discrete, Qualitative, Nominal.

Question 3: You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

(a) Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction?
Ans) The boss is right. The correct measure to calculate customer satisfaction would be to consider both the sales of the product and the complaints for the same product.
So, Customer Satisfaction (Product) = Number of complaints for product / Total Number of sales for the product.

(b) What can you say about the attribute type of the original product satisfaction attribute?
Ans) It is hard to determine the attribute type as different products may have different number of complaints but same level of customer satisfaction.

Question 7:

Q) Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?
Ans) Daily temperature shows more temporal autocorrelation than daily rainfall due to the factors contributing to them being the same. When measured over a time of a month the temperature is bound to show more similarities when it is measured over a closer range of time.

Question 12: Distinguish between noise and outliers. Be sure to consider the following questions.
Ans) Noise does not provide any useful information. They usually occur as a result of any errors during data transmission or collection.
Outliers are data points that are far away from other points in the dataset. They are valid points that are present in the dataset. They provide useful information as opposed to noise.

(a) Is noise ever interesting or desirable? Outliers?
Ans) Noise- No
Outliers-Yes, as they can be used to detect anomalies.

(b) Can noise objects be outliers?
Ans) Yes, sometimes noise can appear as outliers in a dataset.

(c) Are noise objects always outliers?
Ans) No, all noise objects are not always outliers. They can also appear as normal values.

(d) Are outliers always noise objects?
Ans) No, outliers are never noise objects

(e) Can noise make a typical value into an unusual one, or vice versa?
Ans) Yes, noise can convert a normal data point into noise and vice versa.

Section 1.3 – ISLR 7e(Gareth James, et al.)

Question 1:

Q) 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.

Ans) It is null hypothesis if the coefficient corresponding to Intercept, TV, radio and newspaper i.e., B_0, B_1, B_2, B_3 is zero. The p-values for intercept, TV and radio is less than 0.05 and hence we can reject the null hypothesis concluding that the coefficient is non-zero and are significant in predicting sales. But the p-value for newspaper is greater than 0.05 (alpha level of the test) suggesting we fail to reject the null hypothesis and conclude that B_3 is zero. Thus, we can conclude that TV and radio are significant in predicting sales, but not newspaper.

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 $\hat{\beta}0 = 50$, $\hat{\beta}1 = 20$, $\hat{\beta}2 = 0.07$, $\hat{\beta}3 = 35$, $\hat{\beta}4 = 0.01$, $\hat{\beta}5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA, males earn more on average than females.
ii. For a fixed value of IQ and GPA, females earn more on average than males.
iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

Ans) The least square line is given by

$$\hat{y} = 50 + 20GPA + 0.07IQ + 35Gender + 0.01(GPA \times IQ) - 10(GPA \times Gender)$$

Hence, for males the substituting gender = 0.

$$\hat{y} = 50 + 20GPA + 0.07IQ + 0.01(GPA \times IQ)$$

For females the substituting gender = 1.

$$\hat{y} = 85 + 10GPA + 0.07IQ + 0.01(GPA \times IQ)$$

So, the starting salary for males is higher than for females on average if and only if

$$50 + 20GPA \geq 85 + 10GPA$$

which is equivalent to $GPA \geq 3.5$.

Therefore, [**option (iii)**]- for a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.
Ans) Substituting the given values in the least square line for females given above, we obtain

$\hat{y} = 85 + 40 + 7.7 + 4.4 = 137.1$

From this we can predict that the starting salary would be $137000.

(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.
Ans) False. It's possible to have a lot of evidence for a small effect. Also, a small coefficient doesn't even mean the interaction effect is small, since it is very sensitive to the units of the two variables. To verify if GPA/IQ has an impact on the quality of the model we have to test the null hypothesis H_0: $\hat{\beta}_4 = 0$ and look at the p-value associated with *F* statistic to conclude if there is any interaction effect.

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. $Y = \beta0 + \beta1X + \beta2X2 + \beta3X3 + €$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta0 + \beta1X + €$. Consider the training residual sum of squares (RSS) for the linear regression, and 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.

Ans) With the given data it is not possible to decide which RSS is lower between linear or cubic. But since it is mentioned that there is a true relationship between X and Y which is linear, the least squares line may be close to the regression line, and consequently the RSS for linear regression may be lower than RSS for cubic regression