CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Jithin Joyson

Spring 2019: Homework 1

1 Exercises

1.1 1.

a. **No**, since this can be brought up using a simple SELECT statement/query with a gender condition in a database.
b. **No**, since this can also be brought up using a simple SELECT statement/query with a condition for each profitability range in a database.
c. **No**, since this can also be brought up using a simple SELECT statement/query with an aggregate function over sales in a database.
d. **No**, since this can be brought up using a simple SELECT statement/query with an ordering function over identification numbers in a database.
e. **No**. Even though there is prediction involved, this is a fair dice where the probabilities can be computed without mining data.
f. **Yes**, since the stock price of a company can only be predicted after in depth analysis of many variables in the data. This falls under predictive modeling of data mining and uses models to get a prediction.
g. **Yes**, since this is an application of data mining in the medical field to use feature analysis of many entries (patients) in the data. This falls under anomaly detection of abnormalities by monitoring outliers in the data.
h. **Yes**, since monitoring waves can be a result of classifying a disturbance as a seismic wave using predictive models.
i. **No**, since this is data preprocessing which can be later used for mining if necessary.

3.

a. **No**, since this is public information.
b. **Yes**, since IP addresses are private and gaining access to who visits your website at a certain time can give insight to the user's personal life.
c. **No**, since this is public information and it does not put anyone's privacy in danger.
d. **No**, since this is public information, and everyone has an equal amount of information put out.
e. **No**, since this is easily accessible information that everyone can be identified with.

1.2 2.

a. Binary, qualitative, ordinal
b. Continuous, quantitative, ratio
c. Discrete, qualitative, ordinal
d. Continuous, quantitative, ratio
e. Discrete, qualitative, ordinal
f. Continuous, quantitative, interval/ratio
    a. Depends what sea level is since it can be (+, -) or have meaningful heights (*, /).
g. Discrete, quantitative, ratio
h. Discrete, qualitative, nominal
i. Discrete, qualitative, ordinal
j. Discrete, qualitative, ordinal
k. Continuous, quantitative, interval/ratio
    a. Depends if center is (0,0) giving way to (+, -) or have meaningful distances (*, /).
l. Discrete, quantitative, ratio
m. Discrete, qualitative, nominal

3.

a. The **boss is right** since the marketing director only looked at the number of complaints without giving regard to the volume of responses due to it being the best-selling product. We can transform the measure of satisfaction to a **ratio of the number of complains over the number of sales** for a given product to normalize the satisfaction results.
b. Since the original product satisfaction does not explain how it can be placed verses other responses, it **cannot be categorized** as ratio or any other attribute type.

7. Temporal autocorrelation is the closeness of two measurements if they are closer in time. In this scenario, **daily temperature shows more temporal autocorrelation** since daily rainfall is more susceptible to change than temperature; one day it can be rainy and the next be sunny. Spatial autocorrection is analogous to temporal autocorrelation. Daily temperature also shows more spatial autocorrelation since daily rainfall at a specific location is more variant compared to daily temperature.

12.

a. **Noise is not interesting or desirable** since it counteracts what data mining is doing to analyze patterns; noise introduces variance into the data. **Outliers**

are interesting since it can help with detect anomalies in the data and produce better understanding of it.

b. **Yes**, since outliers are objects that vary from the data and often are random (noise).
c. **No**, since randomness can result in normal objects (masquerading).
d. **No**, since outliers are always variant from normal objects (noise can be normal objects).
e. **Yes**, since it randomly interferes with normal objects.

1.2.1 1. The null hypothesis is that **TV, radio and newspaper has no impact on sales**. After looking at the p-values, it can be concluded that TV and radio do have impact on sales (reject null hypothesizes) but newspaper do not have impact on sales (keep null hypothesis). Since we are testing for the null hypothesis, we can certainly **state newspaper do not impact sales**.

3. y = B0 + B1(GPA) + B2(IQ) + B3(GENDER) + B4(GPA)(IQ) + B5(GPA)(GENDER)
y = 50 + 20(GPA) + 0.07(IQ) + 35(GENDER) + 0.01(GPA)(IQ) -10(GPA)(GENDER)
Male (0): y = 50 + 20(GPA) + 0.07(IQ) + 0.01(GPA)(IQ)
Female (1): y = 85 + 10(GPA) + 0.07(IQ) + 0.01(GPA)(IQ)

a. Male: y = 50 + 20(FIXED VALUE)
Female: y = 85 + 10(FIXED VALUE)
B2 and B4 variables drop since they are fixed values. Now the equation is:
50 + 20(FIXED VALUE) > 85 + 10(FIXED VALUE)
10(FIXED VALUE) > 35
FIXED VALUE > 3.5
Thus, if GPA > 3.5, males earn more on average that females. Therefore**, iii is correct**.
b. y = 85 + 10(4.0) + 0.07(110) + 0.01(4.0)(110)
= 137.1
= $137,100
c. To test impact of variables (IQ and GPA) on salary, null hypothesizes need to be created and tested using p-values; the weights (or coefficients) should not be looked for the impact. Thus, **False**.

4.

a. Since the true relationship is linear, it is expected that the RRS for linear regression to be lower than RRS for cubic regression; since RRS are higher when is more off the true regression line.