CS422 - Data Mining

Homework 1

Problem 1.1:

Question 1:

a) Dividing the customers of a company according to their gender

Answer: No

b) Dividing the customers of a company according to their profitability

Answer: No

c) Computing the total sales of a company

Answer: No

d) Sorting a student database based on student identification numbers

Answer: No

e) Predicting the outcomes of tossing a (fair) pair of dice

Answer: No

f) Predicting the future stock price of a company using historical records

Answer: Yes. Because we are predicting the future data using historical information. (predictive Modeling)

g) Monitoring the heart rate of a patient for abnormalities.

Answer: yes. Here we are trying to identify the patient's abnormalities which is completely different from rest of heart rate data. (anomaly Detection)

h) Monitoring seismic waves for earthquake activities.

Answer: Yes. Here we are monitoring the seismic waves to predict the future earthquakes. (prediction)

I) Extracting the frequencies of a sound wave.

Answer: No

Question 3:

a) Census data collected from 1900–1950

Answer: No

b) IP addresses and visit times of Web users who visit your Website

Answer: Yes.

c) Images from Earth-orbiting satellites

Answer: No

d)Names and addresses of people from the telephone book

Answer: No

e) Names and email addresses collected from the Web

Answer: No

Problem 1.2:

Question 2:

a) Time in terms of AM or PM.

Answer: Binary, Qualitative, Ordinal

b) Brightness as measured by a light meter.

Answer: Continuous, Quantitative, Ratio

c) Brightness as measured by people's judgments.

Answer: Discrete, Qualitative, Ordinal

d) Angles as measured in degrees between 0◦ and 360◦.

Answer: Continuous, Quantitative, Ratio.

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

Answer: Discrete, Qualitative, Ordinal.

f) Height above sea level.

Answer: Continuous, Quantitative, Interval.

g) Number of patients in a hospital.

Answer: Discrete, Quantitative, Ratio.

h) ISBN numbers for books.

Answer: Discrete, Qualitative, Nominal.

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

Answer: Discrete, Qualitative, Ordinal.

j) Military rank.

Answer: Discrete, Qualitative, Ordinal.

k) Distance from the center of campus.

Answer: Continuous, Quantitative, Ratio.

l) Density of a substance in grams per cubic centimeter.

Answer: Discrete, Quantitative, Ratio.

m) Coat check number.

Answer: Discrete, Qualitative, Nominal.

Question 3:

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?

Answer: Boss must be right because the marketing director did not consider the total number of products sold. Rather he just took the number of complaints on each product.

Fix for measure of product satisfaction:

Product satisfaction = # of complaints on a product / Total number of sales on the product.

b) What can you say about the attribute type of the original product satisfaction attribute?

Answer: Using product satisfaction attribute we can't arrive at a conclusion because any pair of products can have same number of complaints but doesn't necessarily need to have same level of customer satisfaction.

Question 7:

Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

Answer: Daily Temperature shows more temporal autocorrelation than daily rainfall.

Objects that are physically close to each other tend to have similar properties than that are far apart. Here Locations that are close to each other tend to have similar daily temperature than daily rainfall because the amount of rainfall can change drastically from one location to another location. But the daily temperature seems to be similar when the locations are close to each other.

Question 12:

a) Is noise ever interesting or desirable? Outliers?

Answer: noise ever interesting or desirable - No

Outliers - Yes

b) Can noise objects be outliers?

Answer: Yes, because more amount of noise can lead to outliers.

c)Are noise objects always outliers?

Answer: No because small amount of noise to an object can result in another object like the current object.

d)Are outliers always noise objects?

Answer: No because outliers are objects that are entirely different from the normal object characteristics.

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

Answer: Yes

Problem 1.3:

Question 1:

Answer: The null hypothesis for TV feature is that when both radio and newspaper are present TV feature has no effect on sales attribute. The null hypothesis for Radio feature is that when both tv and newspaper are present Radio feature has no effect on sales attribute. The null hypothesis for Newspaper feature is that when both radio and tv are present Newspaper feature has no effect on sales attribute. From the P values given in the table 3.4 we can infer that P value for newspaper is high (close to 1) so Null hypothesis for Newspaper Feature hold. On the other hand, TV and Radio have low P value which states that Null Hypothesis for Tv and Radio doesn't hold, and these variables have high correlation towards the sales.

Question3:

a)

Answer: iii)

Result(y) = 50 + 20(GPA) + 0.07(IQ) + 35(Gender) + 0.01(GPA -> IQ) – 10(GPA -> Gender).

Here Beta3 and Beta5 both have gender. Beta3 is positive and Beta5 is negative. Beta3 won't have an effect on males as male gender is represented as 0. So once Beta5 multiplier increases after a certain value it will have negative impact for females. 35/10 = 3.5. So, when GPA is greater than 3.5 Male students earn more on average than the female students after graduation.

b)

Answer:

Result(y) = 50 + 20(GPA) + 0.07(IQ) + 35(Gender) + 0.01(GPA -> IQ) – 10(GPA -> Gender).

= 50 + 20(4) + 0.07(110) + 35(1) + 0.01(4*110) – 10(4*1)

=137.1

Result = 137.1.

c)

Answer: False

Even though the value of IQ is high compared to the other predictors, the regression coefficients could be small. So, we can't comment on the interaction effect.

Question 4:

a)

Cubic Regression has more number of predictors and it will be a better fit to the data than the linear regression model. So, the Cubic regression model will have low RSS compared to the linear regression model.