

Data Mining Midterm, Fall 2009

Professor David Mease

Name: _____

*By signing my name below I attest under penalty of the University that I have done my own work on this exam and have not been assisted by other individuals or references (except the single page of notes) while taking this exam. Further, I attest that I have followed all the exam rules as listed below.

*Signature: _____

Exam Rules:

There are 100 points. There are 32 questions. You may not use any references other than the single 8.5" by 11" page (both sides) of notes which you have brought. You may not use a computer but you may use a hand held calculator.

1) (3 points) What is the definition of data mining used in your textbook? Write the letter of your answer here: **B**

- A) the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data
- B) the process of automatically discovering useful information in large data repositories
- C) an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data

2) (3 points) Which of the following is NOT an example of data mining according to the lecture? Write the letter of your answer here: **C**

- A) discovering that certain names are more prevalent in certain US locations using an electronic phone book
- B) grouping together similar documents returned by a search engine according to their context
- C) querying a search engine to learn information about “Amazon”
- D) predicting future tax fraud from past tax returns

3) (3 points) According to the lecture, data mining draws ideas from which of the following? Write the letter of your answer here: **A**

- A) machine learning, AI, pattern recognition and statistics
- B) machine learning, AI, pattern recognition, statistics and experimental design
- C) pattern recognition and statistics
- D) the military and the government
- E) congress and the supreme court
- F) Al Gore

4) (3 points) A .csv file containing 60,000 rows is read into Excel. A column is added and populated with values from the RAND() function. The data set is sorted by this column and the first 100 rows are selected for analysis. This is an example of what? Write the letter of your answer here: D

- A) Manipulating data to obtain the answer you want
- B) Visualization
- C) A simple random sample with replacement
- D) A simple random sample without replacement
- E) Classification
- F) Clustering

5) (3 points) Which of the following is true of Microsoft Excel? Write the letter of your answer here: A

- A) The row limit is a number which is less than 2 million
- B) The column limit is a number which is less than 7
- C) It can only handle numerical data and not categorical data
- D) It does not contain a function to calculate quantiles
- E) It does not contain a function to calculate standard deviation
- F) It can not read in .txt file formats

6) (3 points) If my data frame in R is called "data", which of the following will give me the third column? Write the letter of your answer here: D

- A) data[2,] B) data[3,]
- C) data[,2] D) data[,3]
- E) data(2,) F) data(3,)
- G) data(.2) H) data(.3)

7) (3 points) If a person's height is measured in inches then it is what kind of attribute? Write the letter of your answer here: **D**

- A) Nominal
- B) Ordinal
- C) Interval
- D) Ratio

8) (3 points) Which of the following is true according to the lecture? Write the letter of your answer here: **A**

- A) Attributes are sometimes called variables and objects are sometimes called observations
- B) All continuous variables are ratio
- C) Binary variables are sometimes continuous
- D) Diet Coke causes people to gain weight

9) (3 points) If a data set is space delimited, what should be done to allow a text string that includes a space so that R or Excel will not split the string into 2 columns? Write the letter of your answer here: **A**

- A) Escape it
- B) Remove the space
- C) Use all capitals in the string
- D) Select "Fix the spaces" from the menu bar

10) (3 points) If I want a zip code to be treated as a categorical variable in R, I should make sure it is of what type? Write the letter of your answer here: **E**

- | | |
|-------------------------|-----------------------|
| A) a vector | B) a list |
| C) a regular expression | D) an integer |
| E) a factor | F) this is impossible |

11) (3 points) As you increase the sample size, the sampling error for the sample mean does what? Write the letter of your answer here: E

- A) It remains constant
- B) It increases proportional to the square of the sample size
- C) It decreases proportional to the square of the sample size
- D) It increases proportional to the square root of the sample size
- E) It decreases proportional to the square root of the sample size

12) (3 points) What is the R command to change the default directory which we used in class? Write the letter of your answer here: F

- A) chdir()
- B) ls()
- C) ls-a()
- D) cd()
- E) cd-a()
- F) setwd()
- G) go()

13) (3 points) We often need to add noise to the points in a scatter plot in what situation? Write the letter of your answer here: C

- A) when we are uncertain of our values
- B) when both attributes are nominal
- C) when both attributes are discrete
- D) when neither attribute is ratio
- E) when neither attribute is nominal
- F) when we sample to reduce the number of points

14) (3 points) If I plot the ECDF for male incomes and the ECDF for female incomes on the same plot and the ECDF for the women is decreasing, then I know what? Write the letter of your answer here: A

- A) I did something wrong
- B) women's salaries were decreasing over some time interval
- C) women's salaries were increasing over some time interval
- D) women make less than men in this data
- E) some of the percentiles for the women are larger than those for the men, but not all

15) (3 points) If I plot the ECDF for male incomes and the ECDF for female incomes on the same plot, and the two ECDF's cross then I know what? Write the letter of your answer here: E

- A) I did something wrong
- B) women's salaries were decreasing over some time interval
- C) women's salaries were increasing over some time interval
- D) women make less than men in this data
- E) some of the percentiles for the women are larger than those for the men, but not all

16) (3 points) A histogram using points and lines instead of bars is called what? Write the letter of your answer here: C

- A) bar chart
- B) pie chart
- C) polygon
- D) pareto diagram
- E) ogive

17) (3 points) The number of telephones in your house is what kind of attribute? Write the letter of your answer here: **D**

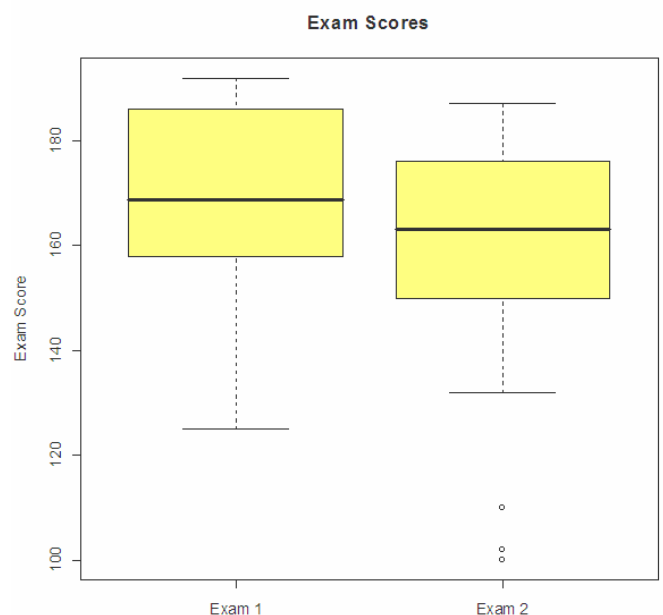
- A) nominal
- B) ordinal
- C) interval
- D) ratio

18) (3 points) The bronze, silver or gold medal awarded at the Olympics is what kind of attribute? Write the letter of your answer here: **B**

- A) nominal
- B) ordinal
- C) interval
- D) ratio

19) (3 points) The boxplots below were produced in R using the default values of boxplot(). What can we tell from this? Write the letter of your answer here: **E**

- A) the minimum of exam 2 is larger than the minimum of exam 1
- B) the median of exam 2 is larger than the median of exam 1
- C) the mean of exam 1 is larger than the mean of exam 2
- D) the mean of exam 2 is larger than the mean of exam 1
- E) the median of exam 2 is in between the the median and 1st quartile of exam 1
- F) the 1st quartile of exam 1 is larger than the mean of exam 2



20) (3 points) What is wrong with the presentation on the right as mentioned in class? Write the letter of your answer here: D

- A) There should also be animation showing a minimum wage job.
- B) In 1970 the minimum wage was actually \$2.80.
- C) A pie chart should be used instead.
- D) The last dollar bill looks more than 3 times as big as the first.
- E) The numbers should be given in percents.

Minimum Wage



1960: \$1.00



1970: \$1.60



1980: \$3.00

21) (3 points) If the mean is larger than the median then this might be an indication that the data is what? Write the letter of your answer here: F

- A) discrete
- B) continuous
- C) filled with a lot of missing values
- D) observational
- E) experimental
- F) right-skewed
- G) left-skewed

22) (3 points) If I have 100 values in my data and I add 5.0 to all of the values, then how will this change the median? Write the letter of your answer here: E

- A) it will not change
- B) it will become larger than the mean
- C) it will become smaller than the mean
- D) it will increase by 0.5
- E) it will increase by 5.0
- F) it will increase by some amount, but we do not have enough information to determine by how much
- G) it may increase or may stay the same

23) (3 points) If I have 100 values in my data and I add 5.0 to all of the values, then how will this change the standard deviation? Write the letter of your answer here: A

A) it will not change

B) it will become larger than the median

C) it will become smaller than the median

D) it will increase by 0.5

E) it will increase by 5

F) it will increase by some amount, but we do not have enough information to determine how much

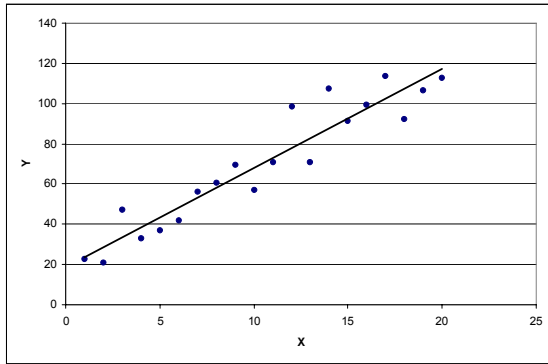
G) it may increase or may stay the same

24) (3 points) Compute the standard deviation for the numbers 23, 25, 30. Show your work below and write your answer here: $\sqrt{13} = 3.6056$

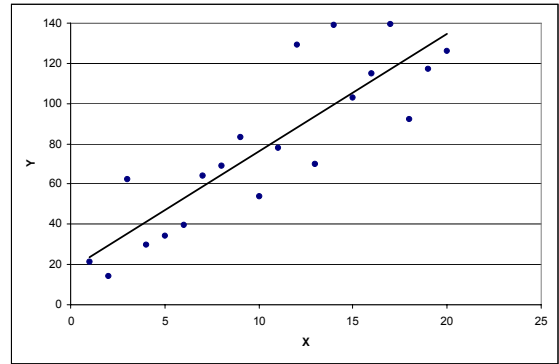
25) (5 points) For these five questions, determine the value of the correlation for each scatter plot. The choices are listed below. You may use each choice at most once.

Choices: $r = -3.20$, $r = -0.98$, $r = 0.86$, $r = 0.95$, $r = 1.20$, $r = -0.96$, $r = -0.40$

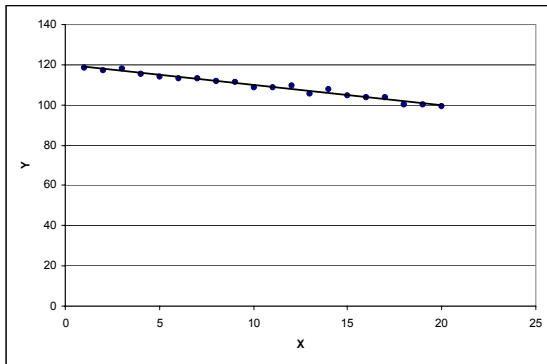
a) Write correlation here: $r = 0.95$



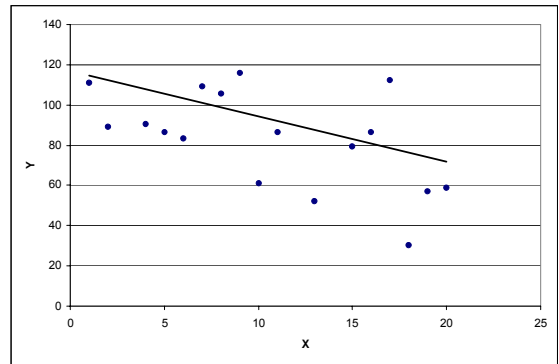
b) Write correlation here: $r = 0.86$



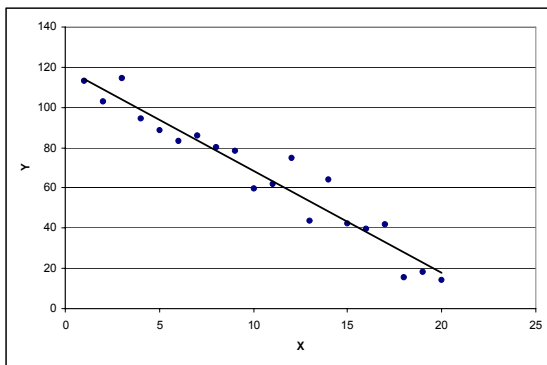
c) Write correlation here: $r = -0.98$



d) Write correlation here: $r = -0.40$



e) Write correlation here: $r = -0.96$



26) (3 points) The confidence for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean? Write the letter of your answer here: E

- A) We rejected the null hypothesis of no relationship at the 5% significance level.
- B) We rejected the null hypothesis of no relationship at the 2.5% significance level.
- C) 95% of all the transactions contain all three of bread, milk and diapers.
- D) 95% of all the transactions contain at least one of bread, milk or diapers.
- E) 95% of the transactions which contain bread also contain both milk and diapers.
- F) 95% of the transactions which contain both milk and diapers also contain bread.

27) (3 points) The support for the association rule {bread} → {milk, diapers} was determined to be 0.95. What does the value 0.95 mean? Write the letter of your answer here: C

- A) We rejected the null hypothesis of no relationship at the 5% significance level.
- B) We rejected the null hypothesis of no relationship at the 2.5% significance level.
- C) 95% of all the transactions contain all three of bread, milk and diapers.
- D) 95% of all the transactions contain at least one of bread, milk or diapers.
- E) 95% of the transactions which contain bread also contain both milk and diapers.
- F) 95% of the transactions which contain both milk and diapers also contain bread.

28) (3 points) Which of the following best describes the two step approach for finding all association rules with a specified minimum confidence and a specified minimum support? Write the letter of your answer here: D

- A) First find all association rules which meet the support requirement and then compute the confidence of each one.
- B) First find all association rules which meet the confidence requirement and then compute the support of each one.
- C) First find all itemsets which meet the confidence requirement and then compute the support for the rules that can be obtained from binary partitions of these.
- D) First find all itemsets which meet the support requirement and then compute the confidence for the rules that can be obtained from binary partitions of these.

Use the table below for the following two questions.

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

29) (4 points) Treating each transaction as a market basket, compute the confidence for the rule $\{a\} \rightarrow \{b, d\}$. Write the letter of your answer here: **B**

- A) .10 B) .14 C) .17 D) .20
E) .30 F) .40 G) .50 H) .60
I) .67 J) .68 K) .80 L) 1.00

30) (4 points) Treating each customer as a market basket, compute the confidence for the rule $\{a\} \rightarrow \{b, d\}$. Write the letter of your answer here: **L**

- A) .10 B) .14 C) .17 D) .20
E) .30 F) .40 G) .50 H) .60
I) .67 J) .68 K) .80 L) 1.00

31) (3 points) Simpson's Paradox is said to occur when which of the following is true? Write the letter of your answer here: **C**

- A) Two variables have a positive correlation but there is no causation because the data is observational rather than experimental.
B) Two variables have a negative correlation but there actually is causation because the data is observational rather than experimental.
C) A 3rd variable causes the observed relationship between a pair of variables to disappear or reverse direction.
D) A 3rd variable causes the observed relationship between a pair of variables to become stronger.

32) (3 points) Explain one of the two reasons mentioned in lecture for favoring a 2-dimensional pie chart over a 3-dimensional pie chart.

3-dimensional pie charts convey no more information than 2-dimensional pie charts and they can be misleading due to the perspective and the rotation.