**SECTION A: (Sample Questions)**

Which of the following does not conform to the “Knowledge discovery in databases” (KDD) process?

* \* Data Source, Data Cleaning, Data Integration, data Selection, Data Mining, and Pattern Evaluation
* \* Data Sources, Data Preprocessing / Integration, Data Exploration, Data Mining, Data Presentation, and Decision Making. (Slide 33 / Page 15 of lecture 0 & 1)
* \* Data Source, Data Preprocessing, Data Mining, Data Post Processing, and Knowledge.
* \* Data Source, Data Mining, Data presentation, and Decision Making.

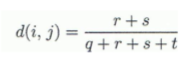
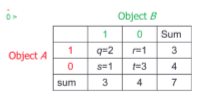
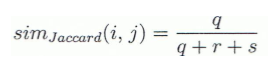
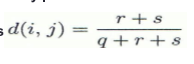
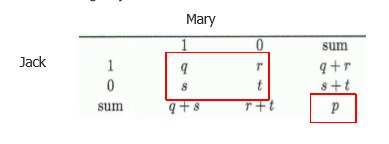
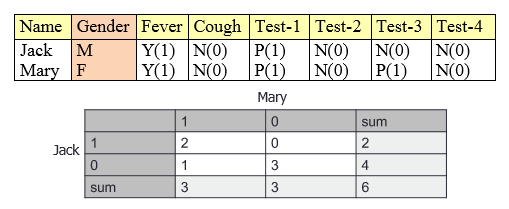
Which of the following is not a proximity measure?

* \* Euclidean distance.
* \* Manhattan distance.
* \* Correlation.
* \* Covariance. (Only tells direction)

Compare and Contrast the following terms in your own word:

* **A. Deterministic Models :: Probabilistic Models (Page 5 & 6 of Lecture 2)**
  + Deterministic model: all data is known before. Once you start the system, you know exactly what is going to happen.
  + Probabilistic model: Element of chance is involved. know the likelihood that something will happen, but don’t know when it will happen.
* **B. Correlation :: Covariance (Page 35 Lecture 4)**
  + • Both correlation and covariance are measures of relation between two random variables. Correlation is the measure of strength of the linearity of the two variables and covariance is a measure of the strength of the correlation. A measure used to indicate the extent to which two random variables change in tandem is known as covariance. A measure used to represent how strongly two random variables are related known as correlation.
  + • Correlation coefficient values are a value between -1 and +1, whereas the range of covariance is not constant, but can either be positive or negative. But if the random variables are standardized before calculating the covariance, then covariance is equal to the correlation and has a value between -1 and +1
* **C. Binary Variables :: Ordinal Variables (Lecture 5 page 4, 11, 15)**

• Binary variables are:

* + - Nominal (i.e., only have two states, 1 and 0).
    - Symmetric binary: both outcomes are equally important (i.e. gender)
      * Distance measure for symmetric binary variables:
        + 
        + 
        + 
    - Asymmetric binary: both outcomes are not equally important (i.e. in health, positive and negative tests)
      * Similarity measure for asymmetric binary variables (AKA **Jaccard coefficient**):
        + 
      * Distance measure for asymmetric binary variables (Dissimilarity):
        + 
        + 
        + 

• Ordinal variables are:

* + - Discrete or continuous
    - Order is important (i.e. freshman, sophomore, junior and senior)
    - Can be treated like interval-scaled
      * You can map the range of each variable onto [0,1] by replacing the i-th object in the f-th variable by:
        + 
        + In other words, freshman:0, sophomore: 0.33, junior: 0.67, senior: 1.

d(freshman, senior) = 1, d(junior, senior) = 0.33

* **D. Dimensionality Reduction:: Numerosity Reduction (Lecture 3 page 25)**
  + Dimensionality reduction, e.g., remove unimportant

variables

• Feature subset selection, feature creation

• Principal Components Analysis (PCA)

• Wavelet transforms

• Removing attributes that are too similar to another attribute

• \*\*Removing attributes that are redundant or irrelevant\*\*

* + Numerosity reduction (some simply call it: Data Reduction)

• Regression and Log-Linear Models

• Histograms, clustering, sampling

• Data cube aggregation

• Removing the rows (the data points), since data can be dirty

* **E. Overfitting :: Multi-collinearity**
  + Over-fitting: Adding more IV’s (Independent Variable) to a regression procedure does not mean that the predictions are going to get “better”; In fact it can make things worse. This is referred to as overfitting.
    - IVs that are too similar may overfit the model
    - Refers to the number of dimensions
  + Multicollinearity: Adding more IV’s creates more relationships between IV’s (apart from the relationship they share with the DV), when this happens it is called multicollinearity.
    - (It is the occurrence of high intercorrelations among independent variables in a multiple regression model. It can lead to skewed or misleading results)
    - Refers to the relation of the dimensions

Multicollinearity arises when 2 or more predictors are correlated with each other; this results in highly variable regression coefficients that strip them of their easily interpretable meaning and ruins any attempt at variable selection via stepwise selection. Overfitting occurs if too many predictors are in the model; an overfitted model detects not only the underlying trend in the data, but also the noise.

<https://chemicalstatistician.wordpress.com/2013/02/20/presentation-slides-overcoming-multicollinearity-and-overfitting-with-partial-least-squares-regression-in-jmp-and-sas/>

Eg. You have multicollinearity with overfitting, but you cannot have overfitting without multicollinearity.

* **F. Equidepth binning ::Equiwidth binning (starts on lecture 5 page 5)**

Equal-depth (frequency) partitioning

• Divides the range into *N* intervals, each containing

approximately same number of samples

• Good data scaling

• Managing categorical attributes can be tricky

Equal-width (distance) partitioning

• Divides the range into *N* intervals of equal size: uniform

grid

* If *A* and *B* are the lowest and highest values of the feature, the width of intervals will be: *W* = (*B* –*A*)/*N.*
* The most straightforward, but outliers may dominate presentation
* Skewed data is not handled well
* **g. Measure::Metric** 
  + A "measure" is a number that is derived from taking a measurement. Your height, weight or temperature would all be measures.
  + In contrast, a "metric" is a calculation between two measures. Typically, the calculation is a form of division. The format of the calculated result can be a percentage, a ratio, a fraction, a decimal or a multiple.
* **H. Similarity::Dissimilarity**
  + Similarity measure or similarity function (e.g. Correlation)

• Numerical measure of how alike two data objects are• Value is higher when objects are more alike

• Often falls in the range [0,1]

* + • Dissimilarity measure or Distance function

• Numerical measure of how different two data objects are

• Lower when objects are more alike

• Minimum dissimilarity is often 0

• Upper limit varies [0,1] or [0, inf) depending on the definition

**I. Positively Skewed Distribution::Negatively Skewed Distribution**

* A positively skewed distribution has a longer tail to the right
* A negatively skewed distribution has a longer tail to the left
* A distribution with no skew (e.g. a normal distribution) is symmetrical: In a perfectly symmetrical, non-skewed, distribution the mean, median and mode are equal.

Match the following Data to its corresponding data types (Nominal, Ordinal, Interval, or Ratio) (Lecture 4 Part 1: Data Preprocessing, pg. 12)

Interval is not linear, like temperatures, the temperature can be 31 one day, 45 the next, and 27 the next.

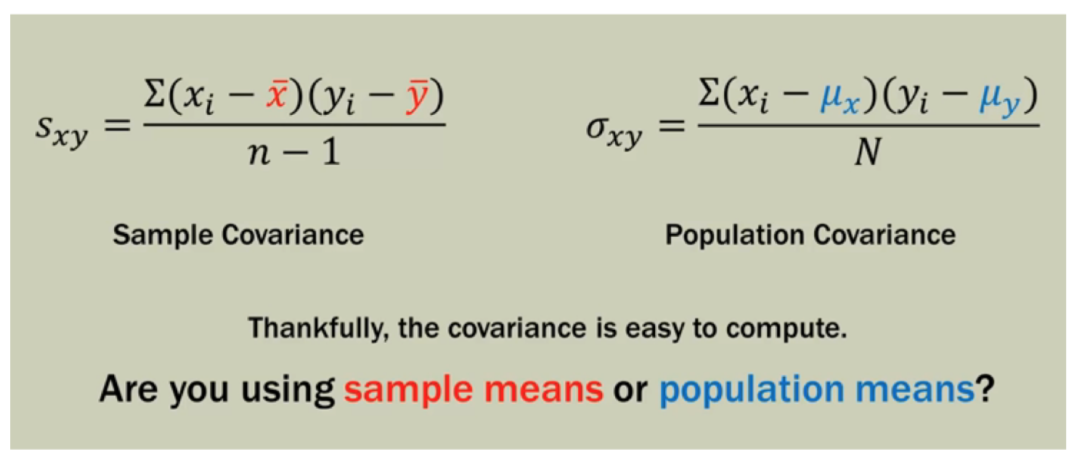
Ratio is linear, like age. You can be 12 one year, 13 the next, but not 11 the next.

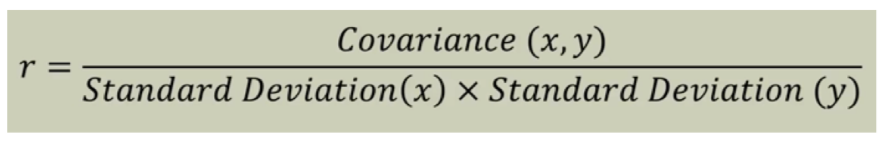
* (a) Commute times (40 mins, 20 mins, 15 mins): Ratio
* (b) Types of Floorings (Hardwood, Tile, laminate) : Nominal
* (c) Household income levels (50k, 35k, 22k): Interval
* (d) Doneness of meat (Rare, Medium, Well-done): Ordinal

**SECTION B : (Sample Questions)**

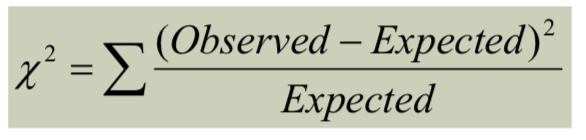
1. Problems on computing Covariance and Correlation.

Covariance:



Correlation: 

1. Problems on Chi-Square Tests of independence.



1. What is the function of Data Transformation? And carry out the min-max normalization on the following salary field of the data with a new min of 15 and a new max of 20. (Lec 4 part 3, page 3)

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Gender** | **Age** | **Salary** |
| 1 | F | 27 | 19,000 |
| 2 | M | 51 | 64,000 |
| 3 | M | 52 | 100,000 |
| 4 | F | 33 | 55,000 |
| 5 | M | 45 | 45,000 |

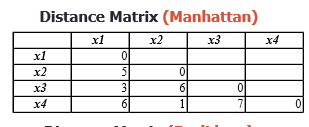
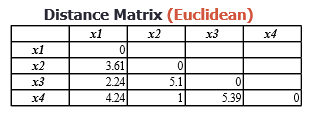
Data transformation is a mapping function that maps the entire set of values of a given feature to a new set of replacement values. NOTE: Each old value will be uniquely identified with a new value. Some methods of data transformation include normalization (scaled to fall within a smaller, specified range: min-max and z-score), smoothing (removing noise from data using discretization), and concept hierarchies(AKA Feature Engineering, new features or attributes are constructed from known ones). (Lecture 4 part 3, page 2)

|  |  |
| --- | --- |
| **Salary** | **Min-Max Normalization** |
| 19,000 | 15 |
| 64,000 | 17.778 |
| 100,000 | 20 |
| 55,000 | 17.222 |
| 45,000 | 16.605 |



Given the following data matrix (Lec 6, page 16)

|  |  |  |
| --- | --- | --- |
| **point** | **attribute1** | **attribute2** |
| ***x1*** | 1 | 2 |
| ***x2*** | 3 | 5 |
| ***x3*** | 2 | 0 |
| ***x4*** | 4 | 5 |

* a) Compute the Manhattan Distance Matrix
* 
* b) Compute the Euclidean Distance Matrix.
* 