UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 2
Chapter 2a - Data
Chapter 2b - Data Exploration

1. 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.

   Example: Age in years. Answer: Discrete, quantitative, ratio

   a) Time in terms of "AM" or "PM".

   b) Brightness as measured by a light meter.

   c) Brightness as measured by people's judgments.

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

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

   f) Height above sea level.

   g) Number of patients in a hospital.

   h) ISBN numbers for books. (Look up the format on the Web.)

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

   j) Military rank.

   k) Distance from the center of campus.

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

   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.)

2. *Proximity* (nearness) is typically defined between a pair of objects. Proximity may be measured in terms of how alike objects are to one another (object similarity) or how unlike they are (object dissimilarity). The term distance measure is often used instead of dissimilarity measure.

   a) Define two ways in which you might define the proximity among a group of objects.

   b) How might you define the *distance* between two sets of points in Euclidean space?

   c) How might you define the *proximity* between two sets of data objects? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

3. Suppose that the data for analysis includes the attribute *age*. The age values for the data tuples are (in increasing order) :
   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

   a) What is the arithmetic *mean* of the data? What is the *median*?

   b) What is the *mode* of the data? Comment on the data's modality (i.e., *bimodal*, trimodal, etc.).

   c) What is the *midrange* of the data?

   d) Can you find (roughly) the *first quartile (Q1)* (lower quartile) and the *third quartile (Q3)* (upper quartile) of the data?

   e) Give the *five-number summary* of the data.

4. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result:

age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

   a) Calculate the *mean*, *median* and *standard deviation* of age and %fat.

   b) Draw the *boxplots* for age and %fat.

   c) Draw a *scatter plot* and a quantile-quantile / *q-q plot* based on these two variables.

   d) Calculate the *Pearson correlation coefficient*. Are these two variables positively or negatively correlated?

5. Construct a *data cube* from Table 3.1 below. Is this a *dense* or *sparse* data cube? If it is sparse, identify the cells that are empty.

Table 3.1. Fact table

Product ID	Location ID	Number Sold
1	1	10
1	3	6
2	1	5
2	2	22

6. This exercise compares and contrasts some *similarity* and *distance measures*.

   a) For binary data, the L1 distance corresponds to the *Hamming distance*; that is, the number of bits that are different between two binary vectors. The *Jaccard similarity* is a measure of the similarity between two binary vectors.

   Compute the *Hamming distance* and the *Jaccard similarity* between the following two binary vectors.

   b) Which approach, Jaccard or Hamming distance, is more similar to the *Simple Matching Coefficient (SMC)*, and which approach is more similar to the *cosine measure*? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

   c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

   d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share > 99.9% of the same genes.)

7. For the following vectors, x and y, calculate the indicated similarity or distance measures.

   a) x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlation, Euclidean
   b) x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlation, Euclidean, Jaccard
   c) x = (0,−1, 0, 1), y = (1, 0,−1, 0) cosine, correlation, Euclidean
   d) x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlation, Jaccard
   e) x = (2,−1, 0, 2, 0,−3), y = (−1, 1,−1, 0, 0,−1) cosine, correlation

Discussion Questions

1. *Proximity* is typically defined between a pair of objects.

   a) Define two ways in which you might define the proximity among a group of objects.
   b) How might you define the distance between two sets of points in Euclidean space?
   c) How might you define the proximity between two sets of data objects?

2. Describe how you would create *visualizations* to display information that describes the following types of systems.

   Be sure to address the following issues:

   - **Representation**. How will you map objects, attributes, and relationships to visual elements?
   - **Arrangement**. Are there any special considerations that need to be taken into account with respect to how visual elements are displayed? Specific examples might be the choice of viewpoint, the use of transparency, or the separation of certain groups of objects.
   - **Selection**. How will you handle a large number of attributes and data objects?

   a) Computer networks. Be sure to include both the static aspects of the network, such as *connectivity*, and the dynamic aspects, such as *traffic*.

   b) The distribution of specific plant and animal species around the world for a specific moment in time.

   c) The use of computer resources, such as processor time, main memory, and disk, for a set of benchmark database programs.

   d) The change in occupation of workers in a particular country over the last thirty years. Assume that you have yearly information about each person that also includes gender and level of education.

3. Explain why computing the *proximity* (distance) between two attributes is often simpler than computing the *similarity* between two objects.

4. Discuss the differences between dimensionality reduction based on aggregation and dimensionality reduction based on techniques such as principal component analysis (PCA) and singular value decomposition (SVD).

The End