Computer Science and Networking, School of Computing and Data Science, Wentworth Institute of Technology

Data Mining

Dr. Salem Othman

Summer 2023

Assignment 3: 2.6 exercises

Deliverables:

1. Answer all the following (2.6 exercises) problems and submit PDF/DOCX file.

Problems:

1. In the initial example of **Chapter 2**, the statistician says, "Yes, fields 2 and 3 are basically the same." Can you tell from the three lines of sample data that are shown why she says that?

2. 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.)

3. You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our bestselling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

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?

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

4. A few months later, you are again approached by the same marketing director as in **Exercise 3**. This time, he has devised a better approach to measure the extent to which a customer prefers one product over other similar products. He explains, "When we develop new products, we typically create several variations and evaluate which one customers prefer. Our standard procedure is to give our test subjects all of the product variations at one time and then ask them to rank the product variations in order of preference. However, our test subjects are very indecisive, especially when there are more than two products. As a result, testing takes forever. I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations 1 and 2, then 2 and 3, and finally 3 and 1. Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest product evaluations, yesterday. I should also mention that he was the person who came up with the old product evaluation approach. Can you help me?"

a. Is the marketing director in trouble? Will his approach work for generating an ordinal ranking of the product variations in terms of customer preference? Explain.

b. Is there a way to fix the marketing director's approach? More generally, what can you say about trying to create an ordinal measurement scale based on pairwise comparisons?

c. For the original product evaluation scheme, the overall rankings of each product variation are found by computing its average over all test subjects. Comment on whether you think that this is a reasonable approach. What other approaches might you take?

5. Can you think of a situation in which identification numbers would be useful for prediction?

6. An educational psychologist wants to use association analysis to analyze test results. The test consists of 100 questions with four possible answers each.

a. How would you convert this data into a form suitable for association analysis?

b. In particular, what type of attributes would you have and how many of them are there?

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

8. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.

9. Many sciences rely on observation instead of (or in addition to) designed experiments. Compare the data quality issues involved in observational science with those of experimental science and data mining.

10. Discuss the difference between the precision of a measurement and the terms single and double precision, as they are used in computer science, typically to represent floating-point numbers that require 32 and 64 bits, respectively.

11. Give at least two advantages to working with data stored in text files instead of in a binary format.

12. Distinguish between noise and outliers. Be sure to consider the following questions.

a. Is noise ever interesting or desirable? Outliers?

b. Can noise objects be outliers?

c. Are noise objects always outliers?

d. Are outliers always noise objects?

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

Algorithm 2.3 Algorithm for finding *k*-nearest neighbors.

1 : **for** $i = 1$ to *number of data objects* **do**

2 : Find the distances of the i^{th} object to all other objects.

3: Sort these distances in decreasing order.

(Keep track of which object is associated with each distance.)

4: **return** the objects associated with the first *k* distances of the sorted list

5: **end for**

13. Consider the problem of finding the *K*-nearest neighbors of a data object. A programmer designs Algorithm 2.3 for this task.

a. Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will return a distance of 0 only for objects that are the same.

b. How would you fix this problem?

14. The following attributes are measured for members of a herd of Asian elephants: *weight, height, tusk length, trunk length,* and *ear area*. Based on these measurements, what sort of proximity measure from Section 2.4 would you use to compare or group these elephants? Justify your answer and explain any special circumstances.

15. You are given a set of *m* objects that is divided into *k* groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

a. We randomly select $n \times m_i/m$ elements from each group.

b. We randomly select *n* elements from the data set, without regard for the group to which an object belongs.

16. Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and *m* is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} \times \log \frac{m}{df_i}, \tag{2.31}$$

where df_i is the number of documents in which the i^{th} term appears, which is known as the **document frequency** of the term. This transformation is known as the **inverse document frequency** transformation.

a. What is the effect of this transformation if a term occurs in one document? In every document?

b. What might be the purpose of this transformation?

17. Assume that we apply a square root transformation to a ratio attribute *x* to obtain the new attribute x^*. As part of your analysis, you identify an interval (*a*, *b*) in which x^* has a linear relationship to another attribute*y*.

a. What is the corresponding interval (*A*, *B*) in terms of *x* ?

b. Give an equation that relates *y* to *x*.

18. 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.

$$\mathbf{x} = 0101010001$$
$$\mathbf{y} = 0100011000$$

b. Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, 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.)

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

a. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$ cosine, correlation, Euclidean

b. $\mathbf{x} = (0, 1, 0, 1)$, $\mathbf{y} = (1, 0, 1, 0)$ cosine, correlation, Euclidean, Jaccard

c. $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean

d. $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard

e. $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0, -1)$ cosine, correlation

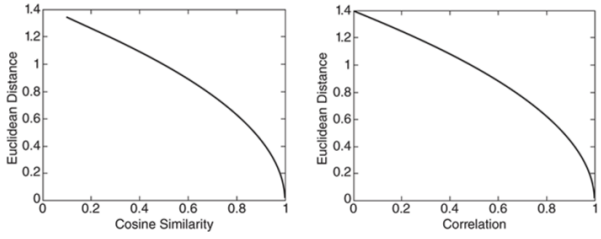
20. Here, we further explore the cosine and correlation measures.

a. What is the range of values possible for the cosine measure?

b. If two objects have a cosine measure of 1, are they identical? Explain.

c. What is the relationship of the cosine measure to correlation, if any? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

d. Figure 2.22(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?



(a) Relationship between Euclidean distance and the cosine measure.

(b) Relationship between Euclidean distance and correlation.

Figure 2.22.
Graphs for Exercise 20.

e. Figure 2.22(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

f. Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L_2 length of 1.

g. Derive the mathematical relationship between correlation and Euclidean distance when each data point has been been standardized by subtracting its mean and dividing by its standard deviation.

21. Show that the set difference metric given by

$$d(A,\ B) = size(A - B) + size(B - A) \tag{2.32}$$

satisfies the metric axioms given on page 77. *A* and *B* are sets and $A - B$ is the set difference.

22. Discuss how you might map correlation values from the interval $[-1,\ 1]$ to the interval [0, 1]. Note that the type of transformation that you use might depend on the application that you have in mind. Thus, consider two applications: clustering time series and predicting the behavior of one time series given another.

23. Given a similarity measure with values in the interval [0, 1], describe two ways to transform this similarity value into a dissimilarity value in the interval $[0,\ \infty]$.

24. 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? (Make no assumption about the data objects, except that a proximity measure is defined between any pair of objects.)

25. You are given a set of points *s* in Euclidean space, as well as the distance of each point in *s* to a point **x**. (It does not matter if $\mathbf{x} \in S$.)

a. If the goal is to find all points within a specified distance ε of point $\mathbf{y},\ \mathbf{y} \neq \mathbf{x}$, explain how you could use the triangle inequality and the already calculated distances to **x** to potentially reduce the number of distance calculations necessary? Hint: The triangle inequality, $d(\mathbf{x},\ \mathbf{z}) \leq d(\mathbf{x},\ \mathbf{y}) + d(\mathbf{y},\ \mathbf{x})$, can be rewritten as $d(\mathbf{x},\ \mathbf{y}) \geq d(\mathbf{x},\ \mathbf{z}) - d(\mathbf{y},\ \mathbf{z})$.

b. In general, how would the distance between **x** and **y** affect the number of distance calculations?

c. Suppose that you can find a small subset of points $S\prime$, from the original data set, such that every point in the data set is within a specified distance ε of at least one of the points in $S\prime$, and that you also have the pairwise distance matrix for $S\prime$. Describe a technique that uses this information to compute, with a minimum of distance calculations, the set of all points within a distance of β of a specified point from the data set.

26. Show that 1 minus the Jaccard similarity is a distance measure between two data objects, **x** and **y**, that satisfies the metric axioms given on page 77. Specifically, $d(\mathbf{x},\ \mathbf{y}) = 1 - \mathbf{J}(\mathbf{x},\ \mathbf{y})$.

27. Show that the distance measure defined as the angle between two data vectors, **x** and **y**, satisfies the metric axioms given on page 77. Specifically, $d(\mathbf{x},\ \mathbf{y}) = \arccos(\cos(\mathbf{x},\ \mathbf{y}))$.

28. Explain why computing the proximity between two attributes is often simpler than computing the similarity between two objects.