B565-Data Mining
Homework #1

Due on Monday, Jan 31, 2023 08:00 p.m.

Dr. H. Kurban

Student Name

January 24, 2023

Problem 1

The following problems have to do with metrics. In each case, prove or disprove the distance is a metric ($\mathbb{R}$ is the set of reals, and $\|X\|$ is the size of a finite set X.)

(a) Let $X \subset \mathbb{R}^n$ for positive integer $n > 0$. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = max\{|x_i - y_i|\}, \forall i, 1 \leq i \leq n.$$

(b) Let $c : \mathbb{R}^{2n} \to \mathbb{R}_{\geq 0}$ be defined as

$$c(x, y) = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{o.w.} \end{cases}$$

Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as

$$d(x, y) = \sum_{i}^{n} \frac{c(x_i, y_i)}{i}, \forall i \ 1 \leq i \leq n$$

(c) Suppose $d_0, \ d_1$ are metrics.

i. $d_0 \times d_1$

ii. $(d_0 + d_1)/(d_0 d_1)$

iii. $max\{d_0, \ d_1\}$

iv. Let X be a finite set. Define a distance $d : X \times X \to \mathbb{R}_{\geq 0}$ as $d(x, y) = \frac{||x \ \cap \ y||}{||x \ \cup \ y||+1}$

Problem 2

Curse of Dimensionality: Generate m-dimensional n data points from a uniform distribution with values between 0 and 1. For an arbitrary m value

$$f(m) = \log_{10} \frac{d_{max}(m) - d_{min}(m)}{d_{min}(m)}$$

where $d_{max}(m)$ and $d_{min}(m)$ are maximum and minimum distances between any pair of points, respectively. Let m take each value from $\{1, 2, \ldots, 99, 100\}$. Repeat each experiment multiple times to get stable values by averaging the quantities over multiple runs for each m. For four different n values, e.g., $n \in \{150, 1500, 15000, 150000\}$, plot $f(m)$. Use Euclidean as your distance metric. Label and scale each axis properly and discuss your observations over different n's.

R or Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

Discussion of Experiments

Answer here...

Plot/s

Place images here with suitable captions.

Problem 3

For the following data, give the best taxonomic type (interval, ratio, nominal, ordinal):

1. A section of highway on a map.
2. The value of a stock.
3. The weight of a person.
4. Marital status.
5. Visiting United Airlines (https://www.united.com) the seating is: Economomy, Economy plus, and United Business.

Problem 4

You are datamining with a column that has physical addresses in some city with the same zipcode. For example,

```
55 WEST CIR
2131 South Creek Road
Apt. #1 Fountain Park
1114 Rosewood Cir
1114 Rosewood Ct.
1114 Rosewood Drive
```

What structure would you create to mine these? What questions do you think you should be able to answer?

Problem 5

For this problem you will be using a data set with total 81 attributes describing every aspect of residential homes of Ames, Iowa. You can download the data from here [link]. The downloadable file already comes with the corresponding names of the attributes. Also a document describing the data is available here [link].

Discussion of Data

Briefly describe this data set–what is its purpose? How should it be used? What are the kinds of data it's using?

R/Python Code

Using R/Python, show code that answers the following questions:

1. How many entries are in the data set? Write the R or Python code in the box below.

Problem 5 continued on next page. . .

R/Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

2. How many unknown or missing data are in the data set? Write the R or Python code in the box below.

R/Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

3. Find 10 attributes influencing the target attribute `SalePrice`. Use coherent plotting methods to describe and discuss their relation with `SalePrice`. Place images of these plots into the document. Write the R or Python code in the box below.

R/Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

Discussion of Findings

Answer here...

Plot/s

Place images here with suitable captions.

4. Make a histogram/bar plot for each of those 10 attributes influencing `SalePrice` and discuss the distribution of values, *e.g.*, are uniform, skewed, normal of those attributes. Place images of these histograms into the document. Show the R/Python code that you used below and discussion below that.

R/Python script

```
# Sample R Script With Highlighting
```

```
# Sample Python Script With Highlighting
```

Discussion of Attributes

Answer here...

Problem 5 continued on next page...

Histograms/Bar Plots

Place images here with suitable captions.

Discussion of simply removing tuples

Quantify the affect of simply removing the tuples with unknown or missing values. What is the cost in human capital?

Problem 6

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

1. Is noise ever interesting or desirable? Outliers?

2. Can noise objects be outliers?

3. Are noise objects always outliers?

4. Are outliers always noise objects?

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

Problem 7

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 sampling schemes? (Assume sampling with replacement.)

1. We randomly select $n * m_i/m$ elements from each group.

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

Problem 8

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

$$tf_{ij}^{'} = tf_{ij} * \log \frac{m}{df_i},$$

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 inverse document frequency transformation.

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

2. What might be the purpose of this transformation?

Problem 9

This question compares and contrasts some similarity and distance measures.

1. 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.
   - $\mathbf{x} = 0101010001$
   - $\mathbf{y} = 0100011000$
2. 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 distance is a distance, while the other three measures are similarities, but don't let this confuse you.)
3. 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. (Note: 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.)
4. 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: Two human beings share $> 99.9\%$ of the same genes.)

Problem 10

For the following vectors, $\mathbf{x}$ and $\mathbf{y}$, calculate the indicated similarity and distance measures. Show detailed calculations/steps.

1. $\mathbf{x} = (1, 1, 1, 1)$, $\mathbf{y} = (2, 2, 2, 2)$ cosine, correlation, Euclidean.
2. $\mathbf{x} = (0, 1, 0, 1)$, $\mathbf{y} = (1, 0, 1, 0)$ cosine, correlation, Euclidean, Jaccard.
3. $\mathbf{x} = (0, -1, 0, 1)$, $\mathbf{y} = (1, 0, -1, 0)$ cosine, correlation, Euclidean.
4. $\mathbf{x} = (1, 1, 0, 1, 0, 1)$, $\mathbf{y} = (1, 1, 1, 0, 0, 1)$ cosine, correlation, Jaccard.
5. $\mathbf{x} = (2, -1, 0, 2, 0, -3)$, $\mathbf{y} = (-1, 1, -1, 0, 0 - 1)$ cosine, correlation.

Submission

You must use LaTeX to turn in your assignments. Please submit the following two files via Canvas:

1. A .pdf with the name `yourname-hw1-everything.pdf` which you will get after compiling your .tex file.
2. A .zip file with the name `yourname-hw1.zip` which should contain your .tex, .pdf, codes(.py or .R), and a README file. The README file should contain information about dependencies and how to run your codes.