Computer Science
Data Mining Core
Homework II Fall 2013
Indiana University
Bloomington, IN

Mehmet M. Dalkilic

September 11, 2013

Definitions

$\mathcal{R}$ is the set of reals

The size of a finite set X is written $||X||$.

1. The following problems have to do with metrics. In each case, prove or disprove the distance is a metric.

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

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

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

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

   Define a distance $d : X \times X \rightarrow R_{\geq 0}$ as

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

   (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 $d : X \times X \rightarrow R_{\geq 0}$ as

   $$d(x, y) \quad = \quad \frac{||x \cap y||}{||x \cup y|| + 1}$$

Foo			
X	Y	Z	A
1	1	Y	abcd
3	255	N	bcde
4	4	N	bcd
2	1	Y	acde
20	1	Y	bdf
5	4	T	fg
5	3	Y	abf

Figure 1: An instance Foo. The domains of X, Y are natural numbers. The domain of Z is Boolean. The domain of A is set of characters $\{a, b, c, d, e, f\}$.

2. Consider the relation in Fig. 1. Partition this data into three blocks using exactly three attributes (or features). For attributes X, Y, use L_2. For A use Jaccard Index. For attribute Z you are free to pick a metric. The table has not been cleaned nor transformed.

3. From Everitt, exercise 2.3,2.4.

4. We have provided data from the FHA, "FHA Single Family Loan Performance Trends, Credit Risk Report," June 2013.

   (a) Put the table *Share By Reason for Delinquency in Percent* into an `R` data frame.

      i. Plot *Reduction of Income* against *Unemployed*. Discuss the results.

      ii. Plot *Death* against *Unemployed*. Discuss the results.

   (b) Put the subtable *Credit Score Range* in the *Delinquency Rates* into an `R` data frame.

      i. Using `R` find the number of loans with credit scores less than 620.

      ii. Of these loans, how many are past due?

      iii. Plot *Credit Score* against *All Past Due*

5. From Tan, exercises Chapter 2: 2,3,6,12,13,16,19,24.