CIS 635 Data Mining

Homework 2

Objectives

- reinforce concepts related to data
- practice the concepts of linear algebra
- explore the R programming language

Instructions

Part A – concepts

1. categorize the following attributes as nominal, ordinal or numeric: age, gender, weather (sunny, cloudy, rain, snow), shirtSize (large, medium, small)
2. describe the following data sets (dimensionality and sparsity, either low or high)
   a) list of colleges in the US with the programs offered (each row is a college and each column is a program – there are 400 possible programs, 0=not available, 1=available)
   b) list of students at GVSU with their GPA, IQ, SAT score, age and number of credits completed)
   c) books of Charles Dickens (0/1 matrix where each book is a row and each word is a column)
3. indicate the type of data set for the following (record, graph, transaction data or ordered data)
   a) stock quotes data
   b) customer list with yearly sales numbers
   c) FaceBook data with friendship links
   d) grocery store receipts
4. answer the following questions about noise and outliers:
   a) can an object be an outlier but not noise?
   b) can an object be noise but not an outlier?
   c) if my class list included a student with a GPA of -7.12 would that be noise or an outlier?
5. what would you suggest for the following data sets (1000 objects, 20 attributes) with missing data (eliminate the objects, eliminate the attribute, estimate the missing values):
   a) 4 missing values in 1 attribute
   b) 950 missing values in 1 attribute
   c) 100 missing values in 1 attribute
6. how do the organizers of the Irish Jig aggregate the age by using categories in the results (https://runsignup.com/Race/Results/27515/#resultSetId-107651;perpage:10) ?

Part B – linear algebra

Given the vector $v = [5 \quad 6 \quad -4 \quad 7]$ and the matrix $A = \begin{bmatrix} 1 & 2 & 3 \\ 6 & 5 & 4 \\ 12 & 11 & 10 \end{bmatrix}$, calculate the following expressions:

1. $v + 3$
2. $v / 3$
3. $v \cdot v^T$
4. $A \cdot 2$
5. A^T
6. $A \cdot A$

Part C – R

Follow the instructions below in R.

1. create a vector called a with the values 4, 7, 2, 10, 5
2. create a vector called b with the values 1, 0, 0, 1, 1
3. calculate the element by element multiplication of a times b
4. calculate the dot product a · b
5. read the file hw02data.txt into the matrix x
6. show just the value of x at row 25, column 3
7. show just first row of x
8. show just rows 1 through 10 of the second and third columns sideways (transpose)
9. calculate the column means of all rows, and columns 2 through 5 using colMeans
10. calculate the column means of columns 2 through 5 using colMeans for only those observations where column 6 is n
11. calculate the column means of columns 2 through 5 using colMeans for only those observations where column 6 is y
12. create a vector called *ind* of length nrow(x) that contains a random sequence of the numbers 0-2 (hint, use sample and modulo)
13. using ind, calculate the average weight of all the records where ind==1