CS 4232/5232 – Fall 2023 – Homework 2
Assigned: 10/14/2023
Due: 10/20/2023 at 11:59 p.m.
Maximum Grade: 100 pts.

Objectives: The objectives of this homework are the following:

- Learn about the different similarity measures.
- Learn how to impute missing data.
- Learn about the different categories of missing data problems.

Instructions:

- This homework is to be done individually.
- Submit **a single HTML file** named: yourlastname_hw2.html **and also the source R markdown file** (with a suffix of .Rmd) with the answers to your homework. To generate this HTML file, you **must use R markdown**. Check the video here for more information https://rmarkdown.rstudio.com/lesson-1.html.

Activity 1: (10 pts.) (Similarity measures) Given the vectors *x* and *y* as below, compute by hand the indicated similarity or distance measure. Show all your work. You can do this exercise by hand, then scan your solution and include it in your markdown file:

a) (2 pts.) $x = (1,0,1,1,0,0)$ and $y = (1,0,1,0,1,1)$, SMC and Jaccard coefficient
b) (3 pts.) $x = (-1,9,3,2)$ and $y = (10,6,2,4)$, Euclidean distance, cosine similarity, and correlation
c) (5 pts.) $x = (4,-1)$ and $y = (-1,6)$, Mahalanobis distance, where the covariance matrix is:
$\Sigma = \begin{pmatrix} 12 & 2 \\ 2 & 25 \end{pmatrix}$.

Activity 2: (30 pts.) (Similarity measures) Given the adult dataset (https://archive.ics.uci.edu/dataset/2/adult), perform the following activities:

a) (5 pts.) Design a dissimilarity function such that, given two rows of this dataset as input, it returns a real number indicating how different the rows are. This question does not ask to implement the function; it asks you to write down this mathematical function in your markdown.
b) (5 pts.) Explain in detail the thought process that led to the creation of this dissimilarity function. Why is this function you proposed reasonable? Why did you choose the coefficients you chose?
c) (5 pts.) Based on your dissimilarity function of part a, design a similarity function, following what is indicated in slide 49 of Topic 2.
d) (5 pts.) Explain in detail the thought process that led to the creation of this similarity function.
e) (3 pts.) Implement your dissimilarity function of part a in R.
f) (3 pts.) Implement your similarity function of part c in R.
g) (4 pts.) Write code in R to compute the similarity between every pair of rows in the adult dataset.

Activity 3: (20 pts.) (Missing data imputation with kNN in Caret) Do the following tasks in R:

a) (1 pt.) Type help(iris) at the R command line and explain this dataset in your own words. What are its attributes and what are their meanings? Display the first ten rows of iris.

b) (1 pt.) Let's introduce some missing data into this dataset by running

```
iris.missing <- iris[,-5]
iris.missing[1:4,1] <- NA
```

Display the first 10 rows of the resulting data frame.

c) (1 pts.) Run the `preProcess` function of the Caret package, providing the arguments k=3 and method = "knnImpute". This will build a model for k-nearest neighbor missing data imputation using $k = 3$. After this, the model is built, but the data is not yet imputed. For more information about this, see http://rismyhammer.com/ml/Pre-Processing.html#imputation.

d) (1 pts.) Run the *predict* function using the model obtained in step c and find the imputations for the missing data. Print out the first ten rows of the resulting data frame.

e) (2 pts.) Notice that *preProcess* and *predict* normalized all columns. Undo this transformation by multiplying every quantitative attribute by its standard deviation and adding its mean. You can find these values in the fields of the output of *preProcess*.

f) (4 pts.) After unnormalizing, notice that *preProcess* also modified the Sepal.Width and the other attributes, which is an undesirable effect. Take only the imputed values for Sepal.Length and copy them —and only them—onto `iris.missing`. Then, subtract the resulting data frame from the original iris. Make sure that the two data frames have the same size; you might need to drop an attribute to do the subtraction. Then, compute the average relative error of the imputation $\frac{1}{4}\sum_{i=1}^{4}\frac{|true\ value_i - imputed\ value_i|}{|true\ value_i|}$, where *true value*$_i$ is the Sepal.Length of the ith row, and the imputed value is the one produced by the procedure above. This measures the error of this procedure.

g) (4 pts.) Make a plot with k as the variable on the x-axis and the total relative error on the y-axis. Your plot must have labeled axes and a title. Choose values of $k = 1$, 2,…,8. To obtain this plot, you have to repeat steps c–f, once per value of k.

h) (6 pts.) What do you observe in your plot? What conclusions do you reach from it? Explain in depth.

Activity 4: (40 pts.) (Missing Data Imputation) Perform the following tasks using R:

a) (9 pts.) Read online about the following three categories of missing data problems: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Explain them **in your own words,** giving **your own** detailed example of each. Provide abundant detail; to get full credit, the example must be amply explained.

b) (4 pts.) In which of the three categories above can we drop entire rows containing at least a single missing datum without negatively biasing our dataset? Why? Explain in detail.

c) (4 pts.) Write R code to generate a synthetic MCAR dataset. You are free to choose the attributes and the number of rows.
d) (4 pts.) Explain in detail why the dataset in c) is MCAR.
e) (3 pts.) Write R code to generate a synthetic MAR dataset. You are free to choose the attributes and the number of rows.
f) (4 pts.) Explain in detail why the dataset in e) is MCAR.
g) (1 pt.) Run the following commands in R and show the summary of the resulting data frame

```
data <- airquality
data[4:10,3] <- NA
data[1:5,4] <- NA
```

h) (2 pts.) Run the command

```
aggr(data, numbers=TRUE, labels=names(data), ylab=c("Missing data Histogram","Missing Data (in Red) Pattern"))
```

You will need to install the VIM package for this. The output of the right-hand side shows the pattern of missing data; this figure is read bottom-up: 68% of the data is complete (blue), 22.2% of the data is missing the Ozone (red square), 1.9% of the data is missing Temp. What percentage of the data is missing Solar.R and Wind? How did you deduce this from the right-hand side figure?
i) (1 pts.) Install the MICE package (for missing data imputation) and run the mice function on `data` frame with meth = 'pmm', seed = 500 to impute the missing data.
j) (6 pts.) Run the command xyplot(tempData,Ozone ~ Wind+Temp+Solar.R. This will plot a scatter plot of Ozone versus each of the other attributes, having in blue the points of the observed data and in red those of the imputed data. What conclusions do you draw from this figure? Explain in depth.
k) (2 pts.) Read about the `mice` function. What assumption does it make about the missing data? Does it assume it is MCAR, MAR, or MNAR? Write down your source and quote it.

Important Notes:

- **You cannot have all your code in a single R markdown chunk**. You need to break down your markdown file into meaningful chunks. For example, one chunk can be a function definition, another a function call, another the definition of a set of related variables, etc.
- **You must not print an entire data frame, especially if it is large**. You only print a few rows at a time.