

|  |
| --- |
| Assignment #3: Multidimensional Scaling and Self-Organizing Maps  *MSDS 411* |

# This assignment has two components parts, one for each of these methods. They use different data sets. Please label your write-up appropriately to keep these analyses separate. The components are worth equivalent number of points.

# COMPONENT 1: Multidimensional Scaling

# In this component of the assignment, you will be conducting MDS analyses.

# Data: The RECIDIVISM dataset is an 18 variable dataset with n=1445 records. Please see the data description file for the variable definitions and additional information about the dataset. The data consists of a random sample records on convicts released from prison during 1977/1978.

## Assignment Tasks

1. Perform a basic Exploratory Data Analysis on the Recidivism data. Report what you have learned through this activity. Prepare the data as best you can for an upcoming MDS analysis.
2. Obtain a dissimilarity matrix using Euclidean Distances. There are a lot of cells in this matrix, but can you see any patterns at this point?
3. Conduct a classical multidimensional scaling using the Euclidean Distances dissimilarity matrix. Graph a 2-dimensional solution and interpret the result.
4. Conduct 2 similar analyses using nonmetric scaling and Ramsey’s method. Graph and interpret the two dimensional solutions. How do these solutions compare with the classical approach?

# COMPONENT 2: Self Organizing Maps

# In this component of the assignment you will fit Self-organizing Map [SOM] models.

# Data: The data for this assignment is the college acceptance data set. The dataset contains information for college acceptance into various engineering programs for 400 students. A simple data dictionary for the dataset is as follows:

* admit (binary) 0 = Not admitted, 1 = Admitted
* gre (numeric) Student’s GRE score
* gpa (numeric) Student’s GPA
* rank (numeric) College ranking

## Assignment Tasks

1. Exploratory Data Analysis [EDA] and Data Preparation for the College Acceptance Data.

* Perform EDA on the data set and report your findings.
* Prepare the dataset for modeling as appropriate. Should scaling or normalization be applied? Why or why not?
* Use only the variables provided in the dataset or variables you create by modifying or combining the variables provided.

1. Fit the SOM model. In the process you need to:

* Determine and report the number of epochs that will be used to train the model.
* Determine the appropriate grid size for the SOM. Report the method that you used.
* Fit the model using the R *kohonen* package or similar to the dataset that you prepared in *PART A*. Use the grid size and epochs that you selected in 1 and 2. Be sure to set the seed before fitting the model so that the results may be reproduced.

1. Evaluate the SOM model. To do this you need to address the following:

* Was the epochs value selected in PART B adequate to train the model? Include a copy of the visualization that was used to make that determination. If the model needs additional training, adjust the epochs value and retrain the model before continuing.
* Was the grid size selected in PART B adequate? Explain why the grid size was or was not adequate and attach the visualizations used to make that determination.
* What is the average number of observations assigned to the nodes?
* Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?
* Generate a *codes* plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.

1. Experiment with the SOM model. To accomplish this task, you will need to:

* Change the grid size for the SOM and retrain the model. Discuss whether you increased or decreased the grid size and why.
* Compare this new SOM to the SOM created in PART B. Does the new grid size improve the SOM? Discuss how grid size impacts the SOM.
* Generate a distance map and attach a copy of it here. Are any nodes quite distant from their neighbors?
* Generate a *codes* plot and attach a copy of it. Discuss what this plot tells us about the applications and college acceptance.

1. Please write a reflection on your MDS and SOM modeling experiences.

# Assignment Document:

All assignment reports should answer each of the questions separately. Please be sure to clearly indicate which question is being addressed. Results should be presented and discussed in an organized manner with the discussion in close proximity of the results. The report should not contain unnecessary R-code, intermediary computations, R-results, or non-essential information. The document should be submitted in pdf format. Name your file Assign3\_LastName.pdf.