Assignment 5: Self-Organizing Maps

Instructions

This assignment will be conducted in Matlab, exceptions are only allowed for those who already use the method in other languages and it should be noted that those who use a different programming language will have significantly less resources.

The code and data necessary for this lab are located at: `/ulteosrv1/s0/meteo515/data` You should have read and execute permissions in this directory. For this assignment, access the code and data in this folder instead of copying it to a different system.

Data: The datasets that will be used in this lab are the propagating sine waves that were used in Assignment 4. The data can stay in this folder and the scripts I have given you will access them.

Code: Two scripts are provided that are either entirely or nearly complete, somcreate_computations.m and Assignment5_Q1.m. To complete this assignment, you will be making modifications to the copy of Assignment5.m that you have place in your own directory. These two pieces of code can be used as a base for the more complex SOM analysis you will be conducting for your final projects.

Software: the Matlab SOM toolbox is located at `/ulteosrv1/s0/meteo515/data/somtoolbox` and the SOMPAK program is located at `/ulteosrv1/s0/meteo515/data/SOM_PAK3.2_MG`, you will have read and execute permissions only for this software. Both of these code repositories were originally from the Laboratory of Computer and Informational Science (CIS) Adaptive Research Center obtain at http://www.cis.hut.fi/research/software. Modification have been made by Melissa Gervais to the SOMPAK program in order to include the Epanechikov neighborhood function. Modifications have also been made to the Matlab SOM tool box, and all modified files in this tool box have MG in their title. Documentation for this software has been provided in a folder called SOM Documentation Materials. There is also a copy of Gervais et al. 2016 whose methodology section may be a useful resource for understanding the method.

Assignment Set-up: Create a folder for this lab within your home directory and copy somcreate_computations.m and Assignment5_Q1.m into this directory. Inside this lab directory, create a folder called som and within som generate folders called data and figures. This is where the data and the figures for each of the exercises will be outputted to. When the SOM analysis is conducted a new folder will be created in each of data and figures with output from the SOM analysis.

In this assignment, you will be conducting a self-organizing map (SOM) analysis of the same 5 different batches of test data as was used in Assignment 4. The goal is to explore how SOM analysis identifies patterns of variability in various datasets and the sensitivity of the analysis to user defined parameters.

1. Running a SOM

- Read through the Assignment5.m and somcreate_sompak.m code to understand each of the main steps
- **Add code in the designated location in Assignment5.m** that loads in the sine_wave_data1.nc from Assignment 4. Create a string variable data_name (ex. data_name= 'sine_example1'). This is used for saving purposes when conducting different experiments and so should reflect which experimental data is being used (ie. 1,2,3,4, or 5). Create a second string variable data_name_title (ex. data_name_title = 'Sine Example 1'). This is used in the titles of some automatically generated plots to distinguish the datasets used and should similarly change depending on the experimental data being used.
- Run the Assignment5.m code.
- A figure with a Sammon map should be generated that represents the relationship between each of the map nodes.
- Four figures demonstrating the progress during SOM training should be generated the quantization error and the topological error through training time for each training period. These will all be saved in the ./figures/exercise1/ folder.
- The best match units are computed using som_bmus and output as the variable bmus.
- The SOM node patterns are the values in sM.codebook which is a matrix the size of (number of nodes, length of x vector). **Add code in the designated location in Assignment5.m** to plot each of the SOM Node patterns in the designated location at the end of the Assignment5.m code. To do so, you can use the subplot function in Matlab to generate a grid of maps of the same size as the SOM grid. Include the SOM node pattern numbers (ex (1,1) or (3,4)) and the mean number of hits in the individual map titles. In the title for the entire SOM map (Hint: use suptitle) include the value of the final quantization and topological errors.
- For each SOM map node, **Add code in the designated location in Assignment5.m** to compute a composite of the data vectors associated with this SOM map node. To do so you will use the best match unit vector `bmus` that contains information about which map node was the best match unit for each datavector. Plot these composites in the same manner as was done for the SOM node patterns above. Note: these figures will be very similar to those created above.

2. Impacts of SOM Parameters

In this question, you will conduct sensitivity tests of several SOM parameters for the first data set `sine_wave_data1.nc` . For each of the following tests, if you do not initially have a flat Sammon map re-run the Sammon map algorithm up to 5 times to try and generate a SOM with a flat Sammon map.

- Without altering any other parameters, re-run the SOM 3 times changing the trial number each time. Is the SOM that has been generated identical to the initial SOM. Why? Include one of the re-run SOM node figures in this report.

- Change the NgridX parameter, the number of SOM patterns in the x direction, from 3 to 2 and any other parameters you think should be changed as a result of the change in SOM size. Re-run the Assignment5.m code. How does this impact the SOM patterns, QE, TE, Sammon map? Why? Was it difficult to get a flat Sammon map? Include the SOM node patterns, QE plots, TE plots, and Sammon map in your report.
- Change NgridX to 4 and NgridY to 5 and any other parameters you think should be changed as a result of the change in SOM size. Re-run the Assignment5.m code. How does this impact the SOM patterns, QE, TE, and the Sammon map? Why? Include the SOM node patterns, QE plots, TE plots, and Sammon map in your report.
- Return NgridX to 3 and NgridY to 4. Test the influence of the trainLmult variable, which is the number of times the input data is presented to the SOM for each training period by changing trainLmult to 10 and then 5. How does this impact the SOM node patterns, the quantization error, and the Sammon map. Include the SOM node patterns, QE plots, TE plots, and Sammon map in your report for each of these new training length multiples.
- I encourage you to play with other parameters, but there is no need to include more tests in your report.

3. Application to Different Data Sets

- Conduct the analysis in step 1 for each of the sine wave datasets. For each of these datasets determine a set of SOM parameters that lead to the development of good SOM maps. To do this, test various parameters such as number of SOM nodes and training length multiple. In your report, indicate what parameters were ultimately chosen and why.
- Include the SOM map node figure in this report. How does the SOM represent the various phases of a sine wave? How often do each of these maps occur on average? Why? Are there patterns that do not occur, what do these look like, and why?
- Include the Sammon map figure in this report. What is the shape of the Sammon map and why?
- Include the best match unit figure in this report. What are the best match units and how do they relate to both the SOM nodes and the SAMMON map?
- Include the 4 figures of the quantization errors and topological errors during the training period. What do each of these mean? How did the quantization errors and the topological errors changing over the first training time and over the second training time in your various tests and your final SOM?
- Compare and contrast how the SOM algorithm handles a sine wave that pauses, a noisy sine wave, changes in the sine wave amplitude, and changes in the sine wave shape.