Data Analytics 874
Post-Block Assignment 3: Self-Organizing Maps
Department of Industrial Engineering

Deadline: 19 November 2021, 23:59

Instructions

For the purposes of this assignment, note the following instructions:

- Assignments are to be completed by students individually.
- Submit your own work.
- You will submit a pdf document addressing all of the aspects listed in the specifications below. Name your pdf document `????????PBA3_874.pdf`, where the question marks are replaced with your student number.
- Make sure that your pdf document has a title page containing the assignment number as the title, and your initials, surname and student number.
- Submit your pdf document no later than the deadline.

Exploratory Data Analysis using SOM (140)

For this assignment you are going to use a self-organizing feature map (SOM) to analyze a real-world data set. The data set, `weatherAUS.csv`, contains weather data for a number of areas in Australia. You can download this data set from SUNLearn.

You will submit only a pdf document containing information as requested below.

If you are adventurous, you can implement your own SOM. However, the objective of this assignment is not to test your ability to implement a SOM, but rather your ability to apply a SOM to conduct an explorative data analysis of a given data set – without having any domain knowledge. Therefore, you may use any SOM library or tool box. You will find SOM libraries for Matlab, R, and python. WEKA is a machine learning and data analytics toolbox implemented in Java, and also offers implementations of a SOM.

The Dataset

The dataset contains about 10 years of daily weather observations from many locations across Australia. The target feature is `RainTomorrow`, and indicates whether it will rain the next day based on the current day weather information. There are 145460 entries in the dataset and 22 descriptive features. Note that the dataset containes many missing values indicated as `NA`.

The Tasks

After you have selected a SOM implementation, you have to do the following:

- Explore the provided data set, and report on the descriptive statistics of the features, and any other visualizations that you find usable to gain a better explorative understanding of the dataset. (20)
- Describe, with motivations, any data pre-processing that you have applied to get the data ready for SOM training. (20)
- Find the architecture and parameterization of the SOM that provides you with the best possible feature map. In your document provide detail on the performance measure(s) that you have used to determine the best SOM, as well as the process that you have followed to decide on the best SOM configuration. Provide full detail on the selected architecture and parameterization of the SOM. (20)
- The next part of the assignment is the most important part, and will test your ability to explore relationships among the features of this data set. Provide descriptive statistics for the different clusters in your feature map. Use these descriptive statistics and the component maps to identify patterns from the data. In your pdf document, present and discuss all of the patterns that you can identify. Provide motivations for these patterns, referring to the descriptive statistics and component maps. As a final step, indicate if any of the included features can be considered irrelevant or redundant. (50)
- For the last part of the assignment, use the code vectors as input to a rule induction algorithm and extract rules to describe each cluster produced by the SOM. Label each cluster wirh the class value that occurs most frequently among the instances that are assigned to each cluster. Provide the extracted rules, with performance measures per rule and for the entire rule set. (30)

Before you attempt the explorative analysis of the data set, it will help if you read articles on the application of SOMs. Such articles have been uploaded on SUNLearn.