Introductory Data Science
Final Project Instructions

Introduction

The final project requires you to perform the various data analyses described below, on two real-world datasets. You must submit a written report that defines the problems you are working on, describes the properties of the datasets, describes how you implemented each analysis, and discusses the results. The assumed audience for the report should be a technical consultant who is familiar with the basics of the techniques that you have used, but has not analysed the data in detail themselves. There is a rubric available in myUNI to guide you in the final preparation of the draft.

The actual analysis is to be completed in the last two practical sessions of the course, and hence should take no more than 4 hours. Extra time outside of these sessions should be used for writing the report, which should not exceed 3,000 words in length (do not feel compelled to saturate the word limit).

Analysis 1: Regression

The UCI Machine Learning repository contains a variety of real-world datasets. For this example, we will analyse forest fire data, taken from here:

https://archive.ics.uci.edu/ml/datasets/Forest+Fires

For each data point in the sample, the amount of burnt land is quantified by the variable area, which gives the hectares burnt. Further information on the dataset is available in the accompanying paper here:

http://www.dsi.uminho.pt/~pcortez/fires.pdf

Split the forest fire dataset into a training set (80% of the data) and a testing set (20% of the data). Include in your report a section on the data preparation, including how you converted the date information from non-numeric to numeric data.

Then:

- Build a regression model using *k* nearest neighbours that can predict the burnt area for new data points. To make the prediction, use a weighted average of the area values for the nearest neighbour points. Be careful to specify in your report which variables were used as the input variables (you can use all of them, but need to describe them).
- For the same dataset, build a multiple linear regression model in Excel that can predict the area.

By evaluating the performance of your models on the test set, explain which model gives more accurate predictions and how this was assessed. When performing the validation on the testing set, it might be useful to use the one-way table approach taken in the by-hand neural network practical. Here, you can consider using the What-If Analysis → Data Table… as a way of evaluating the models on the testing set.

Analysis 2: Clustering

In this example, we will analyse the wine dataset from the UCI Machine Learning repository that is described here:

https://archive.ics.uci.edu/ml/datasets/Wine

The actual csv file for the data is available here:

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

For this dataset:

- Perform a *k* means clustering analysis for different values of *k*, including running the analysis multiple times with different initial centroid positions. Be careful to exclude the "region" variable which labels the known origin of each wine.
- By making a scree plot of the lowest cluster distortion found for each value of *k* vs *k*, determine the optimal number of clusters in the data.
- Does the number of clusters that you obtained above match the region data that was excluded in the training?