Peer feedback INFOMDA2 assignment

Students: Jonathan Koop, Merlin Urbanski, and Martijn van Dam

Authors of paper to be reviewed: Mees van Uden and Lisanne Jannink

Date: December 13th, 2024

Summary

In this study, the authors use a data set consisting of information on economic concepts for trade data in the United Kingdom. As there are more variables than observations, the authors propose using an analysis technique named non-negative matrix factorization to reduce the dimensions in the data while retaining predictive accuracy. The authors will evaluate different configurations of the NMF analysis in relation to a benchmark 'standard' procedure, namely linear regression. The models are evaluated on the Mean Absolute Error and/or Mean Squared Error, where lower values indicate less bias. Finally, the authors will give a conclusion on which model has a better performance on the data.

Feedback

From the introduction it is clear why economic data is interesting to analyze. However, it is not apparent what the data consists of. As an 'entirely correct' explanation of the data is expected in the grading rubric, it would be helpful to add more explanation of what the variables indicate. Most importantly, the draft does not provide a clear definition of the target variable, e.g. describing how 'export trends in the UK' are measured. Regarding the predictors, the 'indicators that track trends' part should be revised, as this is very broad. This would help the reader better understand what data you are using and what export trends are. Additionally, it is also not explained what high-dimensional modelling techniques are, or what high-dimensional means. We suggest adding a definition of high-dimensional (techniques), without going into too much detail, to enhance the interpretability of your introduction. Lastly, we advise you to end your introduction with a clear research question, as this is also mentioned in the grading rubric.

For the method section, you are describing that as reference model you will 'use a linear regression [..] based on all variables to predict the target variables' with 400 predictors and 200 observations. Given that this results in a model with more predictors than observations, the procedure would lead to an unidentifiable model. Therefore, we recommend you reconsider this approach and instead either select important predictors based on theory or statistical approaches to estimate your model.

Additionally, you could enhance your explanation of the NMF method by describing what a latent space is, and what you will do with the output of the NMF, since the analysis methods should be 'concisely explained'. It is currently unclear how the NMF output will be used to predict your target variable(s). Regarding the prediction of target variables, it is unclear which and how many variables you are predicting. Moreover, it is

also unclear how you are doing it in case of multiple variables, as regression only allows one dependent variable per model.

Next, the explanation of how you will handle missing data is rather short. You could give some more explanation on what exactly you are going to do with it. Additionally, you explain that you are going to use NMF, but not why it is a good option for your data. It will give your assignment more credibility if you explain this. Regarding completeness, you can add a description of how many components of W you will be evaluating, instead of leaving that open. It could also be wise to use k-fold cross-validation in this paper, as this will yield estimates of test-MAE and MSE. This makes your model comparison more robust. Lastly, you could explain which statistic (MAE and MSE) you are using in which situation, as this is also mentioned in the grading rubric.

Regarding readability and language: Not every sentence in your assignment has an academic tone. For instance, the sentence “We will be testing NMF with different amounts of components (dimensions), but we need to make sure we have a high enough proportion of variance explained in these models.” You could rephrase this as “We will evaluate NMF with different numbers of components (dimensions), ensuring that each configuration achieves a sufficiently high proportion of variance explained”. In academic papers it is generally not recommended to use contractions of words, such as “we’re” instead of “we are”. The tone of your paper will become more formal if you replace the contractions.

Finally, the draft lacks information regarding the resources used to generate the communicated content. To scientifically back your arguments, we strongly recommend citing academic articles that provide empirical evidence supporting your claims.

Taken together, you are on the right track for an assignment that is formal in tone, transparent, and comprehensive. We hope that the feedback can help you achieve this even better.