Preparations for Master Thesis

Samir Adrik
Student number: 975149

October, 2017 - November, 2017

1 Practical Preparations

Language: The written language for the thesis is **English**. By doing so one does not need to worry about translation difficulties/challenges. Most of the literature on possible subjects is in also written in English.

Version control: Setup a master repository called `NMBU_BIAS_MasterThesis_SamirAdrik` where all the documentation and source code will be stored. The chosen version control system is **Bitbucket** (can also switch to **gitHub** if that is preferred), and each of the respected supervisors (Kristin, Håvard and Trygve) shall be given read access to this repository.

Programming languages: The two main programming languages that the source code is to be written in is **R** (version 3.4.1) and **Python** (version 3.6.3). The source code will be written using the IDEs **RStudio** and **PyCharm**. Furthermore, extensive use of the module `rpy2` (Gautier, 2008) will be used to run R embedded in Python processes. Other programming languages such as **Matlab** are also open for use.

Markup and reference management: The thesis will be written in LaTeX using **ShareLateX** as IDE and **bibtex** as reference management software.

2 Theoretical Preparations

Possible methodologies to use in the thesis: The following is a list of the methodologies that have been taught at courses at NMBU and the ones that are currently researched for use in the thesis.

Dimension reduction: Principal Component Analysis (PCA) which was used extensively in STIN300 and STAT340. Also working on getting a better understanding of addition methods for dimension reduction including Kernel PCA, Graph-based kernel PCA, LDA (used in STAT340) and GDA.

Cluster Analysis: Mostly based on methods covered in STAT340. Hierarchical clustering and Kmean clustering are the main methods currently examined.

Classification: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) (STAT340). Also working on getting a better understanding of additional methods including, Logistic regression (used in STAT340), Naive Bayes classifier, Support vector machines, k-nearest neighbor (STIN300 and STAT340) and Decision trees (INF221).

Multivariate Statistics: Multivariate statistical analysis using Tikhonov regularization, PCA/RCR, PLS, Factor analysis and Multivariate analysis of variance (MANOVA). All methods are mentioned in STAT340 with the exception of Tikhonov regularization.

References

Gautier, L. (2008). rpy2: A simple and efficient access to r from python. *URL http://rpy. sourceforge. net/rpy2. html*.