Compulsory Exercise 2

TMA4268 Statistical Learning V2023

Daesoo Lee, Emma Skarstein, Kenneth Aase, Stefanie Muff
Department of Mathematical Sciences, NTNU

The submission deadline is: Thursday April 20, 23:59h

Introduction

Welcome to the TMA4268 project exercise! In this project, you will be working on analyzing a data set by developing a model using machine learning techniques.

The goal of this project is to give you hands-on experience with the methods learned in the course and to help you develop your skills in data pre-processing, model selection, and evaluation.

To get started, here are a broad guideline for your project (detailed guideline is specified below):

1. Data set: Choose a data set of your choice. Ensure that your chose data set is diverse enough and contains enough data points to train a good model. You can find a variety of data sets in Kaggle. You can try searching for a topic that you find interesting and would like to work with.
   Hint for finding a "well-organized" data set: We recommend choosing a data set with sufficiently-large enough upvotes (i.e., shown as an upward arrow). Otherwise, a data set could be often un-organized or poorly structured with missing values. We don not want you to spend much time on data cleaning!
   If you cannot possibly decide which data set to use, please contact Daesoo Lee by email.
2. Data pre-processing: Before building your model, you might need to pre-process the data (data wrangling), depending on what format your methods require.
3. Model selection: Choose appropriate models and methods for your project. Ideally, you should use methods that you have learned about in the course, such as linear or logistic regression, regression or classification trees, random forests, SVMs etc. Make sure to justify your choice of algorithm and tune the hyperparameters to improve model performance. **Use methods from at least two (ideally more) different modules of the course**. If you are solving a classification task, for example, you can compare various approaches like logistic regression, LDA, QDA, KNN, regression trees/random forests and SVMs, etc.
4. Model evaluation: Evaluate your model's performance using appropriate metrics or evaluation tools, such as accuracy, MSE, sensitivity, specificity, residual plots etc. You should also use techniques such as cross-validation to ensure that your model is robust.
5. Reporting: Present your results and state your findings and interpresentation on the results.

We hope you find this project both challenging and rewarding.

Best of luck!

Grading

Maximal score is 100 points and the number of points given for each section of your report are indicated below. The grading is given by PASS/FAIL. To pass the compulsory exercise, your score must be at least 60.

Supervision

We will use the times where we would have lectures and exercises for supervision.

Supervision hours (in the usual lecture rooms):

- Thursday, April 13, 8:15-10:00 and 10:15-12:00
- Monday, April 17, 10:15-12:00
- Thursday April 20, 8:15-10:00 and 10:15-12:00

Remember that there is also the Mattelab forum, and we strongly encourage you to use it for your questions outside the supervision hours – this ensures that all other students benefit from the answers (try to avoid emailing the course staff).

Practical issues (Please read carefully)

- You should work in the same groups as for compulsory exercise 1.
- Remember to write your names and group number on top of your submission file.
- The exercise should be handed in as **one R Markdown file and a pdf-compiled version** of the R Markdown file (if you are not able to produce a pdf-file directly please make an html-file, open it in your browser and save as pdf - no, not landscape - but portrait please). We will read the pdf-file and use the Rmd file in case we need to check details in your submission.
- In the R-chunks please use possibly both `echo=TRUE` and `eval=TRUE` to make it simpler for us to read and grade. If the output is too long and thus not shown, you may indicate this.
- Please do not include all the text from this file (that you are reading now) - we want your R code, plots and written solutions - use the attached template `Compulsory2_template.Rmd`.
- Please **not more than 14 pages** in your pdf-file! (This is a request, not a requirement.)
- Please save us time and **do not submit word or zip**, and do not submit only the Rmd. This only results in extra work for us.
- **Bonus hint: Neat reports are easier to understand and may result in a better grade - simply because we cannot give full points if things are unclear, ambiguous or messy.**

Guideline for the Template

Please use the template `Compulsory2_template.Rmd` that we provide on the course website (under the *Compulsory Exercise 2* tab).

Title

Replace the placeholder title by an informative title.

Abstract (max. 350 words) (5 points)

The purpose of the abstract it to give a short and concise summary of your project. It is a stand-alone text that is given before the actual report starts. It includes the following components:

1. Begin your abstract by clearly stating the purpose of your project. What problem are you trying to solve? What question do you want to answer? It is important to be concise and to the point.
2. In the next few sentences, describe the data and methods you used to conduct your study. What kind of data set did you use? How did you analyze it? What tools, techniques, or methods did you use? Be specific, but avoid going into too much detail.
3. Summarize your key findings: In the main part of your abstract, summarize the most important results of your project and interpret them shortly (i.e., what does this mean?). Highlight the most significant findings, and provide enough detail to give the reader a sense of what you discovered.
4. (optional) Emphasize the significance of your results: Explain why or/and how your finding(s) is/are important. Highlight any novel or unexpected findings, and explain how they add to our understanding of the topic.

Introduction: Scope and purpose of your project (15 points)

- Briefly introduce the broad idea of the problem or task that you chose and the respective data set that you use. This could be a classification task (e.g., predicting the species of flowers in the Iris data set) or a regression task (e.g., predicting the price of a house based on its features). Clearly define the scope of your project. What specific problem are you trying to solve?
- Describe the source and give a referenc to where the data set is coming from.
- Describe the purpose of your project is some more detail. What are the specific question that you want to answer in your project? Are you trying to find the best performing method or a good performing and light method that is easy to use? Who is your audience? Are you trying to discover the relations between different variables? Are you trying to find important predictors for your classification? Are you trying to draw some insightful understanding in a particular topic/domain? Importantly: Is the main purpose inference or prediction?

Descriptive data analysis/statistics (15 points)

Conduct descriptive data analysis to get an overview over your data (see this example for inspiration).

Examples:

- Report measures such as mean, median, range, standard deviation, and variance to describe the central tendency, variability, and distribution of a data set.
- Scatter plots and correlation matrices across different variables and histograms of variables (see this example).
- Box plots of variables.

Methods (30 points)

- Describe the methods that you are using in your project and explain in detail how you applied them. Depending on the task, these could include methods such as linear or logistic regression, decision trees, random forests, support vector machines etc. You should use several methods for your problem so that you can compare their performance.
- Explain briefly how each method works, what its strengths and weaknesses are, both in general but also in the light of your project (how suitable is the method *in your case*?).

- Describe which hyperparameters are optimized for the methods (e.g., a shrinkage factor is a hyperparameter in Lasso).
- Describe clearly how you evaluate the performance of the different models and methods (accuracy, MSE, misclassification error, CV error,...). Explain how each metric is calculated, and why it is a useful measure of model performance.
- (optional) Consider and describe potential limitations of the methods and the chosen evaluation metrics.

Results and interpretation (30 points)

1. Present your results in a *clear and organized* manner. This could include tables, graphs, or other visualizations that help to convey your findings. Report also all the hyperparameters, the performance (e.g., test error) etc. that you introduced in the Methods section.
2. State your interpretation of the results. **It is important that you compare the different methods in terms of different aspects such as model size, flexibility, bias-variance trade-off, etc.** The interpretation is very much dependent on your initial question. For instance:

- Classification: In a classification problem, interpret the outcome and performance of the different methods and compare them. Which method worked best and why? In addition, the importance of each feature in the model can be analyzed to gain insights into the underlying patterns in the data.
- Regression: In a regression problem, the result interpretation is typically the predicted value for each instance in the test set. The mean squared error (MSE) can be reported to evaluate and compare the performance of the models. In addition, the coefficients of each feature in the model can be analyzed. You may identify the most important predictors for the target variable. Again, the interpretation is depending on your research question (e.g., inference vs. prediction).
- Clustering: In a clustering problem, the result interpretation is typically the grouping of instances into clusters. The characteristics of each cluster can be analyzed to gain insights into the underlying patterns in the data.

3. Discuss any limitations or caveats that are important to keep in mind when interpreting your results.
4. (Optional) Give an outlook on potential alternative/better ways to analyze your data in the future.

Summary (5 points)

Summarize the main findings of your project. What did you discover, and what were the key insights that you gained from your analysis?