Problem Set 3

Luisa Weber & Martin Klarmann

Winter Term 2025/26

Data To make you successful in working with Lasso regression in variable selection settings, in this problem set you will work with two different types of data sets. (1) In the first task, we will provide you with music data through ILIAS. Please know that this music is copyrighted and can only be used for the content of developing your algorithm. (2) For the second task, you will be provided with a data set on track times, track conditions, and car configurations from a racing simulator.

The "Team Analytics" Racing Competition This will be the setup for Problem Set 3 and Problem Set 4. We will explain this more in the classroom today.

Please use the following link to register for "Team Analytics": https://team-analytics.com/f1/registration/register.php?link=ed718ec8defe6ad447122de016d3d5c6 - Select your team number according to your group in ILIAS (Group A = Team 1, Group B = Team 2, etc.) - In ILIAS (in your group folder) you will find a student's manual that gives you a first introduction about: Team formation/registration, Practice/qualifying laps/race strategy submission, effective use of the data center and simulation tools. Please read the manual carefully.

After you registered the team, you can download the simulator data and the track data.

In addition to existing datasets, as part of the next problem set, you will also take the car out on the track for practice laps, for learning more about the car setup and the racing strategy. This will allow you to experiment with various configurations and understand the impact on performance. Experimentation is the big topic of part 4 of the lecture. But be careful, you will only have a limited experimentation opportunity: each team has 80 practice laps per race.

Once you are comfortable with your setup, you have to submit a qualifying lap, which will lock your starting position and car setup for the race. Furthermore, you will also submit the race strategy, including pit stops, tire types, and fuel loads.

Please make a note of the race dates in advance: the warm-up race takes place on a track in France (9 January 26). The chosen tracks for Problem Set 4 are in Silverstone, England (16 January 26), Spa, Belgium (21 January 26) and Monza, Italy (28 January 26).The race will start each time at 8:00pm. It is not required for you to be live throughout the race. To make the race run smoothly, you will need to upload your configuration and your strategy at 5:00 pm on the day of the race. If you don't submit your qualifying lap and race strategy your car will not start the race.

Document This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Relevant R packages For the tasks this week you will need the following packages in addition to the packages included in the standard installation: `glmnet`.

Deliverables and Deadline

1. A Markdown Notebook with all the code and your argumentation. We may apply this code to the original dataset, so please make sure that all changes to the dataset are programmed in the notebook (and not done in Excel etc.).
2. A PDF document with your responses to all tasks (max. 15 pages). We recommend using your Markdown Notebook document to create your PDF file.

Results need to be uploaded to ILIAS in the relevant section on Friday, 19th December 2025, at 5pm.

Task 1 “Song recommendation” - 13 out of 25 points This task is about predicting preferences from unstructured data. The setting is to make a prediction about the preferences for music. The idea is that you learn about the song preferences for a 100 songs, which will then enable you to predict the preferences for a 101st song.

For this task, we will provide you with a more or less representative set of 101 songs across a diverse set of genres (we asked ChatGPT to come up with such a list). They are already provided in WAV format, which enables further processing. Moreover, for 100 songs you will be provided with a single rating on the familiar 5 point star scale. The star ratings come from one individual, so they should be consistent in showing the musical preferences of this individual.

Your overall objective is to predict the rating of this unknown song by extracting relevant audio features from all available files and training a suitable regression model on the 100 labeled examples.

In a first step, you need to extract acoustic features from all 101 WAV files.

Notably, R does not provide a straightforward way to do this. One approach in R would be to use the “voice” package, which will return a set of features. However, it was designed to analyze voice data and so the resulting feature set may not be ideal.

Therefore, please download the separate software “opensmile” (developed at TUM) from GitHub (https://github.com/audeering/opensmile/releases/tag/v3.0.0) and use it to extract the features into a csv. You can call this from R using the system command, but if you want to do this through your system that is okay. Opensmile offers a diverse set of configuration files extracting different feature sets. For our purposes a standard configuration such as compare16 will probably work decently well.

Here is how you can call opensmile to do this:

.../opensmile-3.0-win-x64/bin>SMILExtract -C “.../opensmile-3.0-win-x64/config/compare16/ComParE_2016.conf” -I “.../Adele - 01. Hello.wav” “.../features.csv”. “-C” defines the configuration file, “-I” the input file, and “-O” the output file.

The features.csv is not a straightforward data frame, you will need to have few lines of code to transform it so it works in R.

If your group is feeling more adventurous, you could either try to define a set of features using R’s seewave functionality or work with Python’s librosa or rp_extract. Both can be executed within R. Alternatively, you could just use them for feature extraction in Python and then use the resulting csv’s as input for glmnet in R. By doing this, you will get a feature set generating a rich representation of each song.

Use these features and the 100 song ratings to train a Lasso regression as a prediction model. Finally, apply your trained model to the 101st song in order to estimate its star rating.

Your submission should clearly describe your preprocessing decisions, the extracted features, the modeling strategy, the selected features in the final Lasso model and the predicted rating for the unknown song. Please discuss the limitations and potential improvements of your solution. Can you learn something more general about the preferences of the individual from the features that were selected?

This task explicitly refers to chapters 3.2 and 3.3 of the lecture.

Task 2 “Preparation for Team Analytics” - 12 out of 25 points In the Team Analytics race simulation, the decisions you need to make before the game can be roughly split up into a decision on the car configuration and a decision on the race strategy (particularly fueling tops and tires). As a preparation this task should result in a prediction model that will allow you to determine a starting car configuration and the most important drivers of lap time for each race in Problem Set 4.

The task uses the dataset called “simulator data” that consists of lap time with varying track conditions and car configurations. It includes high data noise.

In a first step, start by carefully exploring the bivariate relationships in the dataset using correlations and visual inspection of scatterplots. Do you already see some evidence for nonlinear relationships?

In a next step, please expand the variables so that they comprise all possible two-factor and three-factor interactions, both linear but also including polynomials. We have no experience with this data set, but we would expect that you do not need to include polynomials higher than the order 3. Notably, there could also be nonlinear interactions. Use some type of vector notation to write up your full model.

Then, please use the simulator data combined with Lasso regression to identify the key drivers of lap times. Which variables and which interactions are particularly relevant for predicting lap times? Please provide some insightful commentary on what you learned—and what you still need to learn from the experiments on the course—for the races.

Using this model, please come up with a prediction for the best car configuration for the warm-up race in France (Le Mans). We will simulate the warm-up race as part of Problem Set 4 in January, but you can already see the weather conditions for the race day and the track characteristics in the simulation tool.

This task explicitly refers to chapters 1.2, 1.3, 1.4, and chapter 3.3 of the lecture.