Karlsruher Institut für Technologie

Institut für
Thermische Strömungsmaschinen
Prof. Dr.-Ing. Hans-Jörg Bauer, Ord.

ITS

Winter Term 2021

Term Project Guide

This is an informative guide intended to help you to structure your project. Please note that you are expected to choose a unique dataset via ILIAS until 24.12.2021.

Highlights

- ✓ The objective of the project is to help you to experience an end-to-end machine learning project. We provided you many problem choices so that you would be dealing with an interesting task.
- ✓ Following the cooking analogy, the best way to learn “cooking” is to try it by yourselves.
- ✓ You are supposed to finish and submit your project before you take the final exam, preferably by the end of the DDE lecture.
- ✓ If you are taking the lecture for credits, you should note that the final exam will revolve around the project you prepared. The best way to prepare is “cooking” and a good way to measure how much you learned is to “taste the dish you serve”.
- ✓ This document is the recipe for your dish: make sure to understand what we expect from you.
- ✓ You can always ask your questions via ILIAS forum.
- ✓ We will spare two lectures at the end of semester for your specific questions.
- ✓ Note that the projects will be collected as Jupyter notebooks.
- ✓ In your code, use reasonable variable names and be generous with the comment lines.

Structure of the Project

Although every project is unique and focuses on a different problem, we expect you to follow a structured outline. Here is the recommend common structure. You can be creative, yet you should present a clean workflow.

1. Analysis of the Problem
2. Data Exploration and Preparation
3. Testing Phase I: Baseline Models
4. Testing Phase II: Model Development
5. Testing Phase III: Model Regularization and Hyperparameter optimization
6. Evaluation of the model predictions
7. Lessons Learnt and Conclusions

Analysis of the Problem (Task)

This section, in principle, consists of three parts.

Understand the Problem:

During the initial phase, you have decided to work on a certain dataset. You should first explore the physical system you have chosen. What was the initial reason attracted your attention? What are you interested in? What is your focus or objective? What are the goals you would like to achieve at the end? These are the first questions you need to ask yourselves. After convincing yourselves, include a brief paragraph describing “your” problem.

The next step is to look at the data with this perspective. Now we are facing the second set of questions:

- What type of problem are you dealing with? Is it classification? Binary or multiclass classification? Is it a scalar regression or a vector regression (set of numbers)? Is it related to clustering, generative learning or forecasting? Are you solving a stationary problem or a dynamic problem?

This is a very critical step. Here you hypothesize your work frame, where your outputs can be predicted given your inputs! Remember that not all problems can be solved. Having a set of data doesn’t mean that your features contains enough information to predict your objective. Select a reasonable target!

Identifying the problem type will help you to think about the model architecture, loss function, and loss criteria. In your report, we expect to see what you think here. Write your working hypothesis here.

Choosing a Measure of Loss / Success:

ML is an optimization problem, which requires a comparison method. After understanding the problem and forming the working hypothesis, you should select a measure of success. To optimize a problem, you make predictions and your observations are valuable only if you can compare them by using a mean (method). “What could it be?” is the question you should answer here. Note that your criteria must be related to your high level goal. For instance, predicting the noise level of an airfoil, type of the noise (low, medium, high) or the product type (anomaly detection: regular, defective).

You should be careful here! For instance, do not use area under the receiver operating characteristic curve (AUC) for class-imbalanced problems. You can use precision and recall. For ranking problems or multi-label classification, you can use mean average precision.

Remember that in the lecture, we also had to write our own functions to create a custom metric, if the problem demands. You got confused? Use the forum to share your troubles. You can always use the readily available tools (active session notes). See also what others did, use Kaggle to look for different perspectives if needed. As long as you understand what you did and why you did it, feel free to explore!

Customize your evaluation protocol:

In most cases, we use labelled data. Depending on your dataset, you must think a process flow diagram for your study: how much data will you spare for testing? How do you measure your training success?

In the lecture, we used one popular method for that: K-fold cross-validation. Make sure that you implement it (or similar ways) to optimize your training in the later stages.

Data Exploration and Preparation

You're almost ready to train / test your models. But first, you should check your data and format it in a way that can be fed into the machine-learning modes you will use. In the lecture, we have seen various ways to do it.

- Check for missing values, NaN values or features,
- Make sure that you look into the uniqueness of the data (is it as it was expected to be)
- Use statistical tools to understand the variations in the data.
- Do you have unphysical outliers?
- Is it possible to combine features?
- Do you have unnecessary features? (a column which gives no information – for instance name column– or a feature you consider unrelated to the problem)
- Check the correlation matrix. It will tell you how much the features are related. You may say, for instance, there is a great potential to reduce number of dimensions.
- Most ML algorithms requires scaled data. Make sure that you scaled it before feeding to the ML model (for example, in the [-1, 1] range).

Testing Phase I: Baseline Models

At this stage, you are ready to use ML models. In the project, you are expected to use models from two families: the ones you learnt in the fundamentals phase (i) and a NN-based (ii) model. In phase I, you will use the former to explore the possibilities in your data. This is quite typical in a real life ML project. Here you are free to try different combinations – nonetheless, you should demonstrate in your project notebook what you have tried and what you have observed.

One important note here is that if you have a high dimensional data (larger than 3), you are expected to use dimensionality reduction for visual exploration. Furthermore, you are supposed to experiment the impact of the coordinate transformation (for instance use PCA) or clustering on the predictive accuracy of these fundamental models. In short, create ensemble learners. This study will be your baseline.

This section should conclude with your observations and conclusions. Please do not extend your discussions indefinitely. A small paragraph would be sufficient to explain your points.

It is possible that **your initial hypotheses are false**, in which case you must go back to the previous step and re-think about the problem. Therefore, this stage is the first battlefront of your initial ideas. You should not delay it to the last weeks!

Testing Phase II: Model Development

In this phase, you will design your base NN architecture. You can start working on this task after Active Session 5 for most problems. If you are dealing with forecasting or transient data, you should be able to work on this task after Active Session 6.

The objective of this phase is to help you to build skills to design a NN / DNN / RNN /AE / VAE. Assuming that your Phase I results are promising, here you will decide the key components of the NN architecture. The type of neurons will be decided based on the task you have. Your loss function, activation functions, network initialization, optimizer and the last-layer activation.

Testing Phase III: Model Regularization and Hyperparameter optimization

In this stage, you will fine-tune the NN architecture to increase the predictive accuracy of your custom model (as in Active Session 5). The question you should answer here is “what is the right complexity for my problem (with this objective function)”.

Remember the universal balance in ML lies between optimization and generalization; underfitting and overfitting; under/capacity and over/capacity. To figure out where this border lies in your problem, you must first cross it.

To figure out how big a model NN you need, you must develop a model that overfits when the volume expands. Follow the tips in the lecture notes while adding more layers, nodes and train for more epochs. Always monitor the training loss and CV loss, as well as the training and validation values of the metrics you care about.

The next stage is to start regularizing and tuning the model, to get as close as possible to the ideal model that neither underfits nor overfits. As a rule of thumb in DNN, you should aim for a large, regularized network.

This step will take the most time: you'll repeatedly modify your model, train it, evaluate on your validation data, modify it again and again, until the model is as good as it can get. You should make notes of your trials, save your plots so that we can follow what you did. What you can do is already in the active session notes.

Note that every time you use feedback from your test data to tune your model, you leak information about the validation process into the model. Therefore, you should use CV to optimize your model! After you fine tune your options, use the final model on the test data.

Add a brief description of this optimization process. You do not need to keep all calculations as code cells in the notebook. Make sure that we understand how much effort you put in here!

Evaluation of the model predictions

At this stage, you will compare the methods you tested in Phase I and Phase III. This will be a brief discussion on your findings.

Lessons Learnt and Conclusions

Finalize the report with a brief conclusion: tell us what you found and what you learned!