Data Science 1 - Final Project

TECHNISCHE UNIVERSITÄT DARMSTADT

Updated 25 April 2022

1 Introduction

SIGNAL PROCESSING GROUP

Dr.-Ing. C. Debes
M.Sc Pertami Kunz

The objective of this final project is to demonstrate all your acquired knowledge in data science throughout a practical problem to be solved. This includes a proper problem formulation and motivation, determining the suitable algorithms and presenting your outcome. It is also a perfect preparation for the project seminar Data Science II.

2 Tasks

- Choose a dataset of your interest.
- Choose between the two:
  1. Separate hypothesis test and regression/classification method:
     Decide which hypothesis you want to test, and separately: what's the response variable you want to predict using some/all other variables.
  2. Hypothesis test on regression/classification models:
     Decide the response variable you want to predict some/all other variables, test multiple models and decide which model is the best statistically (by hypothesis test, not by comparing the performance numbers directly, as the numbers may have occurred by chance).

3 Report Checklist

Your final report should include the items below (ones with ✓). They do not need to be always in this given order. Each item does not need to be in a separate section/subsection. The important point is to make your report **concise and fluent**.

- Introduction
  - ✓ Objective.
    Introduction to the objective of your report. Formulate it in such a way that it is understandable by a non-technical audience. Describe the business/societal or other benefits arising from it.
  - ✓ Introduction to the dataset.
  - ✓ Problem statement.
- ✓ Some preprocessing.
  Do the entries need to be cleaned, filtered to fit your problem of interest?
- Exploratory Data Analysis
  - ✓ Some **useful** plot(s) and/or table(s).
    Plots and tables that lead to your question and/or hypothesis.
- ✓ Some feature extraction/engineering
  Anything to be imputed? Normalized? New features can be derived?
- Hypothesis Testing

SIGNAL PROCESSING GROUP
Dr.-Ing. C. Debes
M.Sc Pertami Kunz

  - ✓ State your hypothesis
  - ✓ Describe your test statistic, why you chose it.
  - ✓ The rest of hypothesis testing steps until conclusion.
    Which values do you need to derive your conclusion: confidence interval/significance level/critical value/p-value? You don't need to mention all of these values, e.g. if you get the p-value from your program, it's enough then to compare it to, say, the significance level you chose.
- One/some regression or classification models.
  - ✓ How did you split your data?
  - ✓ The mathematical expression of your model(s).
  - ✓ Why explore this/these model(s)?
  - ✓ How did you decide the hyperparameters (if any)?
  - ✓ The performance (on the training and test sets)
- ✓ Conclusion
- ✓ References
  Important: cite the dataset. If you mention a python library in the report, then please also cite its reference.

4 Useful links

Here are some links to choose a dataset from (you are free to choose from other sources):

1. UCI Machine Learning Repository
2. Kaggle Datasets

Examples:

1. Amazon's Books EDA and Hypothesis Test
2. A Statistical Analysis & ML workflow of Titanic
3. More hypothesis tests examples
4. Regression examples
5. Classification examples

References for hypothesis test on machine learning algorithms:

1. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms
2. Statistical Comparisons of Classifiers over Multiple Data Sets
3. Inference for the Generalization Error