Data Science Toolbox Assessed Coursework 1: Supervised Prediction

Deadline: Wednesday Noon, Week 12

Group Project description

Comparing model performance is an essential part of data science. You will **choose an application domain** that your group will work with during Assessment 0-1.

Your task for this project is to:

- Consider a **binary classification problem**;
- In which each group member will create a **model submission** that can be evaluated on **left-out test data**;
- together agree and test a **performance metric**;
- compare your models according to that performance metric.

Classification is an open-ended problem and you are invited to interpret this aim in a way that you find interesting, tractable and productive. In particular, it can be valuable to deploy traditional mathematical models to understand how they relate to real-world problems.

Remember that the goal of this is **not** to win, but to learn about the appropriateness of the model. An additional goal is to consider the appropriateness of the performance metric. You should create a test and validation dataset, but you may choose how to do this. You may also choose to limit the model to certain covariates.

Half of the effort should be devoted to exploring appropriate performance measures. Think about the circumstances by which your chosen performance metric will lead to real-world generalisability, and how it might compromise this for the purpose of standardization. Demonstrate this with data and/or simulation; for example, if you believe that you can predict **new** types of data, you could demonstrate this by leaving out some types of data and observing your performance. Examine in what sense your group's best method is truly best.

Appropriate Methods

Your models may include off-the-shelf methodology, including but are not limited to:

- Standard regression techniques such as:
  - transforming the data;
  - creating additional covariates by additional modelling;
  - using variable selection;
  - using penalisation;

    - using non-linear link functions;
- Using more sophisticated forms of regression or classification that you have researched yourself;
- Using models from your personal experience, such as:
    - Time-series models,
    - Point-process models,
    - Multivariate models,
- Any of the advanced topics that are referenced in the Data Science Toolbox lectures.

Advice on assessment

You will be assessed on:

a) whether the model implementation is appropriate, that is:
    - you can be awarded credit for additional implementation if an off-the-shelf implementation falls short,
    - you can be awarded credit for exploring multiple implementations,
    - you can be awarded credit for examining the mathematical details of choices.

b) how well explored the model is.
    - you can be awarded credit for simulation work.
    - you can be awarded credit for sensitivity analysis.
    - you can be awarded credit for plotting or otherwise describing various inputs, outputs, or parameters.

c) the correctness of the methods used to achieve their stated goals.

d) the robustness of the results in supporting the conclusions.

e) whether, after the fact, you can explain the limitations of the approach taken.

You do not need to excel in all areas in order to get a high mark. Instead, you need to perform robustly in all areas and additionally demonstrate insight somewhere to score highly.

You will not be penalised if:

- you choose a model and after testing, the implementation proves deficient,
- your chosen model performs poorly, provided that you have implemented it correctly and understood its limitations,
- the implementation makes it impossible to work with the full dataset, and you have to create a smaller artificial comparison dataset. (you may still be penalised if the model could have been simply fit a different way.)

Topics

Here are suggested example datasets. You are free to find and analyse other datasets on discussion with the course tutor.

- **Cyber Security**: predicting **normal** vs **non-normal** traffic in the KDD99 **(small, 10%) dataset** called `kddcup.data_10_percent.gz` from the week 3 workshop. You are welcome to use additional data from the KDD99 experiment, but this is not required.
- **Network data** based on the Enron email dataset. For example, you could look to predict the next email an account will send, or model the network at a more global level, etc.
- **Epidemiology data**
- Or another topic agreed between the lecturer and your group.

Individual reflection description

- Discuss the rationale behind the model and/or data extensions that you put into your own model;
- Discuss the performance metric that you chose;
- Discuss the mathematics behind your method;
- Reflect on how you might change your own model, were the intention to "win";
- Reflect on how this could work in a competition setting.