Peer–Review Questions for

ML Student Projects CS–C3240 – Machine Learning

Stage 1 – Problem formulation and one ML method

Opens: 11 Sep 2024, 10:00

Closes: 20 Sep 2024, 23:59

Stage 1 objectives

- Choose and formulate a machine learning problem
- Choose **one method** that is suitable for solving the problem

Point distribution (out of 47 point total for the project)

- Submission points: 7
- Peer grading points: 3
    - peer reviews are always done individually

Late submission policy

- Late submission open until **23 Sep 2024, 23:59**
    - Late submissions will incur a penalty of 30% on the stage 1 submission points
- Late submissions after the 23 Sep 2024 deadline will only be accepted due to illness (must provide valid medical certificate)
    - 0 points for peer grading

Overview

The objective of stage 1 is to get started with your ML project and apply what you have learnt in the lectures and exercises to your own ML problem.

You need to choose some everyday application which you will formalize as a ML problem.

Some examples of everyday applications are:

- How much could I sell my car/bike/phone/handbag/flat for?
- What is the population of Finland/Helsinki/Espoo/Vantaa/any other place in 2025?
- What is the average rent of an X square–meters flat in Helsinki in 2025?

Instead of starting from some application, you might also start from some candidate datasets (e.g., check out the list of suggested sources of data provided by the course), because at the end of the day, you need to solve this problem with a data–driven approach. If you are collecting data for your thesis or other research, you are encouraged to use the same dataset for this project as well.

At stage 1, you will choose **one method** that you think is suitable for solving the problem at hand, for example linear regression with mean squared error loss.

Check the report outline for the report format you are expected to submit at stage 1. **Make sure that your report addresses the questions outlined in the peer grading section below, because this is the criteria that your report will be assessed on.**

The page limit for stage 1 report is 3, and the minimal font size is 10. Any content exceeding the limit will not be considered during peer grading and may result in a point deduction.

NOTE: You will be expected to use the same ML problem for the full report at stage 2.

Peer–grading questions

Category 1. Problem formulation

Q1.1 Does the dataset have a **multidimensional input** (i.e. more than one feature variable) and is the meaning of the **data points** clearly explained? Does the model also use more than one feature variable? The report must **explicitly** state what the data points represent.

- 1p – Yes
- 0p – No

Q1.2 Is the **type of data** clearly stated (e.g., binary, categorical, continuous etc.) and does the report clearly state **where the dataset was collected from**?

- 1p – Yes
- 0p – No

Q1.3 Does the report explain the type of machine learning task, i.e. **supervised or unsupervised learning**, and the aims of the project? If the project is a supervised learning task, does the report explain what the **labels** (i.e. the quantities of interest) represent?

- 1p – Yes, the report clearly states the type of machine learning task and aims. If the project is a supervised learning task, the labels are clearly described.
- 0p – No.

Category 2. Methods

Q2.1 Does the report clearly state **the model (hypothesis space)** and explain the **motivation** behind using it for this ML method? Chapter 3 of mlbook.cs.aalto.fi discusses the models used by some well–known ML methods.

For example, *"Linear predictor maps are used as the visualisation shows a linear relationship between the features and the labels."*

- 2p – Yes, the model is explained, and it is also clear to me why it was chosen.
- 1p – The model is discussed but it is not explained why.
- 0p – No, the model (hypothesis space) is not discussed.

Q2.2 Does the report clearly state **the loss function (if applicable)** used and explain the **motivation** behind using it to evaluate the quality of the hypothesis?

For example, *"The logistic loss is chosen as it allowed the use of a ready-made library for logistic regression"*; *"The Huber loss is used as it is robust towards outliers."*

Examples of loss functions can be found in Chapter 2 and Chapter 3 of mlbook.cs.aalto.fi. Note that it might be useful to use a different loss function for learning a hypothesis (e.g., logistic loss) than for computing the validation error (e.g., "accuracy" as the average 0/1 loss).

If unsupervised learning methods without explicit loss functions were selected, does the report clearly state **the distance function and validation metrics** used and explain the **motivation** behind using it to evaluate the quality of the hypothesis?

- 2p – Yes, the function and the motivation behind using it is clearly explained.
- 1p – The function is mentioned by name without discussing the motivation behind choosing it.
- 0p – No, the function is not discussed.

Q2.3 Does the report explicitly discuss how **the training and validation set** are constructed, **the size of each set**, and the **reason behind such design choice**? If the task is unsupervised, does the report outline an alternative validation strategy?

Some examples are (1) using a single split into training and validation set, (2) k–fold cross validation, etc. (See Section 6.2 of mlbook.cs.aalto.fi)

- 2p – The construction of training and validation sets are discussed very clearly. I also understand why the author thinks this is a reasonable design choice.
- 1p – The construction of training and validation sets are discussed superficially.
- 0p – The construction of training and validation sets are not discussed at all.

Q2.4 Does the report explain **the process of feature selection** and any **feature engineering** that was performed?

Features should be properties of data points that can be measured or computed easily (as a highly automated and repetitive task). Some examples are (1) the red, green and blue intensity of the pixels of an image; (2) the size and number of bedrooms of a flat; (3) the weight, blood pressure and body temperature of a person.

Note that theoretical justifications are not necessary, but instead we focus on the process of how the features were selected. It could be based on data visualisation, domain knowledge and other strategies.

Some examples are: (1) *"After visualising the data with scatterplots, feature A and B shows stronger correlation to the labels than others"*; (2) *"Intuitively, the number of bedrooms, flat size and its renovation history are correlated to its price, but data of the last feature is hard to obtain/quantify..."*

- 2p – Yes, the features are explained clearly
- 1p – There is some explanation, but it is still unclear to me how the features were selected
- 0p – No, it is not mentioned at all

Category 3. Other criteria

Q3.1 Is **the code file** submitted as an appendix or does the report contain a link to the code?

- 1p – Yes
- 0p – No

Q3.2 Rate **the quality of scientific writing in the report**. Are the report format and language use professional and clear? Is the report free of typos and incomplete sentences?

- 2p – The report is well-structured and easy to follow, the language is clear and concise, and there are almost no typos.
- 1p – The report is well-written overall, but it could be improved in some respects (please provide examples).
- 0p – The writing is not professional enough for a scientific report, e.g., there are a lot of incomplete sentences and typos.

Q3.3 To the best of your knowledge, does the report contain any existing material – either from this course, Kaggle, or other sources - **without clearly indicating the source**? Please report any suspicions of plagiarism (e.g. direct copying of text from other sources) to course staff.

- 1p – I have not seen the same ML problem or discussion in any of the mentioned places.
- 1p – I have seen the exact same ML problem in one of the mentioned places, but the source is clearly indicated in the report.
- 0p – I have seen the exact same ML problem or discussion in one of the mentioned places, but the source is not indicated in the report.