TENTATIVE – last updated 7 March 2022

Peer–Review Questions for

ML Student Projects CS–C3240 – Machine Learning

Stage 2 – Problem formulation and one ML method

Opens: 03 Mar 2022, 20:00

Closes: 10 Mar 2022, 20:00

Assignment description:

At stage 2, the objective is to let you apply what you have learnt in the lectures and exercises to your own ML problem, without the burden of doing in–depth analysis and comparison of ML methods. You need to choose **one method** that you think is suitable for solving the problem at hand, for example linear regression with mean squared error loss.

NOTE: You are expected to commit to this ML problem for the full report at stage 3.

Point distribution:

- Submission: 60% (12 points)
- Peer review: 40% (8 points)

Final points = (submission + peer review) / 2

Policy for late submissions and peer review

- Late report submission
  - Open until 7 April 2022
  - 0 points for peer review
  - Does not affect submission points
- Missing peer review
  - 0 points for peer review
  - No late submissions allowed

Peer–review questions:

Category 1. Problem Definition.

Q1.1 Is the meaning of **a data point** clearly explained? The report must **explicitly** state what data points represent.

- 1p – Yes
- 0p – No

Q1.2 Does the report discuss what properties of the data points could be used as **features**? On top of that, is the **type of data** also clearly stated (e.g., integers, binary categories etc.)?

- 1p – Yes
- 0p – No

Q1.3 Does the report discuss what properties of the data points are the **labels** (I.e., the quantities of interest)? On top of that, is **the type of data** also clearly stated (e.g., integers, binary categories etc.)?

- 1p – Yes
- 0p – No

Category 2. Methods

Q2.1 Does the report clearly state **where the dataset was collected from**, the **number of data points** and give a **brief description** of the dataset?

For example, *"the dataset is obtained from this Kaggle page <link>. There are x data points in total with no missing data in any fields. The 'price' column will be used as labels. There are 10 other columns which could be candidate features..."*

- 2p – Yes, it is clearly stated where the dataset was obtained, and the description gives me a general understanding of the dataset.
- 1p – The source and the dataset are described very briefly.
- 0p – No, the source and description of the dataset is not mentioned at all.

Q2.2 Does the report explain **the process of feature selection**? 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 process of features selection is explained clearly.
- 1p – There is some explanation, but it is still unclear to me how the features were chosen.
- 0p – No, it is not mentioned at all how the final features were chosen.

Q2.3 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.

Q2.4 Does the report clearly state **the loss function** 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).

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

Q2.5 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**?

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.

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.

Category 4. Overall assessment

Q4.1 Does the report contain existing material – either from this course, Kaggle, or other sources - **without clearly indicating the source**?

- 1p – Yes, I have seen the exact same ML problem in one of the mentioned places, but the source is clearly indicated in the report.
- 1p – No, I have not seen the same ML problem or discussion in any of the mentioned places.
- 0p – Yes, 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.

Q4.2 If you answered 0p - Yes to the question above, does it also use the same model and loss function **without clearly indicating the source?**

- 5p – I chose “1p” in the question above.
- 5p – Different model and/or loss functions are used.
- 5p – The same model and loss function are used, and the source is clearly cited.
- 0p – The same model and loss function are used, but the source is **not** clearly cited.

Q4.3 Does the report contain paragraphs which are **copy-pasted** from other sources - such as the example projects, teaching material, Wikipedia, Kaggle, Stack Overflow and so on.

- 10p – No, I do not suspect any copy-pasting from other sources.

- 5p – Yes, a large part of the report is paraphrased from some source texts, but with indication of the source (students need to use their own words in the report).
- 0p – Yes, some parts of the report are copy-pasted without proper indication of the source (Please report this to course staff!).

Q4.4 (**BONUS**) Do you find the ML problem worth some bonus points? For example, is the problem formulation **highly original** or did the student **explain the use of ML method outstandingly well**?

- 5p – I think the problem is very original
- 5p – I am impressed by how well the author explained the chosen ML method.