# TENTATIVE (updated 17 Jan 2022)

# Peer–Review Questions for

# ML Student Projects CS–C3240 – Machine Learning

# Stage 1 – Problem formulation

**Opens: 27 Jan 2022, 00:00**

**Closes: 10 Feb 2022, 20:00**

**Assignment description**

In stage 1, the objective is to get you started with planning your ML project. You will come up with a problem which you will formulate as an ML problem. Later you will solve this problem with your chosen ML methods. **At stage 2, you can either use the same ML problem or use a completely different one.**

A good place to start is to think of any aspects of your life or any topic of interest that could benefit from making a prediction of some sort. A couple of examples 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?

The possibilities are endless.

Another approach to get started is to look for some candidate datasets (e.g., checkout 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.

Our advice: be creative, but at the same time, keep it simple.

**Point distribution:**

- Submission: 90%
- Review: 10%

**Peer–review questions**

Q1. Is the meaning of a data point clearly explained? The report must explicitly state what data points are representing.

Some examples are data points representing (1) images, (2) flats, and (3) people.

- 0p – No
- 1p – Yes

Q2. Does the report discuss what properties of the data points could be used as features? 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.

- 0p – No
- 1p – Yes

Q3. Does the report discuss what properties of the data points are the labels (I.e., the quantities of interest)? The labels are usually not readily available. To obtain them, it usually requires an expert to examine the data points. The goal of the project is to obtain labels with ML methods to make predictions in place of an expert.

Some examples are (1) whether there is a cat or a person in an image; (2) the maximum price a flat could be sold at; (3) the probability that a particular person needs intensive care next week.

- 0p – No
- 1p – Yes

Q4. Does the report suggest at least one source of data, from which both the features and labels could be found/downloaded.

- 0p – No
- 1p – Yes

Q5. Has this ML problem appeared in the teaching material or the example projects? Is it a direct copy from some well-known ML problem?

If your answer is yes, please specify in the comment box below where you have seen this ML problem appear.

- 0p – Yes, I have seen the exact same ML problem in one of the mentioned places (I.e., same data points, features, and labels).
- 4p – No, this ML problem is either completely original to the author or is formulated differently based on an example in the teaching material.

Q6. Do you find the ML problem very original? Do you consider it a straightforward variation of problems discussed in the lectures or exercises, or well-known ML problems from forums like Kaggle?

Note: The ML problem does not need to be groundbreaking to be awarded 1p. If the project uses data points, features or labels which is mentioned elsewhere, but the problem is formulated completely differently, it should also be awarded 1p.

- 0p – I chose "0p - Yes" for the question above.
- 0p – The ML problem is a straightforward variation. For example, it uses the same data points as an example in an assignment, but with different features.
- 1p – The ML problem is very original.