BIG DATA, MACHINE LEARNING, AND THEIR REAL WORLD APPLICATIONS

Jonathan Wayne Korn

2021-04-16

Project Guidelines

For the course project we will be splitting into groups and tackling a component of research to decide whether further develop into a potential distributed machine learning system. As discussed in the course a distributed machine learning model requires that you build/train/test multiple models targeted at different tasks to address a multi-level decision/problem. Each group will be tasked to develop and research a solution for that component of the distributed machine learning system. As a whole we will we will vote on a problem to address including sub-problems that requires at least 4 components *(machine learning models)* to provide predictions.

The following will guide each group to tackle their component and conduct proper research to help determine if further development in the distributed system would be possible.

(1) Define the Problem

. . . and any sub-problems that may exist within the scope of the component your group is tasked to research by the class comissioning.

Defining the problem will help you identify possible sources of data and any potential ideas on tackling it using machine learning. As a group address the following:

- What is the situation, problem or challenge you are addressing for your group's component?
- What preliminary examination leads you to believe machine learning could help?
- Provide a thorough background for the project; e.g. about the situation, problem or challenge, etc.

(2) The Data

After locating some potential sources of data, you should explore it to make sure it would be suitable to address the problem:

- Give a complete description of the data your group is using in your research, including any you rejected sources.
- Provide a detailed description of your data.
- Provide any exploratory data analyses including visuals, statistical summary, and correlation analysis.

(3) The Technical Development

- Give a detailed description of the process for your entire project.

- Given a detailed description of your approach to the algorithm you have proposed. You do not have to describe well known approaches themselves, *i.e. linear regression. You do have to describe how you applied the approach you when presenting the research*.

(4) Model and Evaluate

- Describe how you tested your approach to ensure that it is valid.
- Discuss the validity of your approach.
- Describe how you will evaluate your results and/or conclusions including any specific metrics, output data, completed machine learning modeling, etc.
- Discuss the baseline you will use to compare your results.
- Discuss how well your approach worked to address the situation or challenge, solve the problem or answer the research question.
- Discuss any potential future work. *(i.e. if you were not able to resolve the situation or problem or answer the research question what will it take to do so? What else needs to be done?)*
- Evaluate and report whether or not someone unfamiliar with your work could accurately replicate it.