CSCE 5063-001: Project Description

Due 11:59pm Sunday, April 21, 2019

The purpose of the final project is to deepen our exploration of machine learning with real-world data. To do this you will need to write code, run it on the data, make some figures, and write a few pages describing your task, the algorithm(s) you used and the results you obtained. You are free to use any online code or third-party sources as long as it is publicly available. However, please make sure that your own entries are developed by your team members - the purpose of this project is to give you practical experience, so it will not be helpful if you just follow instructions from others.

1 Form teams

Please form teams of 1-4 students who share your interests and with whom you will directly collaborate, and send me the list of team members by March 24.

2 Pick a topic

For the topic of the project, there are three basic options:

1. **Kaggle: Titanic: Machine Learning from Disaster:** Use the Kaggle competition dataset to predict which passengers were likely to survive in the Titanic sank, based on features such as class of travel, gender, age, etc.
   `https://www.kaggle.com/c/titanic`

2. **Kaggle: Digit Recognizer:** Use the MNIST dataset to correctly identify digits from a dataset of tens of thousands of handwritten images.
   `https://www.kaggle.com/c/digit-recognizer`

3. **Custom:** Find the topic and dataset on your own but make sure you select a task that can be executed reasonably before the deadline.

3 Choose a technique (or more) to explore

Your project should implement one or more machine learning algorithms and apply them to the data. A key point is that you must explore your approach(es); so you must do more than simply download a publicly available package and run it with default settings, or with a few values for regularization. You must at least explore the method fully enough to understand how changes might affect its performance, verify that your findings make sense, and then use your findings to optimize your performance.

4 Write it up

Your team will produce a single write-up document, approximately 3-4 pages long, describing the problem you chose to tackle and the methods you used to address it, including which model(s) you tried, how you trained them, how you selected any parameters they might require, and how they performed in on the test data. Consider including tables of performance of different approaches, or plots of performance used to perform model selection (i.e., parameters that control complexity).

Within your document, please try to describe to the best of your ability who was responsible for which aspects (which models, etc.), and how the team as a whole put the ideas together.

Please write the document in the NIPS format, which can be found here `https://nips.cc/Conferences/2015/PaperInformation/StyleFiles`.

5 Submission

You final submission must include your write-up and code. The submission deadline is 11:59pm on April 21 2019.

6 Presentation

Each team should give a short presentation about the project on April 23 or 25.

7 Grading

I am looking for several elements to be present in any good project. These are:

1. Exploration of at least one or two techniques in machine learning. For example, using neural networks, support vector machines, or random forests are great ideas; if you do this, explore in some depth the various options available to you for parameterizing the model, controlling complexity, etc. Other options might include feature design, or optimizing your models to provide good ROC behavior. Your write-up should describe what aspects you chose to focus on.

2. Performance validation. You should practice good form and use validation or cross-validation to assess your models' performance, do model selection, combine models, etc. Your write-up should describe how you assess your models' performance using tables or figures.

3. Adaptation to under- and over-fitting. Machine learning is not very"one size fits all" - it is impossible to know for sure what model to choose, what features to give it, or how to set the parameters until you see how it does on the data. Therefore, much of machine learning revolves around assessing performance (e.g., is my poor performance due to under-fitting, or over-fitting?) and deciding how to modify your techniques in response. Your write-up should describe how, during your process, you decided how to adapt your models and why.