Assignment 2

CS780/880: Introduction to Machine Learning

Introduction

The goal of the machine learning project is to get hands-on experience in independently defining, analyzing, and executing a machine learning (or data science) project. It is not necessary (but of course allowed) that the project conducts original research in machine learning.

To make sure that you make progress during the semester, there will be 5 separate deliverables. Each deliverable is a report (a paper) that describes the results and provides appropriate evidence. Please do not simply include a deluge of plots. Be brief and to the point. Only include the most relevant evidence.

The reports can be prepared as R studio notebooks, using LaTeX, Word, or any other typesetting environment.

Projects can be done individually or in teams of **at most 2 people**.

1 Project description and identification of data sources

Due: 2/23/2017

Describe the goals of the project and the data sources that you will use. This report should be at most **1 page** long using a legible format and it should address, among others, the following specific questions:

1. What is the problem? Is it prediction or inference? Is it classification or regression?
2. Why is the problem important?
3. What does success look like?
4. What are the data sources that will be used. Is it likely that they will suffice to achieve the goals?

Good places to look for data are, for example:

- https://www.data.gov/
- https://www.kaggle.com/
- http://worldclim.org

2 Evaluation Methodology

Due: 3/02/2017

Describe the evaluation procedure. This report should be at most **1 page** long. The questions that this report should address are as follows:

- What is the right metric for success? How good does it need to be for the project to succeed? For example, does the prediction error needs to be at most 5%? What about the area under the curve. Argue why.
- Will cross-validation be sufficient? Bootstrapping?
- How to make sure that the results are valid?
- Describe a plan for exploratory analysis

3 Method and Literature Overview

Due: 3/23/2017

Describe the method that will be used. Is it linear regression, is it LDA, or SVN, deep learning, reinforcement learning? Describe in as much detail as possible. Provide any exploratory analysis that supports the use of the selected method. What kind of features are available, do they need to be transformed, etc.

Also describe relevant literature. If the proposed method is new, describe the most relevant existing methods. If the focus of the project is on an application, describe previous work addressing the application. Describe what methods and data sets were used previously. The relevant work should be based in *peer-reviewed* research papers, books. A good resource is the Google Scholar search engine: http://scholar.google.com.

This report should be at most **2 pages** long. Make sure that sources are cited appropriately.

4 Preliminary Results

Due: 4/06/2017

Describe the results of the method. Describe how well the method did in the evaluation and compare with prior work (if applicable). Discuss what the results mean in the context of the problem definition. Is there anything that can be done to improve the results, or are they good enough? What about confidence in the results.

This report should be at most **3 pages** long.

5 Final Report

Due: 4/27/2017

Final report. The page limit is **7 pages**. Shorter report are also fine as long as they appropriately solve the problem.