Machine Learning - Final Project

CS 4731 — Dr. Mihail
Department of Computer Science
Valdosta State University

November 5, 2019

Final Project

The final project will be based on materials from this course. In this project you will compare and contrast two methods covered in this course. The methods can be something explicitly taught in this class, or may be based on something related that we did not cover. They do, however, have to fall in the same category of machine learning algorithms, either supervised or unsupervised.

In order to help with time management, I am requiring separate submissions for each of the following milestones:

1. Project proposal (5% of project grade)
2. Dataset selection (5% of project grade)
3. Script that loads dataset (or parts thereof) into memory (15% of project grade)
4. First method implemented and producing results (15% of project grade)
5. Second method implemented and producing results (15% of project grade)
6. Evaluation implemented (20% of project grade)
7. Writeup (20% of project grade)
8. Presentation (5% of project grade, 50% penalty on failure to present)

Due dates (all deadlines are on the date listed, before midnight):

Milestone	**Due date**
Project proposal	Sunday, November 10th
Loading dataset script	Wednesday, November 13th
First method	Sunday, November 17th
Second method	Sunday, November 24th
Evaluation	Wednesday, November 27th
Write-up	Sunday, December 1st
Presentation	TBA

Late submissions

Milestone deliverables with late submissions (even less than 24 hours) will receive 0% of credit. Most of the milestones are prerequisites of each other, so skipping one or more will have a cascading effect that can lead to failure of the project.

Proposal

This document shall be at most one page, single spaced, 1 inch margin document that describes the following:

- The idea of the project and your motivation.
- The work you plan on doing (e.g., programming, data processing, etc.)
- Clear definition of the problem you're addressing and its motivation (why the problem is important) and a plan to solve the problem.

Dataset

You have to identify the dataset you will work with. A minimum of 1000 total data points are required.

Loading the dataset

For this milestone, you will submit a script that loads the dataset and splits the data points into two sets: training and testing. The training set will consist of 80% of the samples and the testing set will have the rest. Your script is required to print the first 10 data points (with all their attributes) from each set (printing in stdout is fine).

First method

Submit a script that implements the first method. Your script has to train on the train set and test on the testing set. You have to evaluate this method.

Second method

Submit a script that implements the second method. Your script has to train on the train set and test on the testing set. You have to evaluate this method as well.

Evaluation

Modify the script such that an evaluation metric is reported on the test set. For example, in a classification problem, misclassification rate is a useful metric to evaluate the method. This has to be done for both methods.

Writeup

The final report shall consist of 2 single spaced pages of text per person following the same format as your proposal.

Project Policies

1. Each project is to be completed by groups of 1 to 3 students.
2. The members of the team have to contribute roughly equally to the completion of the project.

Project Ideas

There are many interesting projects you can pursue. I prefer you pick something you're personally interested or passionate about. Creative and novel ideas are preferred, even though you may not get awesome results. Some ideas below:

- Comparing methods or algorithms. Many times algorithms need to be adapted for certain datasets and often those adaptations are interesting.
- Missing information. Dealing with missing information is often a necessity in real life. Possible venues of research include how algorithms handle missing data and how algorithm performance degrades with increasingly more missing data.
- Convergence. How long does it take for your method/algorithm to converge? What are some possible ways to speed up convergence and performance trade-offs?
- Novel applications. Has a particular algorithm ever been applied on a dataset? Is there a reason why or why not?

More practical project ideas include:

- Opinion classification. Determining the polarity (negative or positive) of social media posts.
- Decoding brain signals. Magnetic fields from the brain can be used for various tasks, including game controllers.
- Semi-supervised learning. In some cases, large amounts of unlabeled data is easy to obtain but difficult to label.
- Privacy and security. How can we provide security without compromising too much privacy? Can ML algorithms be adapted to make this trade-off?
- Image segmentation. How can we label semantically similar pixels in an image?
- Object detection. Can we detect an object in an image?
- Voice recognition. Can we translate audio signal to text?