

Spring 2020 Project

Machine Learning (CS-GY 6923)

Exact Date TBD
No late assignments accepted

You may work together with one other person on this project. If you do that, *hand in JUST ONE project for the two of you*, with both of your names on it.

The goal of this project is for you to explore how you could improve your implementations of machine learning algorithms.

1 Guidelines

You will make *three extensions* to one or more of the algorithms you implemented in a homework assignment.

Here is how to make *one* extension (the other two extensions are the same):¹

1. Select a machine Learning algorithm you implemented in a homework assignment
2. Learn how to use Scikit-learn's implementation of this algorithm (make sure you set the parameters so Scikit-learn's implementation is close to your homework assignment's implementation²)

Find an *extension* of the basic algorithm that Scikit-learn implements that improves accuracy on one of dataset we used in any of our assignments, in a lecture, or from the list posted on Piazza by the GA's for the course. (An example of an extension would be adding lasso regularization to logistic regression)

3. Implement this *extension* by adding code to your homework assignment implementation.³ (Note: if Scikit-learn's implementation with the extension showed an improvement in accuracy, but your implementation didn't show the same or any improvement with the same extension, that is ok. Please comment on why you thought this was the case)
4. Select a new dataset of your choice and show this data set's accuracy using your homework assignment implementation, and with the new extension you just added

¹You might need to try many different combinations of: algorithm×extension×dataset until you find one where the extension improves the accuracy. Do *not* start step 3 until you have found an extension that improves the accuracy on a dataset.

²E.g. To learn how to use Scikit-learn's implementation of logistic regression you could read https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

³You will write the additional code you add to your existing assignment from scratch.

2 Details

1. If Scikit-learn does not support your algorithm, you may consider other libraries (example: PyTorch, Tensorflow).
2. When implementing the upgraded algorithm, be sure to:
 - not use any external libraries aside from Numpy.
 - do not copy the existing open-source implementations offered by existing machine learning frameworks
3. You will turn in your (working) python notebook and a written report (at most 2 pages for each improvement). In addition to reporting the accuracy, your report should give the rational on why your addition improves our existing algorithm, both from a high level, and then in more detail. In your report, make sure you have a chart to show the accuracy obtained for the different datasets and the different algorithms/implementations

Your grade will depend on:

- the complexity of your extension
- how much you improved in accuracy on a dataset from our existing algorithm
- how well you explained the reason behind your improvement, and your understanding of why your code works