Final Project

CS-UY 4563 - Introduction to Machine Learning
Spring 2022
Due Date: To Be Announced[*]

You and your partner need to submit the following documents to NYU Classes.

- Your presentation slides
- Your project write-up
- Your project code (GitHub link is ok)
- Your dataset - if you made one from scratch

Please note that we will be passing your code and write-up through a plagiarism checker. We know there are ways to cheat on the final exam project. If we suspect you of cheating, you will receive a 0 for your final project grade.

Following are some datasets you can use for your project. Feel free to use any other datasets or even make one on your own:

- https://vision.eng.au.dk/plant-seedlings-dataset/
- https://build.kiva.org/docs/data
- https://www.kaggle.com/competitions

Please send us a link to the data set you will use and who your partner is by April 8th.

Guidelines for your writeup:

1. **Introduction:** You will briefly describe your data set and the problem you are trying to solve.

[*] The projects will be due the evening before the first presentations are given. The date the project will be due depends on the number of presentations. The current estimate is the project will be due April 25 or April 27.

2. **Perform some unsupervised analysis:** Look to see if there is any interesting structure present in the data. If you don't find any interesting structure, describe what you tried.

3. **Supervised analysis:** You must try at least three of the learning models discussed in the class (e.g. Logistic regression, SVM, Neural Networks). For each model you must try different *feature transformations* and different *regularization techniques.* For example, try the linear, polynomial and radial-basis function kernel if you are using support vector machines in your project. Don't forget to illustrate (through graphs) how your feature weights, and error changed when you used different parameters, regularizations and normalizations.

4. **Table of Results:** You *must* create a table that contains the final results for your model. It would be useful to have a table that contains the training accuracy, and the testing accuracy for every model that you created. For example, if you're using Ridge Regression and you're manipulating the value of λ, then your table should contain the training and testing accuracy for every value of lambda that you used.

5. **Why:** Your write-up should analytically discuss your experimental findings; and what conclusions you are able to draw from your work. You should appeal to the concepts discussed in class: *overfitting, underfitting, variance, bias, etc*

You and your partner will give a six minute presentation to the class. The final project presentations will be held during the last 3 or 4 class periods, and during the final exam period for this class. You will be assigned a day for your presentation. If we run out of time the day you are to present your project, you will present the next day reserved for presentations.

You are required to watch the other students' presentations in class. A large part of your project grade for the project will be based on your attendance for everyone else's presentation.

If you have a project idea that doesn't satisfy all the requirements mentioned above, please inform me and we can discuss its viability as your final project.

Practical advice can be found in chapters 1 and 2 of Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, and in `http://cs229.stanford.edu/notes2020fall/notes2020fall/CS229_ML%20advice_presented-slides.pdf` Please Google appropriate topics. A quick first glance at some of these topics can be found here:

- Dealing with unbalanced datasets: `https://www.svds.com/tbt-learning-imbalanced-classes`
- Preparing your dataset: pages 62-69 Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

- Working with time series: `https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm`
- Handling the missing features in training set: `https://web.stanford.edu/~lmackey/stats306b/doc/stats306b-spring14-lecture16_scribed.pdf`