C142/242: Machine Learning, Statistical Models, and Optimization for Molecular Problems

Project Assignment

Assigned March 18 and Due May 1

Undergraduates: For the final project we will develop a supervised learning ANN model applied to the ANI-1 data set. We will do a check-in once per week to see steady progress with appropriate entries of dates in the jupyter notebooks on what has been accomplished. This will be part of the assessment. I.e. this is not a project assignment that should be finished the night before.

Grading breakdown:

Part I: An individual jupyter notebook should be maintained during the course of the project. There will be 5 progress check-ins, each 15% of the grade. You'll submit your notebook to Gradescope at each time point.

(1) (April 1) Data preparation. Show that you're able to load the data, process them into model input format, and split the data into train, validation and test set with batching.

(2) (April 8) Network construction and workflow development. At this point you should have a working code that can train the network, demonstrated on small subset of the data.

(3) (April 15) Regularization strategies and hyperparameter tuning. Use more data to train the network. Play with the architecture, hyperparameters and the regularization strategies. Show your work and defend the final choice of your model.

(4) (April 22) Final production mode. Multiple runs and N-fold cross-validation.

(5) (April 29) Final results. Train your model with all data and compare your results to what's reported in the paper. Finish off Jupyter notebooks and organize results.

Part II: (Due May 1) Submit an individual report on results along with your jupyter notebook. Written reports by the undergraduates should have the following sections: introduction, methods, results, and discussion, and be a minimum of 3 pages written text and 1-2 pages of graphs/figures (25% of grade). Further detail is provided in guidance document.

Graduate students: For the final project apply at least 2 unsupervised and/or supervised learning techniques on a bio/chemical problem of your choice. The finals project has the following expectations for assessment:

a. demonstrate best practices in regards data preparation and analysis of data before starting any subsequent learning strategy; this should be evident in the reporting of results.

b. Demonstrate all necessary regularization techniques, understanding of parameter tuning, bias-variance tradeoffs etc.

c. Report results on your final production runs and comment on whether your goal is achieved in the end.

Grading breakdown:

Part I: An individual jupyter notebook should be maintained during the course of the project to show your process of thinking and experimenting. There will be 6 progress check-ins, each 10% of the grade. You'll submit your notebook to Gradescope at each time point.

(1) (March 21). Project goal write-up. Consult with Prof. Head-Gordon and Joe about ideas for a project on March 20 in class; send a ½ to 1 page write up to thg@berkeley.edu and cc Joe.

(2) (April 1) Data preparation

(3) (April 8) ML model construction and workflow development.

(4) (April 15) Regularization strategies and hyperparameter tuning

(5) (April 12) Production mode

(6) (April 29). Final results; Finish off Jupyter notebooks and organize results

Part II: (Due April 29 or May 1) Final presentation on April 29 or May 1. (40% of the grade) This is expected to be a 10 minute presentation to the class. Turn in final form of your Jupyter notebook and ppt presentation. Further detail is provided in guidance document.