EE 660 **Type 2 Project: Examples and Topic Ideas** Fall 2020
Jenkins

Overall Description

For an overall description of Type 2 projects, and a suggestion on how to devise a topic of your own, see the Project Assignment, pp. 1-2. Below we give some examples and ideas for Type 2 Project topics.

Examples and Ideas

1. A good example of an experiment is Sec. 4.1.2 of AML, especially Exercise 4.2, including the results shown in Fig. 4.3 and some of its interpretation.

   Comment: this is meant primarily as an example, not as a project topic suggestion. If you want to use this topic for your project, first consider how you would change it or add to it. Your project would need to be significantly different (or add significant new material).

2. You could consider addressing the following question: in a model selection procedure (typically finding the best model by a grid search of parameter values using a validation set or cross-validation): what is the probability that the procedure will pick the model (based on validation-set error) that actually is the best model (based on out-of-sample error)? How does this probability depend on different parameters like N_{val}, number of models (or number of parameters and number of values tried for each parameter), complexity of each model, etc.?

3. It is generally accepted that although the VC genearlization upper bound can be very loose, its theoretical dependence on the variables tends to hold for actual out-of-sample generalization error. Design an experiment, or set of experiments, that can test this generally accepted statement.

4. Test the *curse of dimensionality*. Design an experiment, or set of experiments, that will test different ML methods to see how they vary with changing number of dimensions. For example, how much data is required to get a specified performance level, as a function of the number of dimensions? Is it exponential in the number of dimensions? Why, or why not? This will depend on the ML algorithm, and will also depend on the probabiliy distribution of the data that is synthetically drawn.

5. We always assume that the data points are independently drawn. What if they are not? Design an experiment, or set of experiments, to find out. Start by thinking about what affects might be likely to happen, and design experiments to observe, verify, or refute that. Then, what else might you try?

6. If you would prefer to use real-world data (instead of, or in addition to, synthetic data), then the UCI repository is a good source; many of its datasets have little or no preprocessing and feature extraction left for you to do, so those can be used directly in your experiments. But of course, you won't be able to measure out-of-sample error directly. See item (3) of "Tips for Finding Appropriate Real-World Datasets".