Project Workflow
COSC425, Fall 2020

This document serves as a series of steps that you should follow for your COSC425 project assignments. These steps are not strictly linear and may overlap. Their purpose is to illustrate a general procedure for completing your project assignments. (**Note:** Each step may not be relevant to all project assignments.)

Step 1: Data Exploration

1. What is the format of your file? Are columns separated by commas?
2. Is there a listing that describes the names and types of your features and labels?
3. What are the types of your features? Integers? Real Numbers? Boolean? Categorical?
4. Are there missing values in any of your feature data? Are there potentially incorrect values?
5. Document what you have discovered.

Step 2: Data Preparation

1. Compute descriptive statistics on your data (i.e., mean, SD, min, max, quartiles).
    - This will help you identify missing and anomalous values.
2. Devise and apply a strategy for fixing missing / anomolous values in your columns (e.g. NaN).
    - Document and explain your strategy.
3. Convert categorical / nominal features to numerical / binary features.
4. Standardize your data where appropriate (e.g. normalization).
5. Document all of the above.

Step 3: Dimensionality Reduction

1. Consider if you need to reduce the number of dimensions in your data.
    - You may want to start without applying any reduction and retroactively return to this step after exploring Steps 4-8.
2. If you do apply a reduction algorithm, document which method you chose to apply and your motivation.

Step 4: Implement the Algorithm

1. Start with an implementation that is quick-and-dirty, but also correct.
2. Verify its correctness with made-up data for which you know the answer.
    - Make note of opportunities for performance enhancements you can explore later.

Step 5: Data Separation

1. Separate real data into training, validation, and test sets.

Step 6: Train the Model

1. Train and cross-validate your model.
2. Compare training and cross-validation error for various hyper-parameter values.
    - If you have a regularization or complexity parameter, plot training and CV error against it and look a the "elbow".
3. Identify and select the hyper-parameters that give the best generalization.

Step 7: Test the Model

1. Test the trained and validated model on the test data from Step #5.

Step 8: Interpret the Evaluation

1. Examine your results as a collective. Do they make sense?

Step 9: Improve the Model

1. Now that you have a functional model, consider how to improve your model through two lenses:
    - ***Implementation:*** Improve your implementation by vectorizing your data or using optimized libraries.
    - ***Training:*** Improve your training procedure by reducing dimensionality or further exploring parameters.