/hubwayPredict
Code Review Preparation and Framing

Background
Our project is a predictive model for Hubway bike sharing use in Boston.

Key Questions
- How effective is our file arrangement / module system?
    - See what other teams are doing for their organization system
- What ways can we determine similarity between data sets?
    - Comparison of regression lines vs. individual data points
    - Normalized data vs. raw difference comparison
- Supervised or unsupervised learning?
    - If supervised, do we want to use decision trees or a support vector machine?
    - If unsupervised, *k*-means clustering, mixture model, or hierarchical clustering?
- Genetic algorithm?

Agenda
I. Project Introduction (2 min)
    A. Agenda and Objectives (1 min)
    B. Hubway Data Prediction (1 min)
II. Detailed Code Discussion (6 min)
    A. Weather Data (2 min)
    B. Hubway Ridership Data (2 min)
    C. Pretty graphs (2 min)
III. Questions (15 min)
    A. Clarification Questions (5 min)
    B. Feedback Questions (10 min)

Glossary
Supervised Machine Learning: Given labeled training data, the algorithm analyzes the data and generates a function that can be used to map new examples.
Unsupervised Machine Learning: Given unlabeled data, the algorithm will try to find some semblance of structure in the data. There is no "error" or "reward signal" to tell the algorithm whether or not it has succeeded.
Decision Tree: Takes observations about an item and maps those to a specific related conclusion (i.e. If you were on the Titanic and not of the male sex, you had a 36% chance of survival. If you were of the male sex and older than 9.5 years old, you had a 61% chance of dying.)

Support Vector Machine: Given training data, an SVM algorithm builds a model that separates the data into discrete chunks. Given new data, the SVM will see how well that new data fits each discrete chunk and categorize it as whatever chunk matches best. (i.e. If you feed an SVM 100 pictures, 50 of which are tigers and 50 of which are elephants, the SVM will group the elephants and tigers as 2 discrete chunks. Given a new picture of an elephant, the SVM will find that the picture of the elephant matches the elephant chunk most, and categorize the new picture as an elephant.)

K-means Clustering: Groups data into a specific number of clusters, minimizing the distance to the centroid (average) of the cluster's data points.

Mixture Model: a model for creating a probability that a subpopulation exists within a larger population and attempts to characterize the subpopulations.

Hierarchical Clustering: a method of finding clusters while also creating a hierarchical pattern to the clusters. Two methods of hierarchical clustering: agglomerative starts with all clusters in separate hierarchies and begins to group them in partners, divisive starts with all clusters in the same hierarchy and splits them off into different groups.