

SEIS 763 Machine Learning
Assignment 4
Due: midnight 11/3/20 on Canvas

Team Activity

Since everyone has been assigned to a team now, the purpose of this assignment is to decide on a dataset for your team to work on and **submit a project proposal**.

As detailed in the project handout, you should consider the following when writing the project proposal:

- **Dataset size:** What data set is being used - where does the data come from, and what are some characteristics of it (size, missing values, continuous vs. categorical). You should have a minimum of **3000 instances**.
- **Class imbalance:** If you are doing classification, see whether the classes are appropriately balanced. If not, come up with a strategy to overcome the class imbalance.
- **Attributes:** It is preferred to have a dataset that has more numerical attributes compared to the categorical ones. Avoid picking a dataset that has too many categorical attributes.
- **Exploration:** Before you pick one dataset, it might be best to explore the dataset a bit to see if there are any anomalies, missing values, outliers, etc. You can include those exploration results in your project proposal.
- **Hypothesis:** What is the question(s) of interest - be specific. You should avoid generic questions like "*I want to look for patterns in stock prices*". Devise a specific question you can answer with the data.
- **ML Techniques:** What methods do you plan to use - understanding that this might change and that we have yet to cover many methods in class.

Submission: Any one person from the team should submit the project proposal online by the deadline.