CS 4780: Machine Learning	Project Proposal

The Team

We have four members in our team.

Motivation

Previously, we trained machine learning algorithms on relatively small amounts of filtered training data and performed simple classification on test cases. The project that we have selected, General Electric Company's Kaggle competition, presents us with the opportunity to apply what we have learned to a structured output problem. Our motivation to complete this project comes from our desire to overcome the challenges involved with solving a different type of problem than what we are used to (such as working with sizable amounts of data, some of which may be irrelevant) in order to produce an optimization algorithm that can be used to make future real-life processes more efficient.

The Task

The main objective of the solution is to present an optimized flight plan to the pilot, to minimize costs given the training set, which contains information about flight and weather. We are provided with a cost function in a simulator to evaluate the expense undertaken to complete a flight path.

- Since the flight has innumerable variables that can be tweaked to reduce costs, specifying potential suggestions (in other words, the output variables of our solution) is possibly the first design problem that needs to be solved.
- Dealing with data larger than system memory
- Study the feasibility of distributing the load among several clients to increase the speed of learning/prediction
- Generating an optimal path rather than classifying a single instance (structured output)

General Approach

1. We will prune the input data to reduce complexity by assigning weights

based on the relevance of the features. For example, customer dissatisfaction could be considered much less than weather conditions.

2. As the input data is too large to directly work with, we will handle parsing our data using an online learning algorithm.
3. By splitting the data into more manageable chunks, we can construct trained models on subsets of the total data in order to speed up the training process and run our learning algorithm concurrently.
4. We will have a unifying method for the models to reduce time spent on recomputation and allow for quick modifications on certain data sets or partial retraining.
5. We will use the model constructed along with the parameters given to determine the most optimal flight path from a source to destination.

Resources

- Data sets provided to us by the GE Kaggle competition site
- Software:
  (a) FlightQuest Simulator (FQS) provided to us by the GE Kaggle competition site to produce regression values
  (b) Microsoft Visual Studio 2012 to used for read the FQS source code
  (c) Python 2.7 (and libraries) to develop our data processing and machine learning algorithms
- Readings on structured output.

Schedule

- Nov. 1: Complete research and have understanding of tools and design a basic approach to handling huge dataset.
- Nov. 8: Have a working method for generating paths while ignoring most factors
- Nov. 22: Have methods implemented to take into account different factors
- Nov. 29: Begin final report and poster
- Dec 7: Finish refinements on code and report