| CS 4780: Machine Learning | Project Proposal |
|---|---|

# 36: Kaggle GE Competition

## 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 problem statement is that we are given flight information (point of origin/destination/other flights arriving at the same time, etc) and we would have to create paths. Using this, one has to provide the most profitable route for the pilot to follow. We would have to take into account unforseen circumstances and come up with a solution that is robust to withstand eventualities in the future. This leads to a learning problem that is more complex than those encountered in a typical machine learning setting. Adding to the complexity in the problem, the data set is quite rich in information and dimensionality. An ideal algorithm would have to bring out the hidden dependencies in such unrelated data.

If the problem, with such layers of challenges, if solved effectively, its application would bring down costs of flight travel. Development of novel methods to unravel previously intractable problems is one of the leading motivations for us to take up this problem.

## 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 paths actually taken and the weather. We are provided with a cost function in a simulator to evaluate the expense undertaken to complete a flight path given predicted weather and ground conditions. True evaluation can only be achieved by submitting to the contest website, in which five

submissions are allowed per day.

- 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. How we have structured the problem is that we begin from the most direct path to our destination. We then tweak this path in order to satisfy competition rules such as avoiding restricted areas. Once we have our base path, we go along it and check at certain waypoints if deviating from our base path will produce a more optimal solution.
2. We will determine what constitutes a more optimal solution by prioritizing conditions that optimize our fuel consumption via flight speed and altitude (such as turbulence and jetstreams). Conditions that are difficult to decipher and do not provide as much information, like weather, will be considered later if time permits while more irrelevant information, like customer dissatisfaction, will either be weighed less as features or not be considered as features at all.
3. We will model all of the conditions we consider separately using Markov models. In order to do this, we will train on the data given to us and attempt to associate the existence and severity of the condition with a given set of coordinates. We will also be looking into regression techniques in order to generate altitute and airspeed output.
4. As the input data is too large to directly work with, we will transform the data to make it more tractable. 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. Also, by doing so,

we can consider each of these models separately and then determine the relevance of each parameter.

**Progress**

1. We have already managed to generate a naive model where it just tries to coast at a consistent alititude with a consistent airspeed in a great circle path between two points.
2. We are currently working on determining an optimal airspeed and airspeed based on our input data files. After that, we will be trying to extend that predicting the best movement to make at each point in time based on the the training data. We must validate to see which things are important in determining these parameters.

**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
  (d) SciKit Learn for various learning algorithms using regression
  (e) GeoPy for handling spherical coordinates of Earth.
- Readings on structured output.

**Schedule**

- Nov. 24: Have a few methods of training, have data transformed
- Nov. 26: Work on predicting effect on flight speed and altitude by progressing through a certain coordinate
- Nov. 29: Begin final report and poster
- Dec. 3: Finalizing our overall path-prediction algorithm.
- Dec. 7: Finish refinements on code and report