EECS 491 Project

William Koehrsen wjk68

March 26, 2018

1 Applying Probabilistic Classification Models to a Machine Learning Competition

1.1 Problem Introduction

Numerai is a hedge fund that trades based on the results of weekly machine learning competitions. This strategy relies on the wisdom of the crowd and the democratization of machine learning to create a viable trading platform. Every week, Numerai releases a new set of training and testing data to the public and the best performing model predictions (as measured on a validation set) are selected to be used in live stock market trading. Contributors have the chance to earn a cryptocurrency, Numeraire for their predictions based on performance in the market. The competitions run by Numerai are set apart from other machine learning competitions in that competitors must place a stake on their predictions representing confidence that their model can generalize to new data. In theory, this should reduce the common problem of overfitting in machine learning competitions where models are highly tuned to maximize scores on the test set.

2 Project Objectives

My project will apply Probabilistic Classification methods to the Numerai weekly machine learning compeitions. Most winners of these competitions use standard models such as Logistic Regression, Gradient Boosted Classifiers, or Deep Neural Networks. However, based on the discussions I have read, there is little, if any, use of probabilistic programming methods. I intend to use Gaussian Process Classification in PyMC3 to make predictions and submit them to one or several weekly competitions. I also want to examine using Bayesian Neural Networks in PyMC3 for classification. The competitions run every week, which will give me several attempts to experiment with different models. I have had some limited success using standard methods and want to see if there may be some advantage to using probabilistic models for classification. The plan for this project is as follows:

- Develop several standard machine learning classification models as a baseline

- Develop probabilistic classification models using PyMC3
- Submit predictions to Numerai weekly competitions and compare
- Iterate on models to see if Probabilistic Models can be competitive with standard machine learning models

2.1 Data

The competitions are a supervised classification task: the datasets contains features corresponding to market indicators and binary labels corresponding to buy/sell choices. Every week, Numerai releases a new set of training and testing data for a new competition. The data is completely de-identified which means the features can not be linked back to real-world indicators. The training dataset contains several hundred thousand observations, each with a corresponding binary target. For the testing set, there are three separate tags:

1. Validation: contains actual labels which can be used by competitors to evaluate their model
2. Test: used to determine the leaderboard on Numer.ai. Labels are not provided for these samples.
3. Live: predictions made on these samples are used live in the stock market. No labels are provided (as they are not available) for these samples.

Predictions submitted by competitors are to be made in the form of a decimal between 0 and 1 rather than 0/1. Competitors are ranked on a leaderboard based on the test scores and are paid based on how much they stake and performance on the live predictions which are used in actual trading by Numerai.

2.1.1 Metrics

The primary metric for evaluating the performance of a model is log-loss. Predictions are ranked on the leaderboard in terms of log-loss on the test set.

However, in addition to log-loss, to be considered on the leaderboard, the model must meet the following 3 criteria:

1. Consistency: The percentage of eras (an defined grouping of samples) in which a model has a log loss < - ln(0.5). A model must achieve greater than 75% consistency to make it on the leaderboard.
2. Originality: A measure of the correlation of predictions with previously submitted predictions from other users. This is to encourage novel models but also has the side effect of favoring earlier submissions.
3. Concordance: A measure of if predictions made on the validation, test, and live data set are from the same model. Predictions should be from one model to prevent overfitting to the validation data.