Real Time Decision Making on the Kaggle Platform

Michael Kim, Virginia Tech / Booz Allen Hamilton

Motivation

Data driven forecasting can be applied to weather, energy, sales, sports, and stocks.

Can we improve forecasting using real time data streams, crowdsourcing (Kaggle-like platform), and machine learning?

Idea

Data scientists are given a continuously (perhaps daily) updated stream of data with ground truth.

Given the data, make a one time step ahead forecast (can be regression or classification) that minimizes some loss function.

Major Issues

How to give back to users (Luis von Ahn) so that effective Human Computation is possible.

How to ensemble predictions. The best single prediction may not have the best marginal contribution to an ensemble. Users may not always submit predictions daily.

How to build a pipeline system with the client's goals in mind so that the platform solves the client's problems.

Proposed Solutions (HC)

Give back money to users proportional to:

1) marginal leave user out (of ensemble) loss.
2) mean (weighted) forecast loss of individual user's prediction.
3) some linear combination of participation rate and 1) and 2)

Proposed Solutions (ensemble)

1) Allow users to see historical predictions with anonymous user ids. Make the users create ensemble code, and select the best via CV.
2) Use a weighted mean or median of the highest scoring data scientists.
3) Ensemble via glm, gbm, neural net, or random forest.

Proposed Solutions (Pipeline)

The Kaggle platform could be modified for real time crowdsourced forecasting and ensembling. Top Coder is another possibility.

Work still needs to be done in building ensembles given data streams and crowd based predictions. A possible system (Machine Learning on Streaming Data with Storm and MOA) was build by my colleague Paul Yacci (Informs 2014 Big Data).

Conclusions and Future Work

Build out a system that allows real time, crowd sourced forecasting and ensembling. Open source the code and put it on Github.

Test the viability of the system in the wild.

Improve the system given user feedback.