How to Run a Data Mining Competition

Professor Susan Holmes and Nelson Ray
Department of Statistics
Stanford University

April 24, 2011

1 Type of Competition

The first step in organizing a data mining competition is deciding on the type of the competition. Netflix sought to improve their movie recommendation engine by 10%. For teaching purposes, the challenge should match the students' abilities (but never understimate them!) and the level of the class. In a class on linear models, it is best to stick with prediction of a continuous response variable and not attempt to, say, cluster image data. Once the overall type of competition (e.g. supervised versus unsupervised, entirely quantitative data or heterogeneous), has been determined, it is time to find a suitable dataset.

2 Getting Data

There are many excellent repositories available online providing easy access to a wide variety of datasets. Here are just a few to get you started.

2.1 Data Repositories

- UCI Machine Learning Repository (199 mostly multivariate datasets suitable for classification and regression problems, among others)
- U.S. Federal Executive Branch datasets ranging from EPA toxic release data to VA health data
- Infochimps Data Marketplace (huge variety of both free and non-free data)
- CMU Datasets Archive
- A list of about a dozen (social) network datasets.

2.2 Scraping the Web

One of the downsides of using widely available data in a competition is that many datasets have already been publicly "analyzed to death." Fortunately, it's quite easy to scrape the web for novel datasets. Scrapy is a Python-based application framework for crawling websites and extracting structured data. The Twitter API makes it easy to collect tweet data.

3 Hosting

Kaggle-in-Class takes care of all the little details in actually hosting the competition. It is very easy to upload test and training datasets and customize the metrics on which the participants will be judged as well as tweak the frequency of allowed submissions. Detailed contest instructions can be uploaded, and the site takes care of user sign-ups, a forum for contest discussion, and a live leaderboard displaying each team's performance.

4 Class Integration

How should the students be incentivized to participate? Will the competition be a core component of the course grade or extra credit? It's a good idea to integrate the competition with the course. Displaying the leaderboard at the beginning of class is exciting and will motivate the top performers. Giving hints and sample code yielding improvemed performance will encourage participation from the entire class in catching them up and will drive the leaders to work more.