Applied Mathmamatics 221

Project Milestone 1: Proposal

David Freed and Sam Green

February 24, 2016

1 Problem Statement

In this project, we intend to model the predictive properties of Twitter data relating to outcomes of high visibility basketball games in the NBA or NCAA. Previous work has applied tools from statistical analysis to establish that twitter data can be used to predict outcomes of games in sporting events. [1] Previous work that we have discovered, however, leaves unanswered questions about how the predictive quality of Twitter data changes over time. In this project, we will investigate the shape of the "predictive curve" for tweets, with the goal of answering the question: "At what point in a game do tweets become an accurate predictor of outcomes?"

The project has a fundamentally applied goal but notheless depends substantially on optimization theory. We intend to build a predictive model that investigates outcomes of games, but the main question of interest is rooted in sensitivity analysis. Rather than simply asking "can we use twitter data to predict outcomes?", we wish to perform sensitivity analysis that characterizes when the model we produce becomes useful.

The pool of questions we can pose is also very extensible. After building a baseline predictive model, we can alter the character of the input data (limiting to tweets from specific types of Twitter users, correlating predictive quality of tweets with geolocation, testing the number of tweets necessary for the model to become predictive.)

2 Deliverables

2.1 Timeline: Mid-March

- Twitter data collection script written (using the Twitter Streaming API)
- Game data sourced, and collection script written (game data must be converted to time series)

2.2 Timeline: End of March

- Data collected (Twitter and Game)
- Sentiment classification script for Tweets
- Baseline expert prediction set created (for benchmarking)

[1] Previous work on the NFL: `https://www.cs.cmu.edu/~nasmith/papers/sinha+dyer+gimpel+smith.mlsa13.pdf`

2.3 Mid-April

- Statistical Model in process or complete.
- Preliminary sensitivity analysis & results.

2.4 End of April

- Robustness checks.
- Final Analysis.
- Extensions.

3 Collaboration

This project will be completed by David Freed and Sam Green. We intend to work collaboratively on all aspects of the project and will divide specific implementation tasks based on our previous experience as they arise.

4 Data

5 Deliverables

6 Next Steps