Applied Mathmamatics 221

Project Milestone 1: Proposal
David Freed and Sam Green
February 24, 2016

1 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.

2 Problem Statement

3 Data

We will primarily collect two types of data for this project: Twitter-based data and outcome-based data. As illustrated in the above problem statement, our goal is to create connections between the two and see how good Twitter is as an inferential model (i.e. given past game outcomes, how good is Twitter at predicting future outcomes).

3.1 Twitter Data

Our primary aim is to collect data from Twitter regarding an actual game. We will sort our data by time, date, and hashtag. Since we are interested in overall sentiment during the game, we will limit our search to tweets between tipoff and the final buzzer, data that will come from the following dataset.

We will further filter by hashtag/account to get relevant Tweets. In a game between the Cleveland Cavaliers and Golden State Warriors, for example, we might look at all Tweets containing #GSW, #Cavaliers, #Dubs, #GSWvCLE, and any number of relevant hashtags. It is not clear how we will select appropriate hashtags?with luck, we can create a script that will automate this process effectively.

For each tweet, we want to grab the following information:

1. User characteristics (tagline, number of followers, etc.)
2. Tweet characteristics (time, location, no. of retweets/favorites, etc.)
3. Sensitivity score

The last item will be the most salient characteristic of the Tweet–we want to know how the user is feeling about the game. We will get this sensitivity score by developing our own natural language processor. In order to do this effectively, we hope to borrow from the literature here.

3.2 Box Score Data

To connect this Twitter data with outcomes, we need to get a set of outcomes. We will create a simple scraper for box scores so that we can get the outcomes for all the games that we study. We envision scraping sites like `http://espn.go.com/nba/playbyplay?gameId=400828584` to get the following data for every game:

1. Time of every score change (so as to keep track of score over time)
2. Information about each play (who made which play)

Eventually, we will try to map the time that the event occurred in the game to when it appeared in real life. This will allow us to get a sense for what information is included in Tweets and what types of plays prompt people to get excited on Twitter.

4 Deliverables

5 Next Steps

Our next steps all have to do with acquiring data. We break it down here into steps needed to acquire Twitter data and steps needed to acquire box score data:

Twitter Data Next Steps

1. Get Twitter API keys
2. Identify the best ways to get requests and the structure for streaming data from Twitter
3. Read the literature on sensitivity analyses on Tweets
4. Get trial data (ideally with location tags attached)

Box Score Data Next Steps

1. Create a scrape to get all of the box score data
2. Write a win probability model

The win probability model will likely be a replication of what people already do online and will provide a reference point for how good Twitter is at predicting games. We might initialize the model to the Vegas line or some other ex-ante predictor of performance (Pythagorean estimates are a commonly accepted way of measuring team strength).