Tennis Predictions
GA Data Science - Final Project Outline

Rohini Pandhi
February 13, 2014

1. Problem to be solved

Provide predictions on tennis match outcomes and who will win tournaments based on various stats/features.

2. Description of dataset

- What dataset will you be using for your project? **http://www.tennis-data.co.uk/wimbledon.php**
- How will you access this data? API? Scraping? Other? **CSV downloads available**
- What sort of information is in the dataset? In other words, what features are available? **http://www.tennis-data.co.uk/notes.txt**
- How will you turn that data into a training set? (If using a supervised approach) **We know the actual outcomes of the tournaments so I will use specific matches for training data and then apply them to predict other already completed matches.**
- How do you anticipate processing that data to get it into a form to use for your modeling? **Most likely, the data will need to be parsed from its CSV format and/or stored in a database and queried from there.**

3. Hypothesis

What is your hypothesis? In other words, what do you hope to predict or otherwise learn as the outcome of your project?

My hypothesis is that the stats typically captured for matches are great for choosing favorites in tournaments, but predicting wins require additional criteria/features (e.g., weather, injuries, coaches, age, etc.). So I will base my prediction model on the features available to start.

What are some of the features you might use?

To begin, the features that will be included in the classification will be: tournament surface, player ranking, tournament venue, round of match, stats of wins/losses/ties, etc.

4. Statistical methods I plan to use and why

Think back to our 2x2 matrix slide: Is your problem a classification problem or a regression problem? Will your approach use supervised approaches or unsupervised approaches? **Classification and supervised.**

Which of the machine learning algorithms that we have learned do you plan to use for your final project and why? Which do you explicitly NOT plan to use and why?

5. Applications the finding may have

Once you have completed your project, what are some of the applications of your findings? In other words, how might those findings be applied? What is the "practical" value of the model you will have built?

Making money in European betting systems?

Also, what will your deliverable be in addition to your code and data? Will you write a report in the style of CS229? Will you create a visualization? (NOTE: Do NOT attempt to learn D3 on top of everything else unless you are already a javascript ninja! Seriously.)

Since this is a (hopefully) straightforward classification problem, I'd like to try and visualize this data with D3 if I can, but no promises.

REFERENCES:

[Optional] links to relevant sources

Other potential questions to answer:

- **If a favored player is knocked out early, does that improve the chances of an "underdog" winning the entire tournament?**
- **Who will win the upcoming French Open? US Open? Wimbledon?**