Final Project Ideas

Maxwell Bates

DNA assembly prediction

When a researcher is ordering segments of DNA, can we predict what larger organism / gene someone is ordering?

(DNA fragments are ordered in small pieces)

Goal: predict what researcher is making (recommendations, more sales, security)

DNA assembly prediction

Source: publicly available genes (NCBI)

Fragments can be generated programmatically (segment of organism, random, two pieces joined, etc.)

Features generated by running fragments through BLAST algorithm, which performs a sequence comparison and calculates a match score

Bonus: BLAST DNA and protein sequence

DNA assembly prediction

Hypothesis: Using BLAST scores, if a set of fragments comprise a larger sequence, we can predict what that sequence is

Hypothesis: Can make better predictions than just looking at single fragments

Can determine if researcher is trying to order pieces for a particular sequence

Amazon Reviews -> Rating

Given text + metadata of a review, can we predict the star rating?

Using this model, can we predict the sentiment of reviews of different product categories?

May have implications in determining fake reviews, ensuring stars match sentiment (if model accurate enough)

Amazon Reviews -> Rating

Source: Dataset of 500,000+ reviews (general or food or movies)

Example High level information:

Number of reviews: 34,686,770
Number of users: 6,643,669
Number of products: 2,441,053
Users with > 50 reviews: 56,772
Median # of words per review: 82
Timespan: Jun 1995 - Mar 2013

Amazon Reviews -> Rating

Perform sentiment analysis and perform regression / naive Bayesian

Success = predict positive / negative sentiment based on text content

Wild success = accurate star predictions using additional metadata

Scientific Paper Citations

Predict number of citations of published scientific papers

citations is roughly correlated to impact of paper

Matters to scientists, journals, etc…. Publications are academic currency

Scientific Paper Citations

Source: published scientific papers (probably the ones with public access)

Basic sentiment analysis (this probably won't work alone, but may lead to other interesting findings)

Other possible features:

- Popularity of specific keywords
- # of years authors have published
- # authors

Scientific Paper Citations

Hypothesis: Content of the paper will allow us to predict its number of citations

Expect this to be very rough, so any statistically significant correlation would be cool

Spitballing

Weather vs. National health outbreaks vs stock market prices

Word counts over time vs. book’s success

Amazon food reviews, sentiment analysis… other source

Uber pickups

Climate change (landsat?) vs. housing prices - ML

Landsat vs. crop prices

Pollution data vs. ___ (storm prediction? cancer data? energy data? crop prices if regional?)