Ranking YouTube Videos' Popularity

Loc Do and Joseph Richardson

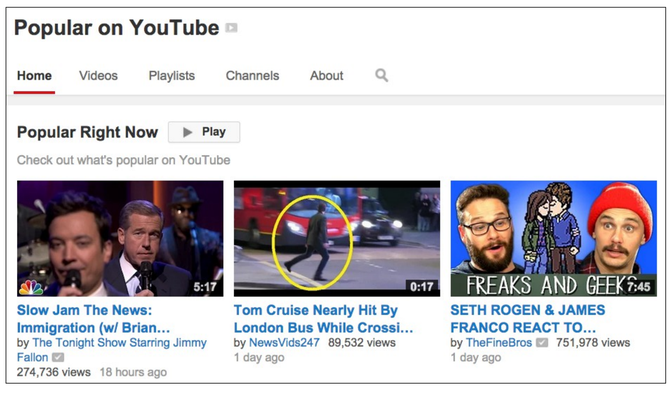
Carnegie Mellon University, Computer Science Department

PROBLEM STATEMENT

- Given two YouTube videos, can we predict which one will attract more views based on the metadata
- Challenges
  - Complex problem by nature
  - Sparseness in the training set
  - Over a million videos ➔ computational challenge

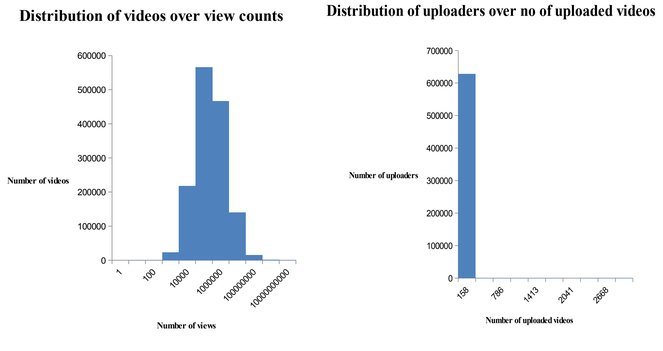
MOTIVATION

- Enhance the video recommendation system
  - Return a list of "popular" and "relevant" videos to users' interests.
- Figure out which features correlate most highly to the popularity of a YouTube video.



DATASET

- Data was crawled from YouTube in Oct-Nov, 2014
  - 1,432,213 videos with metadata (title, view counts, no of likes, no of dislikes, etc.)
  - 628,072 unique YouTube uploaders
- Distributions of videos and uploaders:



TWO METHODS

RANKING BY CLASSIFICATION

1. Problem Formulation

Given two videos i and j, each is associated with a feature vectors X_i and X_j. Let Y_{ij} denote a binary class indicating which video is more popular.

$$Y_{ij} = \begin{cases} 1, viewCount_i \geq viewCount_j \\ 0, otherwise \end{cases}$$

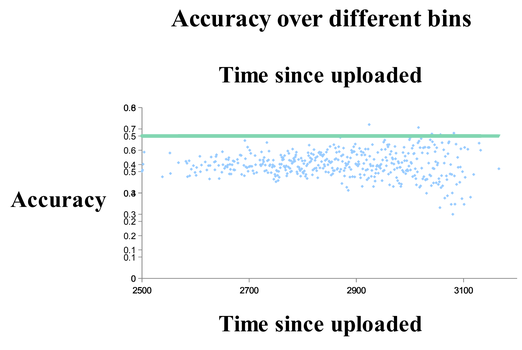
Finding the more popular video is equivalent to predicting the binary class.

2. Approach

- Demote X as representative feature vector $X_{ij} = X_i - X_j$
- Apply logistic regression on X_{ij} vs Y_{ij}
- Ill-conditioning optimization problem: Stochastic Gradient Descent with regularization.
- Large number of videos: Bagging methods by training different classifiers on different parts of the data and using a majority voting scheme on the test set.

3. Results

Accuracy on pairwise comparison in the test set varies with video age:



RANKING BY REGRESSION

1. Problem Formulation

Predict the view count of the videos one at a time and then compare, rather than comparing one pair at a time.

While this will likely perform slightly worse at the ranking problem, it allows us to make stand-alone to give predictions about the order of magnitude for a video's popularity.

$$f(X) = X\beta$$

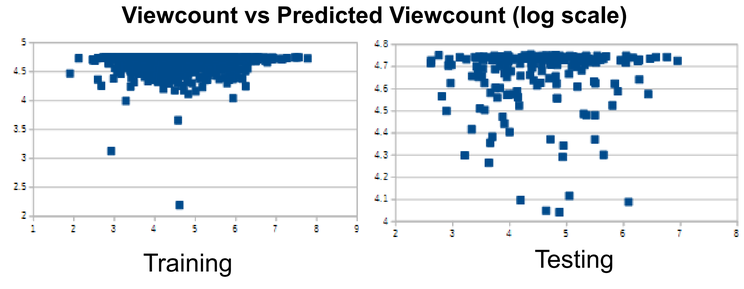
2. Approach

- Linear Regression
- Stochastic Gradient Descent
- Log scale
  - Order of magnitude matter most
  - Avoid dominance of most popular

$$\beta^{t+1} = \beta^t - \eta(x_i(x_i\beta^t - Y_i) + \beta^t)$$

3. Results

- Typically correct only within an order of magnitude
- High error in both training and testing suggests that this problem is not linearly separable
- Ranking accuracy just over 50%
- More sophisticated models needed



FEATURE EXTRACTION

Features are extracted from video's metadata, including:

- Video features
  - Bag-of-words model on the title (2,447,603 unique words)
  - Video length [1 second, 107373 second] (in seconds)
  - Days since first uploaded: [1, 3423]
  - Ratio of Likes/Dislikes
    - Like [1, 8M]
    - Dislike [1, 4M]
- Uploader features
  - Subscriber count
  - Number of videos previously uploaded by uploader

FUTURE WORK AND STRETCH GOALS

- Bag-of-features are not good indicators of ranking videos popularity
  - Find more features such as mining videos' comments.
- Both linear models show bad results on separating the two classes
  - Apply non-linear models such as Gaussian Processes
- Predicting the popularity of videos over time
  - Capturing different snapshots of a video over a period of time
- Examine the effect "the rich get richer" by YouTube recommendation system on the video popularity.