**Youtube Trending Videos Prediction and Latent Semantic Analysis**

**Final Project Report Draft for CEE 690-07, Fall 2019**

**Jiajie Zhang**

## 1. Introduction

Youtube is the most popular video-sharing platform in the USA and even around the globe. Over the years, Youtube created an economy eco-system where people take channel uploaders as an occupation and get pay-back based upon the popularity of their videos. It’s rather important for the uploaders to consider approaches to make the videos that are potentially popular or going to go viral, since their earnings are highly dependent on how much user activities they created on Youtube platform. The good exemplars for those uploaders are the year’s top trending videos. According to Variety magazine, “To determine the year’s top-trending videos, YouTube uses a combination of factors including measuring users’ interactions (number of views, shares, comments and likes). Note that they’re not the most-viewed videos overall for the calendar year”. Therefore, an analysis on Youtube top trending list may cast light on which category and what kind of topics might be useful to help uploaders determine the topics of

* What factors can be considered to predict the total views of a trending videos?
* Will the difference in categories and topics affect the popularity of a Youtube video? i.e. Would the topics and categories of the videos be a significant factor that affects the video popularity?

To answer these questions, we construct a Bayesian regression model combined with the latent semantic analysis on a list of Youtube trending videos with several descriptive and evaluative features, which we assume there exists some relationship between features and the total view counts of each video.

## 2. Data

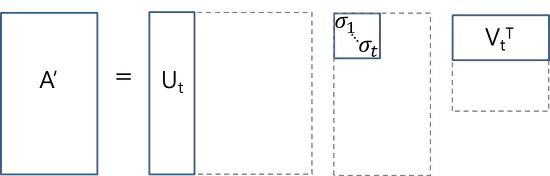
The data was retrieved from Kaggle open Datasets named “Trending YouTube Video Statistics”, which includes data on daily trending YouTube videos from 2017/11/14 to 2018/6/14. The whole dataset in the US region was used in this study. The columns of each observation include descriptive contents such as video id, video title, trending date, channel title, publish time, description, category, as well as numerical values describing the performance of the video, such as tags, views, likes and dislikes, and comment counts. While the video id in raw dataset is not unique, since some of the videos are trending for a consecutive amount of days, we therefore create a new parameter ‘Trending\_days’ that counts the time of occurrence of a video showed up on the list. The row that has the latest trending date was preserved for any video that has multiple trending\_date. All the numerical values were standardized before putting into the model training.

Different data manipulation methods were adopted in different models. In non-Bayesian models, only numerical features were included for the prediction of video total views. In the early data assessment, we calculated the group means of numerical features, and significant discrepancies of those numerical means were observed as in Figure 1. We therefore merged grouped mean of those numerical values in non-categorical models to represent this difference.

Additionally, observations with comments or ratings disabled were removed from dataset. We then engineered a feature of like/dislike ratio which is expected to be another important factor evaluating the performance of a video. Some infinity numbers in this column were replaced by max like/dislike ratio in the dataset to avoid calculating failure.

To explore the effects of categories in Bayesian modeling, we convert video category column into a dummy variable matrix with the first columns dropped out. The reason for using N-1 dummy variable is to avoid each column being the linear combination of other columns, so that the coefficient of each column can be calculated and is relative to the column we dropped out.

In order to analyze the relationship between video tags and total views, the latent semantic analysis (LSA), (also called latent semantic indexing in the NLP field) on video tags was adopted. The key idea of LSA is to map high-dimensional count vectors (McGill, 1983.), to a lower-dimensional representation in latent semantic space (Hofmann, 1999). The implementation of LSA generally contains two steps. We first apply term frequency-inverse document frequency (tf-idf) transformation on the tag column, reflecting the importance of the words in the document. This approach yields a document-term-matrix as well as a dictionary containing all the words in input tags. Then an SVD decomposition was applied in order to reduce the dimensionality of the sparse doc-term-matrix. We preserved 15 dimensions out of the intuition that the number of categories in the dataset was 15 originally. Each dimension was described by a linear combination of 10 words with weights assigned to them respectively. The mathematical representation of this approach is as follows (Brand, 2006):



Where that U and V are orthogonal matrices and is a diagonal matrix. Whereas t here denotes the number of dimensions (topics in this study) we’d like to reduce to. Similar to Principle Component Analysis, LSA finds a low-rank approximation, which means the first few dimensions often explain most of the variance in the original matrix A. With the help of decomposition, we reduced the dimensionality of A to representing the document-topic-matrix, with a covariance matrix representing the covariance between each topic, and representing the term-topic matrix. (Brand, 2006).

## 3. Methods

## 3.1 Support Vector Machine

We selected the support vector machine (SVM) regressor with radial basis function(rbf) kernel as the non-Bayesian model in comparison with Bayesian methods. SVM is a robust machine learning algorithm in that it maximizes the minimum margin. With radial basis function kernel, SVR is able to capture the nonlinearity in some features, while the interpretability is weakened. The tuning parameters ‘C’ and ‘epsilon’ are optimized under 3×3 grid search for the highest R-square score.

In order to compare the coefficient of parameters calculated between Bayesian and non-Bayesian methods, a linear kernel with optimized tuning parameters was then applied for fitting. This approach may sacrifice some of the model accuracies, while parameter coefficients were preserved so that it adds up model interpretability.

To explore the necessity of categorizing data and LSA on tags, input dataset with only numerical parameters, with numerical and dummy variables and with all of the numerical, dummy and LSA topics probability parameters are evaluated and compared by R-squared score.

## 3.2 Bayesian Regression Models with Latent Semantic Analysis and Dummy Variables

After the LSA process, each row of tags was represented by a linear combination of 15 weighted topics (dimensions). For the limitation of computation power, the number of dimensions was not tuned for best model performance, but empirically determined by comparing coherence values for a grid of dimensions (Figure 2). The probabilistic model of the observed scalar data is approximated by the linear combination of all parameters and intercept, with a normally distributed error term, as shown in the equation below:

Where denotes each observation I the dataset, denotes the total number of input parameters, which is the sum of number of numerical values, dummy variable and topics. The prior of the intercept is a normal distribution with standard deviation =100, i,e., ); The prior of s are normal distributions with the means and . The prior of variance is also assigned to an Inverse Gamma distribution, i.e., , (Figure 3). The advantage of this model is that though it contains many aspects of parameters, the model structure remains to be relatively straight forward, therefore it linearly adds up the computation complexity which is effective. However, it’s unavoidable that many hyperparameters from LSA methods need to be tuned when searching for best model performance. Additionally, it is unknown that whether including both video categories and tags topics would increase the performance of prediction, since these two parameters in the raw dataset may contain similar information, which can both be treated as clustering of videos based on the contents.

We utilize No-U-Turn (Hoffman, 2014) sampling method in pymc3 (Salvatier J, 2016) package because of its efficiency. The target acceptance was set to 0.9 since initially some of the parameters have diverged. Other sample hyperparameters are set as: draws = 200, tune =1500, chains = 2. A single run of the Bayesian regression model roughly takes 10 minutes. This enables us to tune hyperparameters for the model in future.

## 4. Results

The R-squared score is compared across both Bayesian and SVM models, the WAIC score is compared within Bayesian models.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Parameters |  |  |
| SVR-rbf-1 | Numerical and Dummy | 0.78 | - |
| SVR-rbf-2 | Numerical and LSA | 0.71 | - |
| SVR-rbf-3 | Numerical, Dummy, LSA | 0.78 | - |
| Bayesian-1 | Numerical | 0.75 | 35490±2375 |
| Bayesian-2 | Numerical and Dummy | 0.75 | 35084±2388 |
| Bayesian-3 | Numerical and LSA | 0.75 | 35055±2438 |
| Bayesian-4 | Numerical, Dummy, LSA | 0.75 | 34763±2439 |

In comparison of SVR model with different parameter inputs, we observed a higher prediction performance when incorporating categories, while the effects of LSA variables are not substantial, given that the of SVR-rbf-1 and SVR-rbf-3 is very similar. More importantly, the performance of SVR-rbf-2 model is significantly weaker than other models, indicating either: (1) the tags of each is a weak predictor to video views, or (2) The process of LSA needs to be further refined. As is shown in the topic weight matrix (Table 1), some terms recursively showed up in many topics such as ‘news’, ‘song’, ‘meme’. While some of topics contain compact terms with similar semantics, for example, the majority of high weighted words in topics 2 is about politics and news, and most high-weight words in topics 3 are about sports.

Surprisingly, the scores of multiple Bayesian models with different input are very alike. With the increasing of parameter numbers, the WAIC score doesn’t seem to reduce significantly. This indicates that no substantial information was added to the model along with the introduction of those parameters. The performance results of Bayesian regression models are somehow contradicted to SVR models.

The coefficients of the parameters generated by both Bayesian and SVR method were compared in Table 2. Although the relative differences are high for many of the parameters, the rank of weights associate to variables in two models are similar. This may be due to the different strictness of regularizations between two models, for example many of the coefficients in Bayesian model is greater than SVR model in absolute value.

Further, we filtered statistically significance parameters in Bayesian models. It turns out that 21 out of 36 parameters were statistically significant in that 0 is not among the 95% confidence intervals. Additionally, only 4 of 14 parameters in topics are statistically significant.

## 5. Discussion

The model we build up above demonstrates that the total views of a video can be predicted by numerical values in the input data set, such as likes dislikes, trending dates. the SVM model shows the categorical discrepancies among the videos. However, none of the SVM model and Bayesian model implies the substantial relationship between latent topics and video views. With regard to the model comparison between SVM and Bayesian, SVM yields better performance in terms of the model accuracy where more parameters are included, while Bayesian regression appears to less affected by the increase of parameters, and is slightly underperformed than SVM. Due to the high interpretability of Bayesian models, more information, such as coefficients and confidence levels associated with each parameter can be collected and analyzed.

One shortcoming for this study is that the performance scores of Bayesian models too similar that comparing them it's unapproachable, one potential solution for that is to introduce other matrices, for example, the MSE scores.

The models are preliminary and are yet to be powerful to draw persuasive conclusions. There are many aspects that can improve. For example, as mentioned before, during the LSA text cleaning process, the number of topics and words are arbitrarily assigned. It is reasonable to conduct great search on these parameters. Additionally, the latent Dirichlet allocation (LDA) topic modeling algorithm, which is similar to the LSA algorithm, but applies Bayesian inference, can be a competitive alternative to the LSA algorithm. When applying that more visualization approaches to the topics can be implemented, and thus provides more insight on the sparse distribution of topics.

## 6. Figures

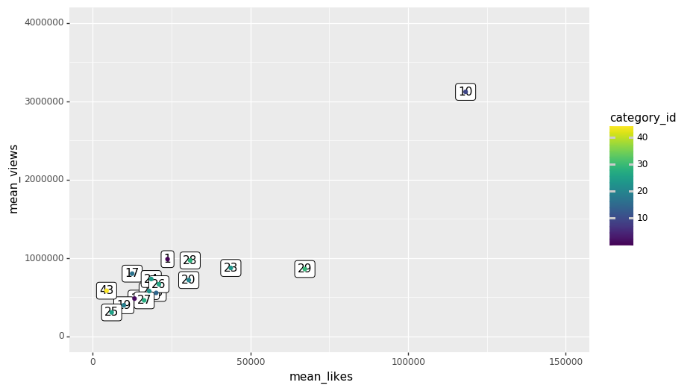


Figure 1: Mean of views and likes of a video grouped by categories

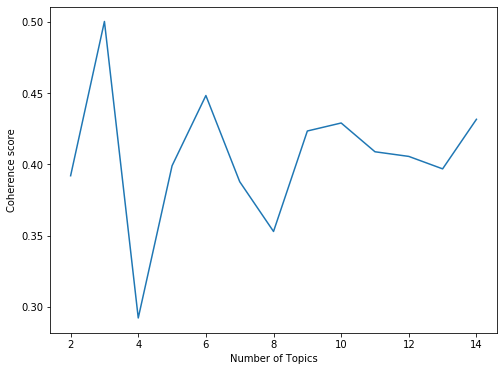


Figure 2: Coherence score vs. number of topics

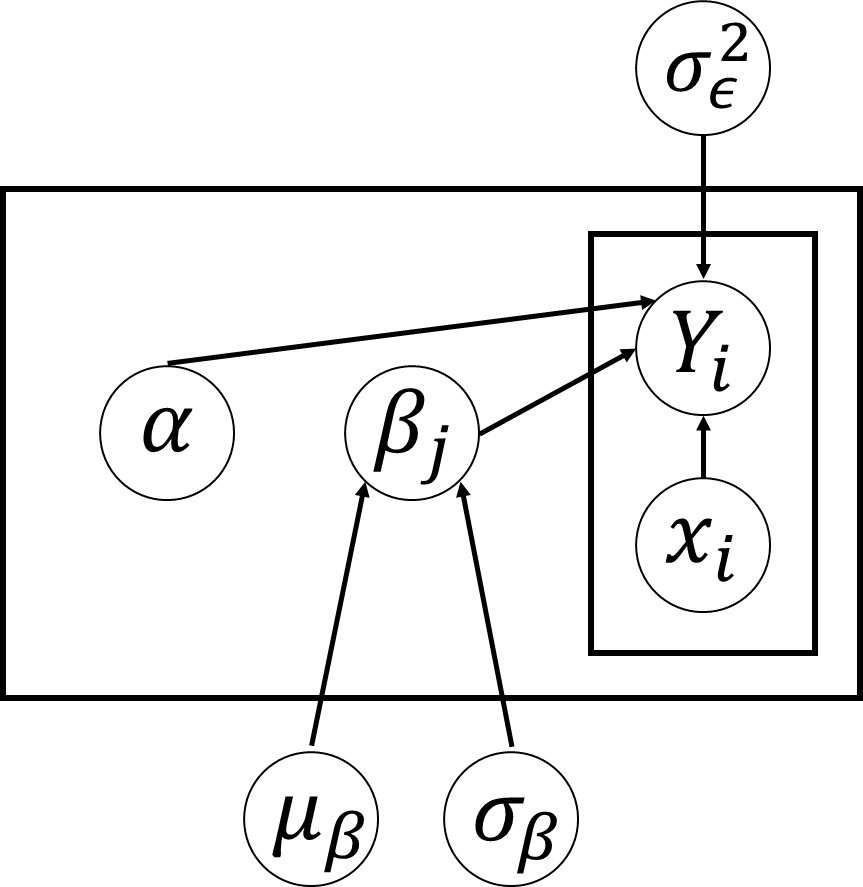


Figure 3: Graphical interpretation of Bayesian regression model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Topic | 0 | 1 | 2 | 3 | 4 | 5 |
| term1 | song | meme | news | vs | crime | food |
| 0.575 | -0.770 | -0.648 | 0.472 | 0.429 | 0.682 |
| term2 | punjabi | dank | trump | highlight | tv | street |
| 0.312 | -0.351 | -0.388 | 0.416 | 0.286 | 0.235 |
| term3 | 2018 | song | song | game | episod | link |
| 0.293 | 0.264 | 0.358 | 0.387 | 0.273 | 0.201 |
| term4 | new | compil | punjabi | news | serial | rhett |
| 0.264 | -0.240 | 0.199 | -0.264 | 0.262 | 0.199 |
| term5 | meme | funni | live | nba | song | mythic |
| 0.258 | -0.171 | -0.175 | 0.213 | -0.262 | 0.156 |
| term6 | latest | punjabi | show | cavali | drama | video |
| 0.200 | 0.147 | -0.142 | 0.178 | 0.206 | 0.149 |
| term7 | news | comment | donald | song | news | 2018 |
| 0.186 | -0.117 | -0.137 | -0.173 | -0.200 | -0.131 |
| term8 | 2017 | award | polit | trump | comedi | react |
| 0.124 | -0.116 | -0.108 | -0.155 | 0.192 | 0.127 |
| term9 | dank | 2018 | msnbc | 2018 | soni | tulfo |
| 0.116 | 0.102 | -0.091 | 0.136 | 0.163 | -0.123 |
| term10 | funni | latest | cnn | warrior | show | good |
| 0.116 | 0.093 | -0.089 | 0.128 | 0.162 | 0.112 |

Table 1. Topic-term matrix of LSA

|  |  |  |  |
| --- | --- | --- | --- |
| Parameters | Weights(SVR) | Weights(Bayesian) | Abs\_Relative\_Diff |
| likes | 0.5891 | 0.8113 | 27% |
| dislikes | 0.3258 | 0.2947 | 11% |
| trending\_days | 0.0678 | 0.1414 | 52% |
| Movies | 1.1119 | 0.0920 | 1109% |
| topic\_0 | 0.0311 | 0.0616 | 50% |
| Sport | -0.0014 | 0.0301 | 105% |
| Shows | 0.0841 | 0.0232 | 263% |
| topic\_3 | 0.0235 | 0.0225 | 4% |
| topic\_5 | 0.0069 | 0.0043 | 60% |
| topic\_2 | -0.0052 | 0.0006 | 920% |
| topic\_1 | 0.0006 | -0.0042 | 113% |
| topic\_4 | -0.0163 | -0.0188 | 13% |
| Cars and Vehicles | 0.0024 | -0.0192 | 113% |
| How to and Style | -0.0232 | -0.0214 | 9% |
| like\_dislike\_ratio | -0.0103 | -0.0231 | 56% |
| Entertainment | -0.0010 | -0.0266 | 96% |
| Travel and Events | -0.0378 | -0.0296 | 28% |
| Science Technology | -0.0146 | -0.0437 | 67% |
| News and Politics | -0.0235 | -0.0489 | 52% |
| Pets and Animals | -0.0375 | -0.0745 | 50% |
| People and Blogs | -0.0246 | -0.0750 | 67% |
| Education | -0.0492 | -0.0923 | 47% |
| Music | -0.0350 | -0.1102 | 68% |
| Gaming | -0.0673 | -0.1283 | 48% |
| Comedy | -0.0926 | -0.2094 | 56% |
| comment\_count | -0.0623 | -0.2126 | 71% |
| Non-Profit | -0.1319 | -0.4541 | 71% |

Table 2. Parameter coefficients of SVR and Bayesian model, with Absolute relative difference

# References

Brand, M. (2006). Fast Low-Rank Modifications of the Thin Singular Value Decomposition. *Linear Algebra and its Applications, 415*, 20–30. Retrieved 12 9, 2019, from http://www.merl.com/reports/docs/TR2006-059.pdf

Hoffman, M. D. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1), 1593-1623.

Hofmann, T. (1999). *Probabilistic latent semantic analysis*. Retrieved 12 9, 2019, from https://dl.acm.org/citation.cfm?id=2073829

McGill, G. S. (1983.). Introduction to Modern Information Retrieval.

Salvatier J, W. T. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science 2:e55 https://doi.org/10.7717/peerj-cs.55*.

Smola, A. J., & Schölkopf, B. ( 2004). A tutorial on support vector regression. *Statistics and Computing*, 14 (3): 199–222.

Mitchell Jolly, “Trending YouTube Video Statistics”. 2017.11.13. Kaggle. version 115. https://www.kaggle.com/datasnaek/youtube-new/metadata

Project Summary

The purpose of this project is to retrieve useful information from Youtube trending videos dataset that provide guidance to uploaders on video preparation such as topics and category selection. To achieve this, we focus on solving these two main problems in the project:

* What factors can be considered to predict the total views of a trending videos?
* Will the difference in categories and topics affect the popularity of a Youtube video? i.e. Would the topics and categories of the videos be a significant factor that affects the video popularity?

To answer these questions, we built up a series of statistical and machine learning models to evaluate the effects of various video attributes, such as likes, dislikes, category and topics to the performance of a videos, which is represented by total views. A text analysis on video tags is also conducted in this research. The result shows that videos in the categories of movies, sport, shows are more favorable to the users on Youtube, while those in non-profit, gaming and comedy categories are relatively less favorable. Topic patterns like sports and news are detected from the analysis of tags, among which topics regarding latest news, entertaining or funny videos as well as memes are popular on Youtube. Therefore, we conclude that videos with contents about movies sports, entertaining shows and up-to-date news are recommended for Youtube Channelers.