Report - Kaggle Competition

ABBA

February 8, 2016

Introduction

Additional variables

The url of the articles include useful information that can potentially imporve our prerformance in the competition. An example of the url is the following:

http://mashable.com/2014/01/12/game-of-thrones-season-4-trailer

From the url we can get the year, day and month then it was published as well as some keywords of the content of the article.

Using text mining techniques we can get all the keywords of the article titles. With this we can create new variables indicating whether a specific word appears in the title of the article. Given that we have many keywords, we create variables only for those keywords that have appeared in at least 150 observations.

Evaluation

In order to evaluate our models, we decided to use 5-fold Cross Validation

explain 5-fold CV

Variable Selection: RFF

Models

- Random Forest
- K-nearest neighbors
- SVM
- Boosting

Results

(plot ?) Kaggle score vs CV score