CS 725

PROGRAMMING ASSIGNMENT 1

AUTUMN 2016, IITB

Problem statement:

With the expansion of the Internet, more and more people enjoy reading and sharing online news articles. The number of shares under a news article indicates how popular the news is. In this assignment, you need to find the best model and set of feature to **predict the popularity of online news ['shares']**, using machine learning techniques.

The aim of this assignment is to help you learn the application of machine learning algorithms to data sets. This involves learning what data means, how to handle data, training, cross validation, prediction, testing your model, etc. We will try to do the complete flow in this assignment.

Data set:

The training data set is provided in the file "train_data.csv". A brief description about the features is given as below.

Attribute Information:
0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)

23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

You will be provided with 2 files train_data.csv and test_data.csv. Your code must take both as input and output predictions for the test data. Formats for deliverables is described below.

Files to be submitted on BodhiTree: A single zip file named ROLLNUMBER.zip, containing below files

- Code files [model.x]
    - One main file should be run to utilize all the code.
    - The main file should be named model.x [x=py,r].
    - Only standard libraries like numpy, pandas, sklearn [in case of python] are encouraged.

- Read me file [readme.txt]
    - Should contain description of your code, model and feature engineering.
    - One line command to run your whole code.
- output file against test data [output.csv]

Files to be submitted on Kaggle :

- **output.csv [to be uploaded]**
  Output format:
  output.csv with two columns containing a header row ["id","shares"] followed by all the predicted shares for each entry in the **test_data.csv**. Order of output entries should be same as the order of test data.

id, shares

0,1000

1,234

2,529

…

Programing language:

R/Python will be allowed to maintain uniformity. You are allowed to use the already available libraries.

Important Note:

Your code should use the dataset provided by us for the training. If that is not done, you will be awarded zero marks for the assignment. **Copying of assignments will be dealt with very seriously**.

Post your queries on BodhiTree. Working on this assignment and figuring about the intricacies of R/Python will help you a lot during the project.

Important References:

- http://scikit-learn.org/stable/
- http://scs.math.yorku.ca/index.php/R:_Getting_started_with_R
- **Good place to start** : https://www.kaggle.com/c/titanic