SMAI Project - Eval II

Predicting the Popularity of Online News

Dipankar Niranjan
Sriharsh Bhyravajjula
Susobhan Ghosh

Feature Extraction

We have used **Fisher score** to extract top 20 features out of 60 given in the dataset. This is done to determine features with the least correlation.

Our results were similar to the one stated in the paper, and these were the **top 20** features:

'kw_avg_avg', 'LDA_02', 'data_channel_is_world', 'is_weekend', 'data_channel_is_socmed', 'weekday_is_saturday', 'LDA_04', 'data_channel_is_entertainment', 'data_channel_is_tech', 'kw_max_avg', 'weekday_is_sunday', 'LDA_00', 'num_hrefs', 'global_subjectivity', 'kw_min_avg', 'global_sentiment_polarity', 'rate_negative_words', 'num_keywords', 'num_imgs', 'LDA_01'

Machine Learning Approaches

As planned before, we have started to implement a basic linear descent method - Widrow Hoff (LMS).

Widrow-Hoff algorithm is a Least Mean Square (LMS) algorithm. It is similar to the relaxation rule above, but the key difference is that the relaxation rule is a correction rule, so $a^T y_k$!=b, and the corrections never cease. In case of Widrow-Hoff, the learning rate eta, is annealed i.e. decreased wrt k to converge. Effectively, $\eta(k) = \eta(1)/k$ is the learning rate. It terminates whenever a k is obtained such that

$\eta(k)(b_k - a^T y_k)y_k < \theta$

Results of Widrow Hoff

We initially faced issues with convergence of LMS using the training data (70-30 cross validation).

After setting very low values of theta and eta - as 0.000000001 and 0.0000001, we were able to get an accuracy of about 49.5%

After tweaks and improvements to eta and theta, the accuracy stood out to be 53.384%