Stack Overflow Query Outcome Prediction

Robbie M. Jones[1], David Lin[2]
[1]Department of Computer Science, Stanford University,
[2]Department of Physics, Stanford University

Motivation

With over 13 million questions and 6 million users, Stack Overflow is an invaluable resource for computer scientists. As the community has grown, the ratio of answers to questions has decreased by over 80 percent. Therefore, **community moderators prevent content dilution by closing questions that are off topic, not constructive, not a real question, or too localized.**

In order to help new users get acclimated to the site and reduce moderator burden by automating quality control, we **build a classifier to predict whether a user's question will be closed along with the reason for closure.**

Data & Preprocessing

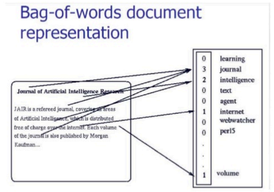
We used the StackExchange Data Explorer to query publicly available Stack Overflow data. Preprocessing included separating closed and non-closed questions, restricting time-sensitive attributes, and applying various transformations to textual data. Our resulting matrix had a question and associated data for each row, labeled as "not-closed" or "closed-[reason]".

Feature Extraction

User Information
- Reputation
- Age of user
- Account age Previous

Content
- Bag-of-words for title
- Bag-of-words for body

Text Metadata
- Number of words, sentences, special characters, StackOverflow URL's, tags, punctuation marks, lines of code, lower case letters, upper case letters



Models

Classification Summary

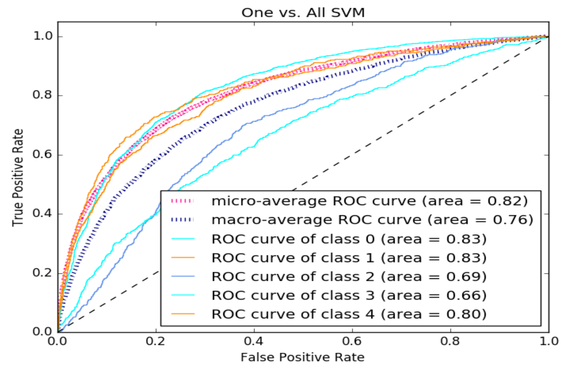
Model	ACC
Lasso Logistic (L1)	.607
Ridge Logistic (L2)	.610
Adaboost	.552
Linear SVM	.763
Polynomial SVM	.618
Radial SVM	.748

Our baseline used L1 and L2 logistic regression, obtaining ACC values of .607 and .610. Our boosting algorithm achieved 0 training error. However, the boosting model overfit the training data and underperformed the baseline on the test set. We can improve boosting results with feature reduction

L2 Baseline Confusion Matrix

		Prediction				
		Class 0	Class 1	Class 2	Class 3	Class 4
Actual	Class 0	2067	54	204	80	60
	Class 1	96	384	28	75	82
	Class 2	363	28	148	45	16
	Class 3	217	83	98	161	99
	Class 4	80	108	37	94	293

Results & Discussion



Our best learner was the linear SVM, followed immediately by the radial SVM. The microaverage and macroaverage ROC curves are shown above. Ultimately, Our classifiers did well separating non-closed questions from closed questions, but differentiating between the different reasons for closure was much more difficult.

Future Work

1. **New Features.** In particular,
   a. Community Scores: Badges, post scores, favorites
   b. Temporal Aspect: Different time periods would account for changing community culture
2. **Neural Networks.** With a good topology, NN can better capture non-linear decision boundaries.

References

[1] D. Correa and A. Sureka, "Fit or Unfit : Analysis and Prediction of 'Closed Questions' on Stack Overflow," in Proceedings of ACM, 2013. [Online]. Available: https://arxiv.org/pdf/1307.7291v1.pdf. Accessed: Dec. 13, 2016.
[2] "Predict closed questions on stack overflow," in Kaggle, 2013. [Online]. Available: https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow. Accessed: Dec. 13, 2016.
[3] L. Ponzanelli, A. Mocci, A. Bacchelli, M. Lanza, and D. Fullerton, "Improving Low Quality Stack Overflow Post Detection," in University of Lugano, 2014. [Online]. Available: http://www.inf.usi.ch/phd/ponzanelli/profile/publications/2014e/Ponz2014e.pdf. Accessed: Dec. 13, 2016.