Methodology

Stack Overflow Question Quality

Ryan Tobin, James Westbrook

In our attempt to classify Stack Overflow questions as “good” or “bad” based on their score, we will use a variety of machine learning techniques. We will use logistic regression, (extreme) gradient boosted trees, and SVM. We will use k-fold cross-validation for each method to reduce variance imposed by our relatively small sample size. We expect gradient boosting and SVM to achieve a better classification error than logistic regression, however as we are interested in inference, we include it as it is far more interpretable than the other two. There is much potential for improvement, particularly with how we select our features. We expect that interaction features may play a more important role than raw features - for example, code to text ratio may be a more important metric than simply the number of words in a post - and it can be very difficult to identify these interaction features.

To evaluate each model, we will use test classification error, and will consider 20% classification error to be successful. Other researchers were able to obtain an accuracy of 80% on similar questions, and obtaining above 80% accuracy would be considered impressive. Along with this, we will compare areas under the ROC curve for each method.

To make sure that we are obtaining legitimate results, we will perform a 70/30% training/test set split for each model. This split will be random for each model so that we can avoid overfitting in our analysis. If we use the same test set for each model, then this may not give a good picture of how each model performs in general, especially with a small test set.