PS6 Machine Learning: Everyone's a Critic

Cai Glencross and Katie Li

March 4, 2018

1 Hyperparameter Selection for a Linear-Kernel SVM

(d) At a higher C, we punish each misclassification more, and are more prone to over-fitting with less generalizable results. Almost all the metrics, accuracy, f1score, auroc, precision, and sensitivity, generally have a higher score at a larger C. Likely the held out validation set and training sets are quite similar, and thus minimizing the total number of misclassification allows for high scores of these metrics.

The only metric whose score decreases with a higher C is sensitivity. A small C maximizes the margin. At an extremely small C, there is a very large margin, and the classifier will resemble the majority-vote classifier. Thus, if we classify everything as positive, then we would have no false negatives and therefore perfect sensitivity, yet poor specificity. This result is shown in the graph, where when we have $C = 0.001$, then specificity is approximately 0 and sensitivity is 1.0. This implies that most of the tweets in the training and test data are positive. (See Figure 1).

The optimum C appears to be 1.0.

2 Hyperparameter Selection for an RBF-kernel SVM

(a) A higher γ determines the shape of the decision boundary, whether it is a more extreme transformation, or a less "smooth" decision boundary.

While a large γ may lead to overfitting, and have higher generalization error, a small γ may lead to underfitting with lower generalization error.

(b) For the grid, we used a logarithmic scale for the C_{values}, similar to the example given in the RBF linear model, in powers of 10 from 10^-2 to 10^2. For the γ, we used the grid in power of 10 from 10^-3 to 10^2. We used logarithmic scales to test on both of these parameters to get a better spread of data. Please find the results in Table 1.

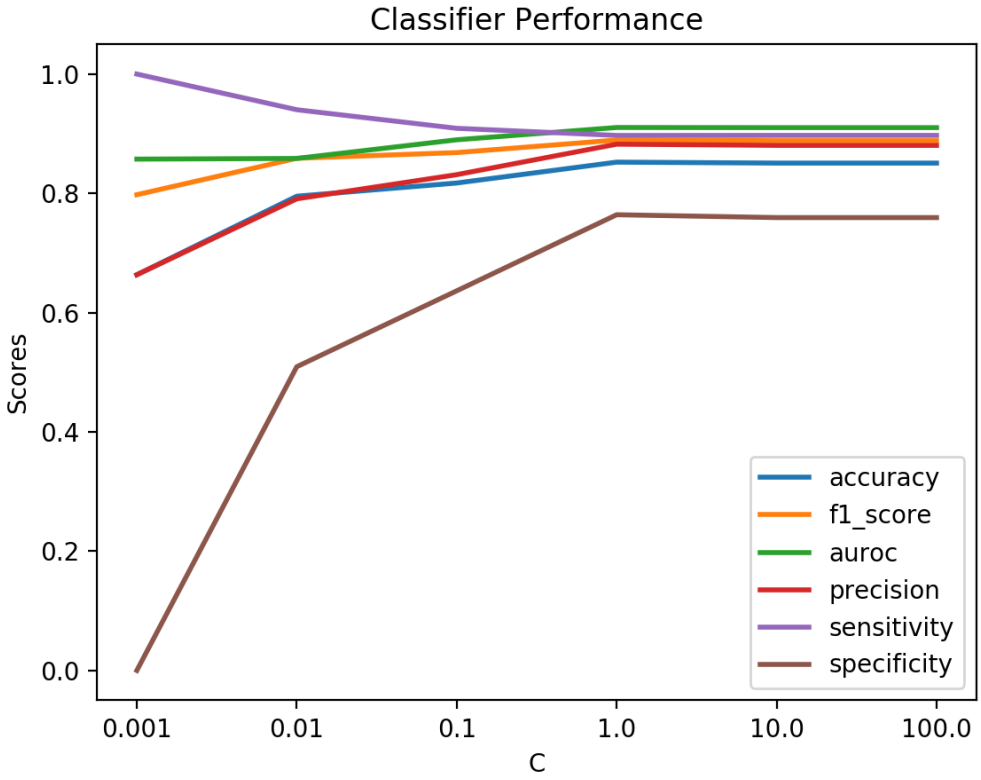

Figure 1: Performance metrics for linear SVM (Problem 1 d)

(c)

metric	score	C	γ
accuracy	0.829	10	0.010
F1-score	0.873	10	0.010
AUROC	0.906	10	0.010
precision	0.854	10	0.010
sensitivity	1.00	0.01	0.001
specificity	0.678	10	0.010

Table 1: Performance Metrics for the RBF_kernel SVM

With the exception of sensitivity and specificity, all the parameters have relatively high scores with $C = 10$ and γ of 0.010. However, the cost for sensitivity is much lower, and the score is 1. A low cost of mis-classification would allow for more slack, and greater regularization. Combined with the evidence of the relatively low score for specificity, 0.678, where there are a high number of false negatives. Potentially the a highly regularized model that is nearly the majority class classifier, would give us something like these results.

3 Test Set Performance – Bootstrap Confidence Intervals

(c) In general the RBF SVM performed slightly better than the linear SVM - which implies that our linear SVM was underfitting the data, because the RBF SVM is a more complex model with a larger hypothesis space. By far the metric in which our model performed the worst was in specificity. This is because the majority of the tweets were labeled as positive. Therefore it is much more difficult to correctly label the negative data. For instance, if we were to just use the majority classifier, while our accuracy would be 73% our specificity would only be 0. We can observe the same bias in our precision metric, which shows how often a positive prediction is correct. These SVM models' precision metric are especially high since false positives are so rare, because there are so few negatively labeled tweets. The most illustrative metric might be the f1-score as it is the harmonic mean between sensitivity and specificity, and therefore would be less susceptible to imbalanced data.

4 Feature Importances

(a) The most important words for distinguishing the positive or negative sentiment are those with the highest coefficients. We take the highest coefficients, which are most influential in our linear-kernel SVM. These map back onto certain words, whose presence will influence a positive or negative.

(b) List of top 10 most important words in descending order from most important to least important.

   (a) Positive: 1. fanfreakingtastic; 2. cool; 3. funny; 4. d; 5. must; 6. loved; 7. awesome; 8. great; 9. good; 10. excellent

   (b) Negative: 1. hilariooooouuss; 2. isn; 3. boring; 4. 2012; 5. skip; 6. absurdly ; 7. disappointing; 8. shit; 9. hated; 10. not

   The features most important for positive sentiment are words that have the connotation of "good," which the features which are most important for negative sentiment mainly have connotation of "bad," which fits the general way that people talk in person about movies. However, there are

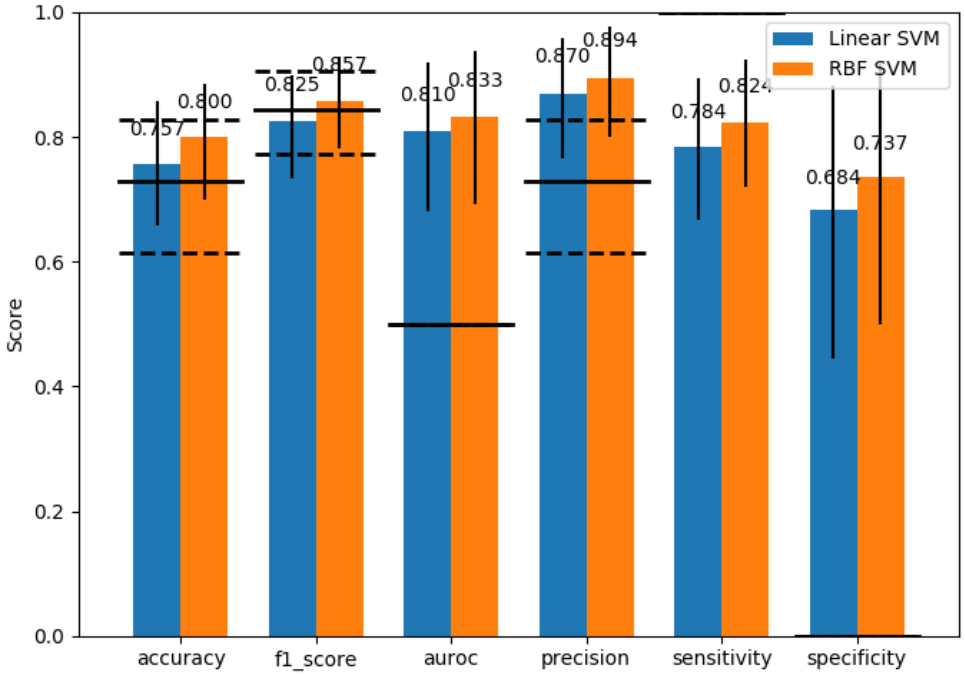

Figure 2: Problem 5c: Performance measures of baseline (majority vote), linear SVM and RBF SVM.

a few features were surprising, such as "2012," and "d." This might due to the fact that the training data is a bag of words, and doesn't have much context surrounding these words.

(c) The words are not placed in context but we can't tell which words surround them. For example, "not bad" is the opposite of "bad," but these SVM classifiers would label both these tweets with the feature "bad."

5 Explanation

In general, people who talk about movies on twitter are happy with what they see. Most people who wrote tweets about movies wrote positive ones. We found that a the presence of a several given words can be highly predictive of whether a tweet was positive. But, tweets about movies that specifically mention "2012", the word "boring," are likely from viewers who did not enjoy the movie. In addition, sarcastic tweets are likely to be negative reviews of the tweets, such as one that uses the word "hilariooooouuss." The tweets with the words "fanfreakingtastic", "cool," "funny," or other similar words with positive connotations are more likely to be positive reviews.