Homework 3

Daniel Hartig

March 28, 2017

a.

Since SVM uses multi-dimensional distance between points to assess similarity, it is important that each dimension be scaled similarly to other dimensions. If one feature has a much larger average magnitude than others, small relative differences in that single dimension can overwhelm larger relative differences in other features. To avoid this situation, all features should be re-scaled so that accurate distances between support vectors can be calculated.

c.

As C is increased, we can see that the number of support vectors decreases. Because C is the weight of the penalty on slack variables, as C increases, the model attempts to reduce the number of slack variables by decreasing the margin. So with a high C value, there is a smaller margin and fewer support vectors.

When C, the penalty for misclassification in the training set, is increased, the model's complexity (or its n-dimensional shape) increases so that every point (if possible) is classified correctly. As the complexity of the model increases, the computational complexity of determining distance to margin also increases. This is what causes increased runtime. At a high enough C value, the model will not terminate.

d.

The value for the entries in w^* is a measure of its relationship with either classification. Positive values indicate a stronger relationship with the positive classification, 1 (spam), while negative values indicate a stronger relationship with the negative classification, -1 (not spam).

The five features most strongly associated with the positive, 'spam' classification are 'capital_run_length_average', 'free', 'your', 'capital_run_length_longest', and 'you'. The five features most strongly associated with the negative, 'not spam' classification are 'edu', 'george', 'hpl', 'hp', and '650.'

Some of the most spam related features are intuitive; the word 'free' and long runs of capital letters are obvious indicators of spam. However, some of the features can potentially cause classification problems, such as the words 'your' and 'you.' These common words are likely less common in academic e-mails, but more likely to produce false positive classifications if applied to personal e-mail.

The least spam related terms are more difficult to interpret without knowing more about the source of the data. However, specific technical terms such as 'hpl' (probably referring to either a distributed linear equation solver or polypeptide hormone) or the 'edu' associated with academic institutions are good indicators that an e-mail is not trying to sell you something.

e.

The false positive and true positives rates increase as C increases. When $C < 1$, both rates are zero; that is, the soft margin is very wide and nothing is being classified as spam. The rates increase until at $C = 10^{4.5}$, 16.5% of the test set is correctly identified as spam (true positives), while 3.7% is incorrectly identified as spam (false postives).

While the true postive rate steadily increased with C, the false positive rate increased until $C = 10^{1.5}$ and then stayed relatively constant. To hold the false postiive rate below 1% while maximizing the true positive rate, $C = 10^{0.5}$ should be chosen. This, however, does not produce a great rate of true positives, only 1.9%.

f.

This can be solved by assigning the `class_weight` parameter in the `sklearn.svm.SVC` model. The `class_weight` will adjust the C value by different weights for different classifications. If we assign a weight of 1 to the 1 (spam) class, but 10 to the -1 (not spam) class, then misclassification of the not spam, as false postives, is penalized more heavily. Doing this with the model built for part b, we can get a true positive rate of 0.4% with no false positives when $C = 10^{4.5}$. This is superior to the true positive rate of 0.1% with no false positives from the un-weighted model when $C = 10^0$.

Running our weighted model with $C = 10^7$, we get 15.8% true positives and 0.4% false positives with good training and test error. We can see that it is possible to optimize the model to reduce false positives. While this reduces true positives as well, the overall result may be more useful as a spam filter.