Homework 3: Support Vector Machine

Joe Arriaga
CS 445: Machine Learning

March 15, 2019

1 Description

Several Support Vector Machines (SVM) were trained using the Python library `sklearn.svm` to classify emails from the University of California at Irvine Spambase dataset [1] as "spam" or "not spam".

Three experiments were performed to investigate the effects of feature selection, each trained an SVM using a different set of features and the results were compared. The first experiment simply trained an SVM with a linear kernel in the standard way using all features to act as a benchmark. The second experiment trained multiple SVMs using only the features identified to be the most important in the first experiment. Important, here, means that the weights in the SVM associated with those features had the greatest absolute value, and thus were the most predictive. The number of features used varied from 2 to all 57 features and the accuracies achieved by each of the 56 SVMs were compared. The third experiment similarly trained multiple SVMs using anywhere from 2 to all 57 features but the features were now selected at random for each instance, meaning that the features used in subsequent instances were not necessarily related, and the resulting accuracies were compared.

2 Results

2.1 Linear SVM using all features

The SVM trained using all the features performed well, achieving an accuracy of 90%, a precision of 86%, and a recall of 89%. The Recall Operating Characteristic (ROC) curve, shown in Figure 1, also depicts a quite competent classifier. The accuracy of the machine shows that it was able to correctly classify most examples, the slightly lower precision indicates that more of the errors were due to legitimate emails being classified as spam, and the recall shows that most of the spam emails were identified but some passed as legitimate.

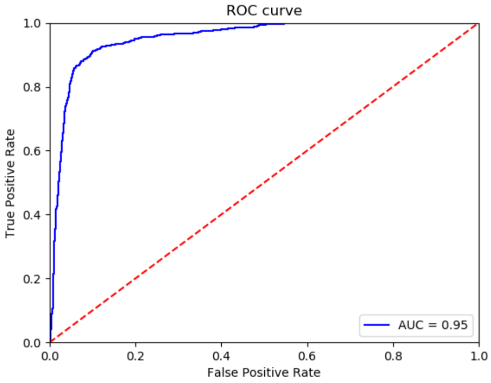


Figure 1: ROC curve for SVM using linear kernel on all features

2.2 Limited Features, Selected by Importance

All SVMs trained for the second experiment achieved a 90% accuracy commensurate with the accuracy of the SVM using all features. Because the same level of accuracy was achieved by all machines, even one using only the two most important features, one must conclude that using only those two features was sufficient and the other 55 features added no more useful information.

The top five most important features, in order, were the frequencies of the words: "george" (it is unclear whether capitalization was considered, likely all words were converted to lower case), "hp", "85", "cs", and "lab". It is noted in the documentation for the dataset that the non-spam examples were taken from personal emails and it is implied that the intention of the dataset is for creating a "personalized spam filter" rather than a general purpose filter [1]. Furthermore, the dataset was created by several people at Hewlett-Packard, one of whom is George Forman. Understanding where the examples came from (the sampling distribution) and the intended use of the dataset begins to illuminate why these particular features were the most predictive.

The top five predictive features, in fact the top eight, were all negative predictors. That is, the weights were negative meaning that if these words occured frequently the message was very unlikely to be spam. Because this was a personalized spam filter it is logical that emails with specific references to the "owner" of the filter or email account would be unlikely to be spam. It can be inferred that George Forman contributed many of his personal, legitimate emails to the dataset and this is to whom the "george" in the feature refers. Similarly, the dataset was created by a group and Hewlett-Packard and this is where "hp" comes from. We can deduce that "85" refers to the phone number of Mr. Forman at the time which is listed as 650-857-7835 in the documentation. It is unclear why the "85" substring is separated and somewhat surprising that it is such a predictive feature when one considers that "650" and "857" were also included features but these were the 37th and 41st most important features. The words "cs" and "lab" are also fairly specific to Mr. Forman and other contributors from the group, one assumes, but are not so tightly coupled as one's name, employer, or telephone number.

In conclusion, the fact that the most important features are all negative indicates that the machine operated largely by assuming an email was spam if it could not be actively identified as legitimate, this aligns with the slightly lower precision found in the first experiment. The mechanism of the filter seems to rely upon the fact that legitimate emails are very likely to contain specific information about the "owner", such as their name, while spam emails are likely to be much more general and contain no such specific information.

It is difficult to draw many conclusions about the effects of feature selection from this experiment because all subsets of features performed almost identically. However, the results are in line with what one would expect and it is only the degree of the effect that is surprising. It is likely that some features are more important than others for classifying an example. Furthermore, it is likely that there is a relatively small subset of key features that are most important for deciding. The difference then, is the number of key features and the relative weight each of them contributes to the final decision. Assuming this is the most probable situation one would expect that if the top few features contributed 80% of the final decision then a machine using only these few features would perform about 80% as well as a machine using all the features.

This case was extreme in that the top two features appeared to account for 100% of the final classification and thus there was no difference in performance for any number of features. Further investigation was performed running the SVM with only the very top feature and the performance was still the same. This indicates that simply the presence or absence of the word "george" determined if the email was legitimate or spam. This is likely due to a bias in the training set were almost all legitimate emails contained the word "george" while it is very unlikely that many spam emails contained "george". However, as stated in the documentation, this is perfectly acceptable for a personalized filter.

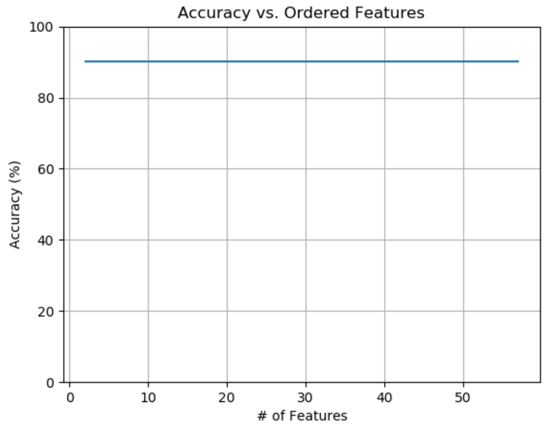


Figure 2: Accuracy vs. m for SVM using m most important features

2.3 Limited Features, Selected at Random

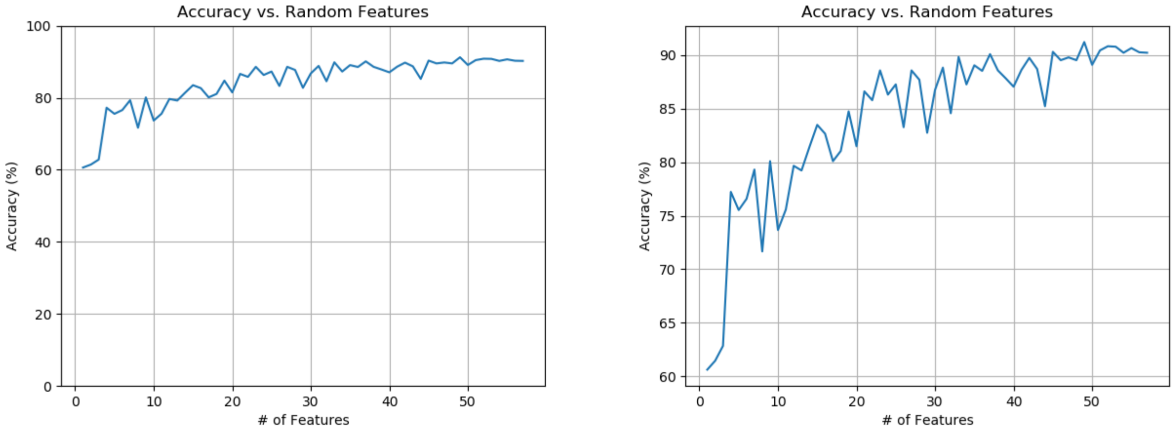
When features are selected randomly the general trend is for accuracy to improve as the number of features increases but the accuracy does not increase monotonically, but rather quite chaotic. In this instance, the accuracy started above 60% indicating that even less important features may indeed contribute to classification but are perhaps overshadowed by the more important features. The learning of the SVM instances can be divided into several stages characterized by the mean accuracy. By visual inspection of the plot we could divide the stages as follows:

Stage	Approximate Mean Accuracy (%)	Approximate Numbers of Features
1	61	1 - 3
2	77	4 - 11
3	82	12 - 20
4	86	21 - 34
5	88	35 - 45
6	90	46 - 57

One would speculate that the event that caused the accuracy to enter the next stage was the inclusion of one of the more predictive features. Within each stage, however, the accuracy is anchored around an average dictated by the few features that are very predictive and fluctuates due to the inclusion of less predictive features which the SVM must rely on but which often predict incorrectly.

Even when most features are being included in the model the range of accuracies remains relatively large, about 5%, and it is only in the last stage when nearly all features are being included that the accuracies become more consistent. It is also interesting to note that some models with fewer features achieved higher accuracies than the final model including all features, suggesting that some features may have been counter-productive and the model would perform better by identifying these and disregarding them.

As would be expected, when all features were included the SVM achieved the same accuracy as the model with weight-ordered features, although the weight-ordered model was able to achieve that accuracy with only a single feature rather than 57. One can only conclude that the top few features contribute the vast majority of the predictive power while the rest of the features are dwarfed in comparison, despite them actually having some predictive power when considered individually.



(a) Full-scale

(b) Zoomed

Figure 3: Accuracy vs. m for SVM using m randomly selected features

3 Conclusions

When one considers the mechanism SVMs use for classification, it is perhaps unsurprising that it relies on only a few features to discriminate between classes. Since only the support vectors have any real influence on the location of the decision boundary and the classification of examples it seems that it would be common for these support vectors to share a few key characteristics. This reliance on key characteristics also makes sense given that it is the same mechanism humans use in almost all cases to determine what a thing could be, that is, to classify some object. We tend to focus on certain characteristics and find similar characteristics of objects we already know. The particular characteristics we focus on may vary from person to person but the mechanism is largely the same. This method of classification is by no means foolproof and some features may actively mislead the decider but thousands of years of human experience has proven that for most situations it is good enough.

References

[1] Mark Hopkins, Erik Reeber, George Forman, and Jaap Suermondt. Spambase data set. https://archive.ics.uci.edu/ml/datasets/Spambase.