CS 545 ASSIGNMENT 5

Paresh Bhambhani

November 17, 2015

Contents

1 Filter methods

In this section we implement a python function that returns the Golub score. The implementation is furnished below:

```
def golub(data,labels):
   pos_examples = np.ones((1, len(data[0])))
   neg_examples = np.ones((1, len(data[0])))
   #Separating positive and negative examples
   for i in range (len(data)):
      if (labels[i] > 0):
         pos_examples = np.vstack((pos_examples,data[i]))
      else:
         neg_examples = np.vstack((neg_examples,data[i]))
   pos_examples = np.delete(pos_examples, (0), axis=0)
   neg_examples = np.delete(neg_examples, (0), axis=0)
   mean_pos = pos_examples.mean(axis=0)
   mean_neg = neg_examples.mean(axis=0)
   std_pos = pos_examples.std(axis=0)
   std_neg = neg_examples.std(axis=0)
   scores=np.zeros(len(data[0]))
   for i in range (len(data[0])):
      std_total=std_pos[i] + std_neg[i]
      if (std_total == 0):
         std_total = 1
      scores[i] = (np.abs(mean_pos[i] - mean_neg[i]))/(std_total)
   return scores,scores
```

2 Embedded methods: L1 SVM

2.1 Features with non-zero weight vector coefficients

Table 1: Number of Features with non-zero weight vector coefficients

Dataset	Total No. of features	Number of features with Non-zero weight vector coefficients
Arcene	10000	559
Leukemia	7129	40

Table 1 lists the number of non zero features for the two data sets in consideration. The number of features with non zero weight vector coefficients are averaged over 10 runs.

2.2 Accuracy of various approaches using cross validation

Table 2: Accuracy via Cross Validation

Dataset	Approach	Average Accuracy
Arcene	L1 SVM	0.6795
	L2 SVM trained on the features selected by the L1 SVM	0.5691
	L2 SVM trained on all the features	0.7845
	L2 SVM that uses RFE	0.7845
Leukemia	L1 SVM	0.9667
	L2 SVM trained on the features selected by the L1 SVM	0.9656
	L2 SVM trained on all the features	0.7923
	L2 SVM that uses RFE	0.9171

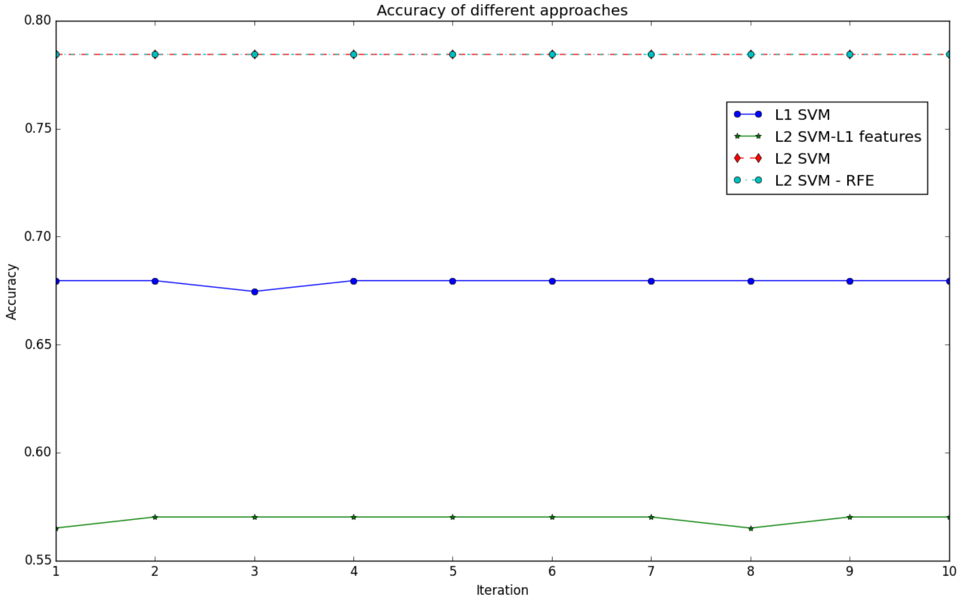

Figure 1: Arcene Data - Accuracy of various Approaches

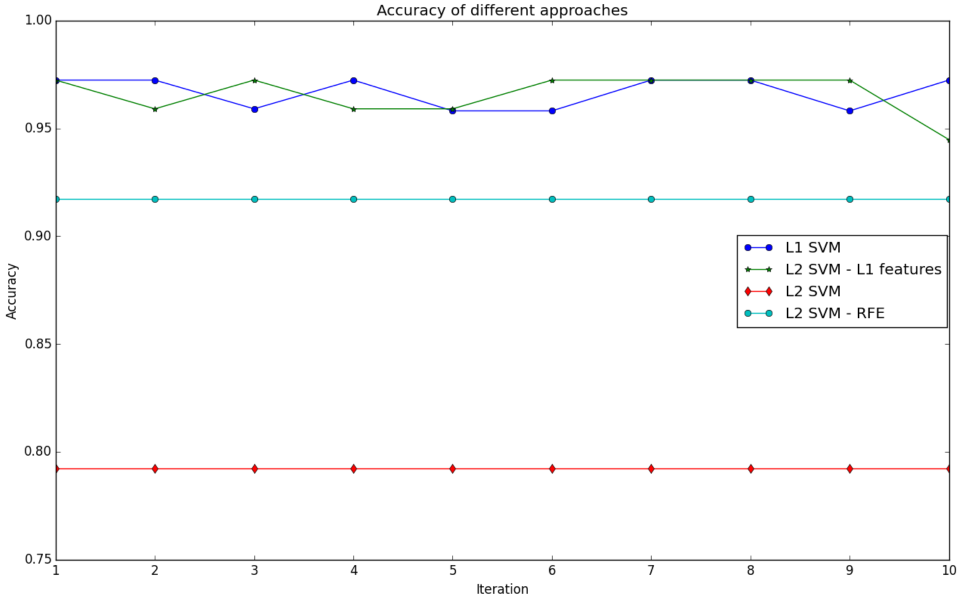

Figure 2: Leukemia Data - Accuracy of various Approaches

Table 2 lists the accuracy averaged over 10 runs for various approaches using the CV method.

Figures 1 and 2 show the accuracy plots for the four methods on the two datasets for the 10 runs.

It is quite clear from Table 2 and the Figures 1 and 2 that for Arcene dataset L1 SVM is not the way to go. The accuracy for L1 SVM and L2 SVM run on the features selected by L1 are pretty low. L2 SVM run on all features and L2 SVM with RFE perform better on Arcene.

On Leukemia dataset we get pretty good results from L1 SVM, L2 SVM trained on features selected by L1 and L2 SVM that uses RFE. Only the L2 SVM trained without any feature selection does not give as high accuracy. This is due to the fact that there are very few features with non zero weight vector coefficients in Leukemia (an average of 40 in 7129 as we saw in Table 1). Therefore without feature selection we are bound to get a lower accuracy.

2.3 Workaround for L1 SVM leading to sparse Solution

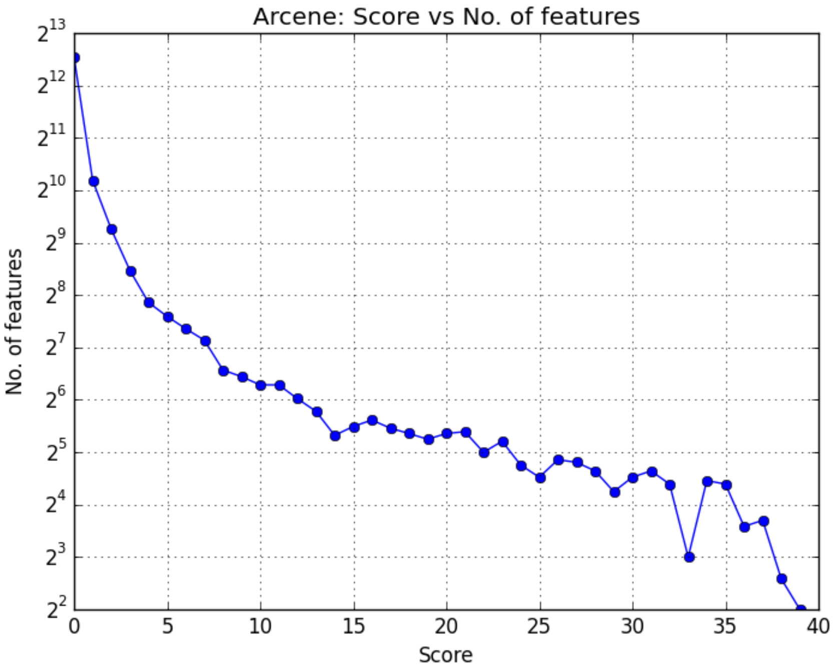

Figure 3: Arcene Data - Score of features

For this part of the problem the data was divided into k=40 subsamples where each sub-sample represents 80% of data chosen randomly. As instructed, for each sub-sample, L1-SVM was trained upon and for each feature of the data, a score was computed which reflected the number of sub-samples for which that feature yielded a non-zero score. The scores of each feature for both the datasets can be viewed when the submitted code is run. Since it is difficult to observe thousands of features, here we plot the number of features per value of score to help with the analysis.

Figures 3 and 4 show a plot of Score of features vs the number of features that have that score. It can be clearly seen that maximum number of features have a 0 score, which indicates that maximum number of features have a 0 weight vector coefficient for all the 40 sub-samples.

A comparison of slope of the curve in Figure 4 with the slope of the curve in Figure 3 reveals that there are comparatively more features in Arcene data having non-zero values than Leukemia dataset where very few features are with a non zero value.

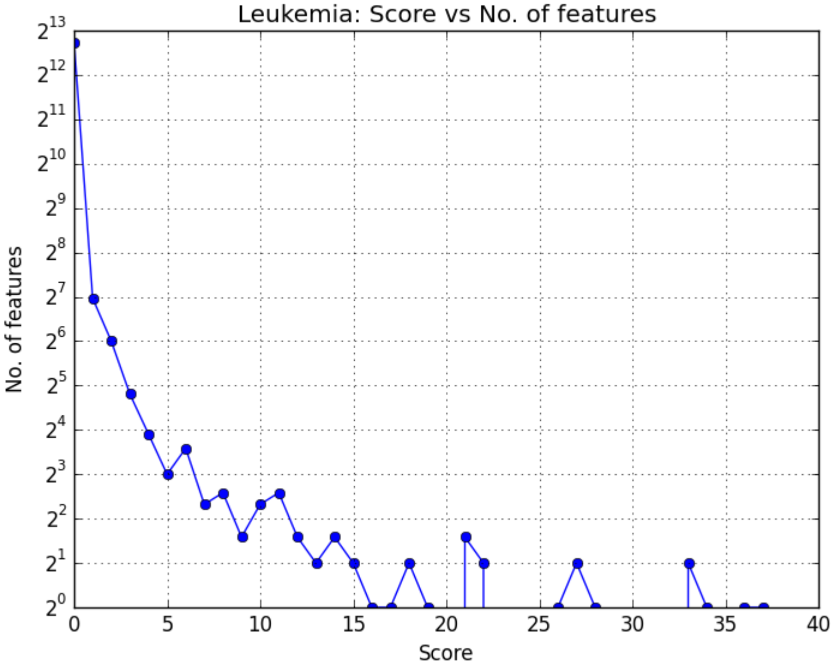

Figure 4: Leukemia Data - Score of Features

3 Method Comparison

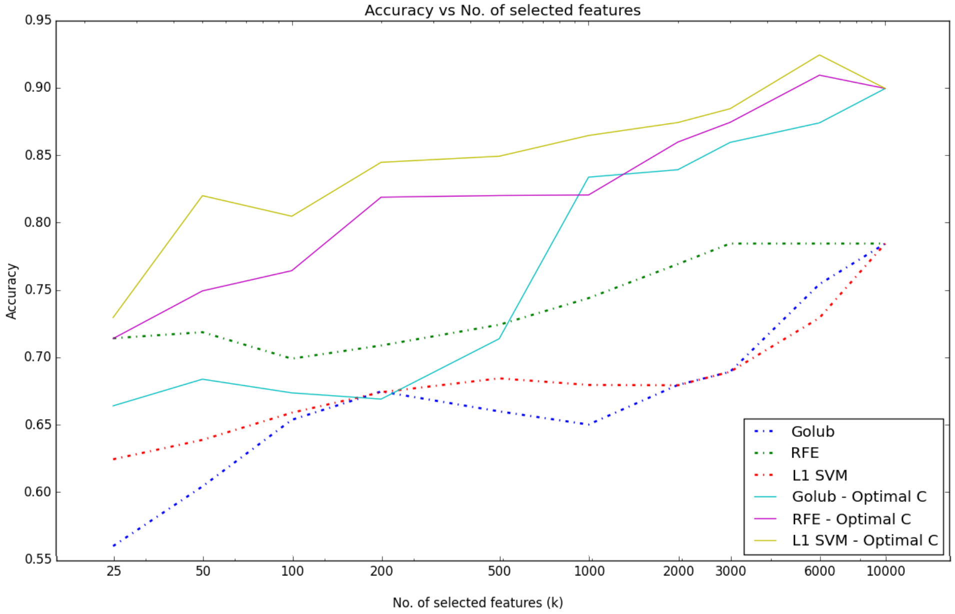

Figure 5: Arcene Data - Accuracy vs No. of features selected

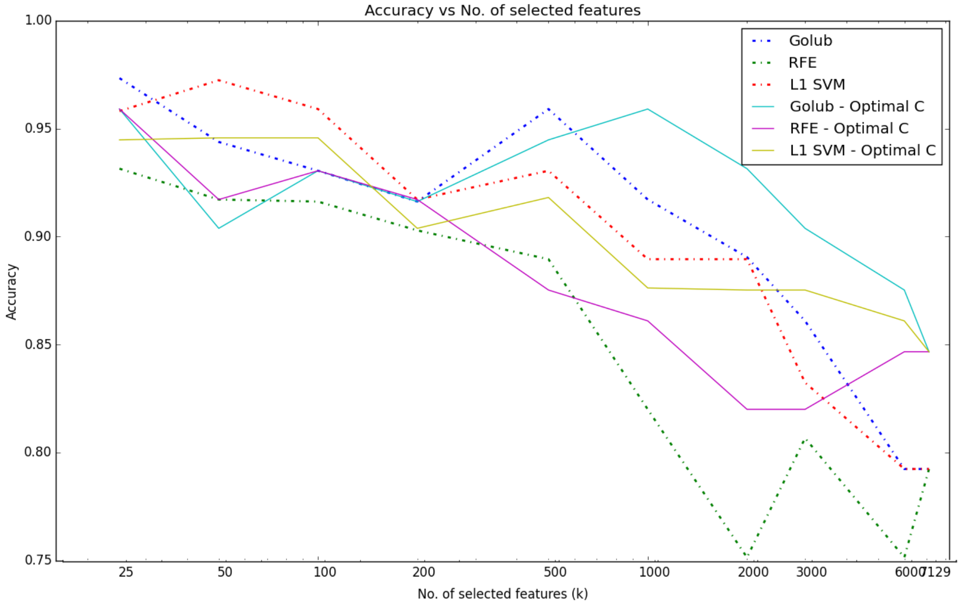

Figure 6: Leukemia Data - Accuracy vs No. of features selected

For Arcene data it can be seen from Figure 5 that the RFE method of feature selection works best, then the L1 feature selection method using sub-samples and lastly the Golub score method.

For Leukemia dataset shown in Figure 6, this order is exactly opposite. Golub scoring method performs better for the most part, then the L1 SVM feature selection method and lastly the RFE method.

The above described results are when we do not select the optimal value for the SVM soft-margin constant. When we select the optimal value of C using internal CV over a Grid of values, we see an increase in accuracy for all the feature selection methods in both the datasets. This is as per our intuition that since we are choosing an optimal SVM soft margin constant there should be a noticeable increase in the performance. The slopes of the three feature selection curves when optimal C is used in Figure 5, for Arcene dataset, shows an improved accuracy across number of features selected. This trend is also loosely visible for Leukemia dataset in Figure 6.

We can see in both Figures 5 and 6 that when all the features are selected, all the three methods result in the same accuracy value for both the dataset and for both the cases of working with and without an optimal C, showing that the feature selection methods will behave same when we are selecting all the available features.

It is interesting to note that the accuracy increases as the number of features selected increases for Arcene dataset whereas it decreases for Leukemia dataset. This can be ascribed to the fact that there are more features in Arcene data which are important which we also saw in Part 2 Figure 3, hence increase in the number of selected features will improve accuracy. Whereas in Leukemia dataset Part 2 Figure 4 we saw that the maximum number of features are with 0 weight vector coefficient and hence when we are increasing the number of selected features we are adding features mostly with small or zero weight vectors hence the accuracy goes down.