Bootstrapping the Support Vector Machine

Thomas Goerttler, Christian Koopmann, Patricia Craja

Department of Computer Science,
Machine Learning Group

Humboldt-University of Berlin, Germany
thomas.goerttler@gmail.com, c.k.e.koopmann@gmail.com,
Patricia.craja@gmx.de

Abstract

The goal of this project is the analysis of the variance of Support Vector Machines (SVMs) and the relationship between this variance and other important aspects of the SVM for both Linear and Gaussian Kernels. We calculate the variance of SVMs using the minimal distance of prediction points to the decision boundary by applying the bootstrap method.

Introduction

In contrast to probabilistic classifiers which provide classification with a degree of certainty, SVMs only predict the most likely class that the sample should belong to.

We apply the bootstrapping method to the SVM algorithm, i.e. drawing random training samples with replacement from the full training data set to train different SVMs. Through the repeated calculation of the predictions of the test data set we can estimate the prediction variance, which is an indicator of the degree of certainty of the SVM.

Since Predictions might be identical across all bootstrap samples, we use a real valued substitute: the minimal distance of each prediction point to the decision boundary (Figure 1).

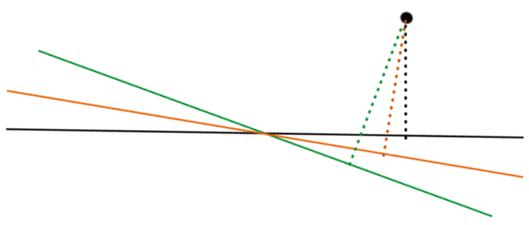

Fig. 1: The minimal distance of a prediction point to each hyperplane is considered.

Implementation

Bootstrapping

In order to calculate the variance of the predictions we use bootstrap samples. We first train the SVM on the full training data and will call this SVM the *Full SVM*. Then we draw N random bootstrap samples([1]) and train N separate SVMs on each bootstrap sample so we get N different set of predictions on the test data. Figure 2 shows, that the variance decreases with better separation and more data.

The n distances of each point in the test data set from the decision boundary can be seen as random variables. For the N SVMs we obtain N different values of these random variables and can calculate their variance. We take the average of these n variances as an indicator for the variance of the *Full SVM*.

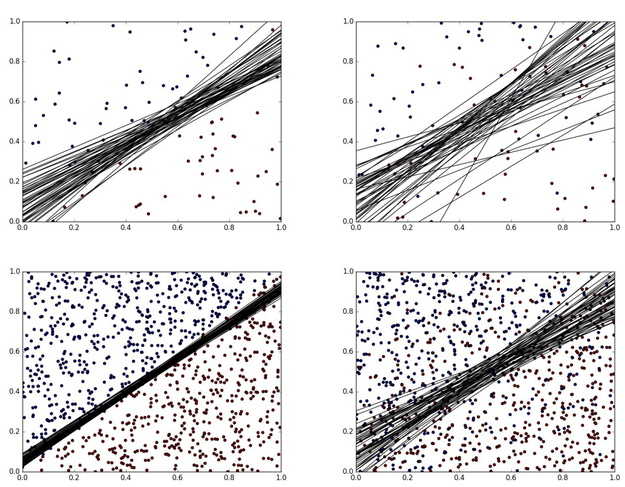

Fig. 2: 100 hyperplanes plotted based on samples,(a) and (c) use perfectly separated data, (b) and (d) not. In (a) and (b) use 100 observations in the training data,(c) and (d) 1000.

Data Simulation

We draw n observations for each of the input variables x from a normal distribution and use the Hyperplane-Approach: labels calculated using the formula $y = sign(c + w^T x + error)$. To change the number of support vectors of a data set the variance of the error is modified, while the intercept controls the balance of the data set.

Results

In this section we will show how the prediction variance is correlated with the regularization parameter C, the balance of the training data set and the number of support vectors of the original *Full SVM*. The size of each data set will be denoted by n, the number of bootstrap replications by N. The first graph corresponds to the Linear, the second to the Gaussian Kernel.

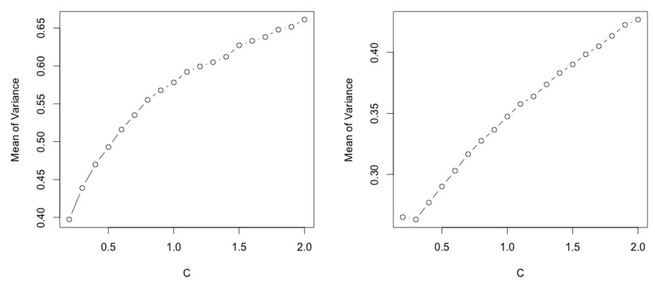

Fig. 3: Relation between the variance mean and the regularization parameter C. The y-axis represents the mean of the variance of 20 different data sets for each value of C(n=100, N=1000).

Influence of the regularization parameter C

Observations: For both Linear and Gaussian Kernel SVMs C is positively correlated with the variance of the SVM-predictions.

Interpretation: C controls the trade off between errors of the SVM on training data and margin maximization. To avoid overfitting, the SVM employs regularization and controls the amount of regularization by the constant C ([4]). A high value of C corresponds to low regularization, i.e. lower size of slack variables and therefore higher overfitting, which means that the SVM fits the training data too well but does not generalize to new data and therefore the variance of the predictions of our test data rises.

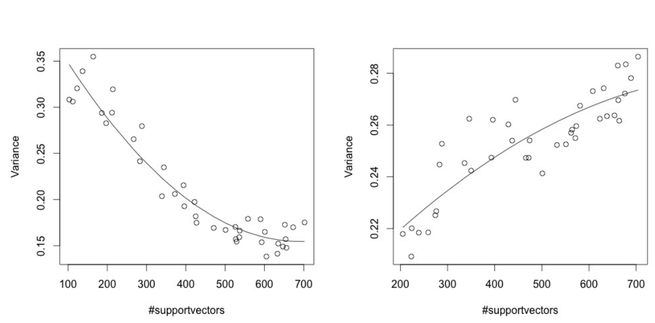

Fig. 4: Relation between the variance and the number of support vectors(n=1000, N=500)

Influence of the number of support vectors

Observations: The number of support vectors is negatively correlated with the variance of the Linear Kernel and positively correlated with the variance of the Gaussian Kernel.

Interpretation: The number of support vectors controls the slack variables of the soft margin SVM. A high number of support vectors implies larger margins and higher slack variables. Training the Linear SVM on the bootstrap samples is therefore more robust and the variance of the decision boundary as well as the distance from each point of the test data set to the decision boundary has smaller variation. For Gaussian Kernels, a higher number of support vectors might imply overfitting on the training data set and therefore bad generalization to new data which could be the reason why the variance of the predictions of our test data rises.

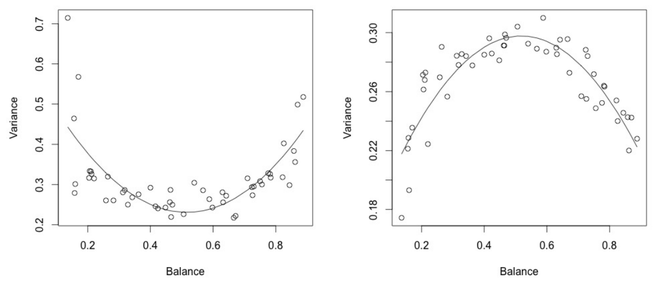

Fig. 5: Relation between variance and balance of the data (n=500, N=1000)

Influence of the balance of the training data set

Observations: Graphs are symmetric around 0.5 for both kernels. For the Linear Kernel, the variance of the SVM rises when the data is unbalanced, it is minimal for perfectly balanced data. For the Gaussian Kernel the maximal variance is attained for perfectly balanced data, it drops for unbalanced data.

Interpretation: In the linear case the variance is especially high, when one class is very small compared to the other class. Depending on which points are sampled from this class, the decision boundary and therefore the distance of test points to this boundary might vary dramatically. The Gaussian Kernel shows exactly the opposite results. This occurs because the Gaussian Kernel tends to overfit the data on the training data sets ([4]). If the data is perfectly balanced, more different samples of the smaller group are possible. Therefore the hyperplanes differ more. If the smaller class is very small, the sample from the smaller class can not differ that much from each other.

Conclusions

- The variance of distances of test points seems to be a good indicator for the robustness of the SVM.
- The results of changing the C-parameter are not surprising but confirm the strength of the variance estimator.
- Interesting effects of support vectors and data balance.
- Bootstrapping method has once again shown its value for situations where classical methods of statistical inference are not available.

References

[1] Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[2] Bradley Efron and Robert J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[4] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.