TCCM Homework:
Training an SVM classifier

Péter Bernát Szabó

1 Introduction

During this exercise I will train and SVM classifiers to differentiate between three kind of irises when given data about a plant's sepal width, sepal length, petal width and petal length. I will use the scikit-learn package to perform the training and analyze the results.

2 Obtaining the data set

I obtained the data from the builtin datasets of scikit-learn, by using the `load_iris` function implemented in scikit-learn. Then I exported the necessary arrays into a CSV file using the `csv` python module.

3 Training the SVM classifier

As this classification problem is relatively simple, my main goal was to obtain a similarly simple classification algorithm. To this end I decided to use only two features of the provided four for the classification. Additionally, keeping the dimensionality of the feature space low helped with the easy visualization of the decision boundary of the SVMs. To asses which two features are best suited to distinguish between the three species of iris I trained SVMs on all six possible feature combinations and compared the results.

After selecting the features to consider, I split the data set into a training and test set, by considering 70% of the available data for training and the rest for testing. For the fitting of the SVM I only used the training data. Next, I used the fitted SVM to predict the classes of the test data and assessed its accuracy. Finally, I created plots from the decision boundaries of all trained SVMs. In these plots (see Figure 1) I also show the data points in the corresponding test set.

4 Results

As can be seen on Figure 1 a linear kernel SVM is already capable of correctly classifying more than 97% of the test data points if the best suited features are used. The best suited set of features is the petal length and width, as this provides the highest accuracies and

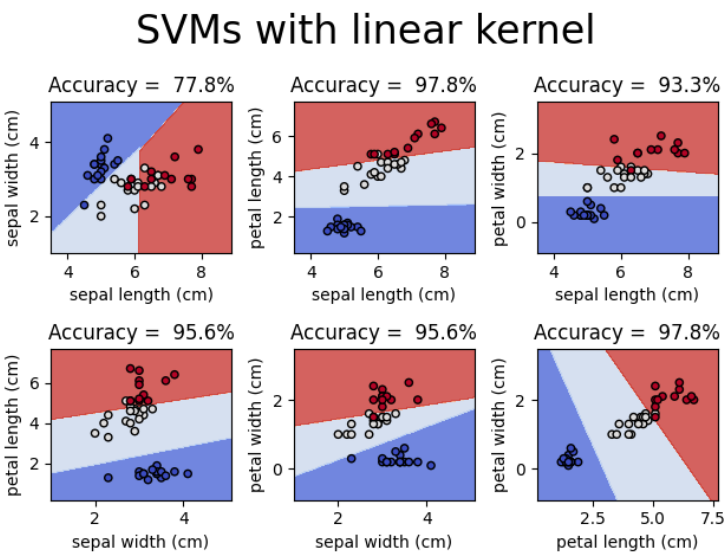

Figure 1: The classification boundaries obtained from various SVMs trained on different combinations of input features using a linear kernel.

furthermore the separation between the three classes is the largest in this case. The linearity of the kernel is verified by inspecting the shape of the decision boundaries: they are all straight lines.

After analyzing the results obtained with the linear kernel I explored a non-linear kernel as well: the radial basis functions (RBF) kernel. I expected that the additional complexity (being able to model curved decision boundaries) might help with the accuracy of the prediction. The plots obtained with these SVMs are shown on Figure 2. As can be seen from the figure however, the difference in accuracy (and the curvature of the decision boundaries) are non significant.

5 Conclusion

It was seen that the problem of classifying the species of irises can be performed quite well (97% accuracy) using a relatively simple model. Two input features and a linear kernel are sufficient to achieve the goal. It was also demonstrated that the additional complexity of using a non-linear kernel is not worth the increased model complexity as it does not provide any improvement in the accuracy.

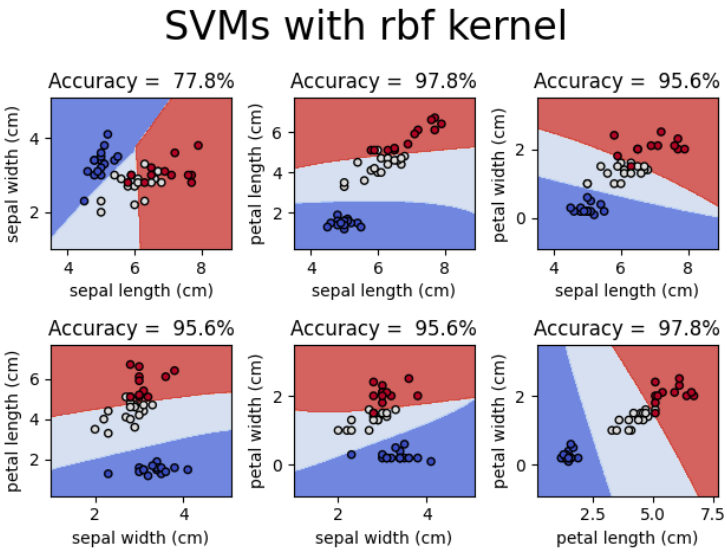

Figure 2: The classification boundaries obtained from various SVMs trained on different combinations of input features using an RBF kernel.