**Support Vector Machines**

*Exercise Session 1 - Classification,*

*report by,*

*Marimuthu Ananthavelu (r0652832), KU Leuven.*

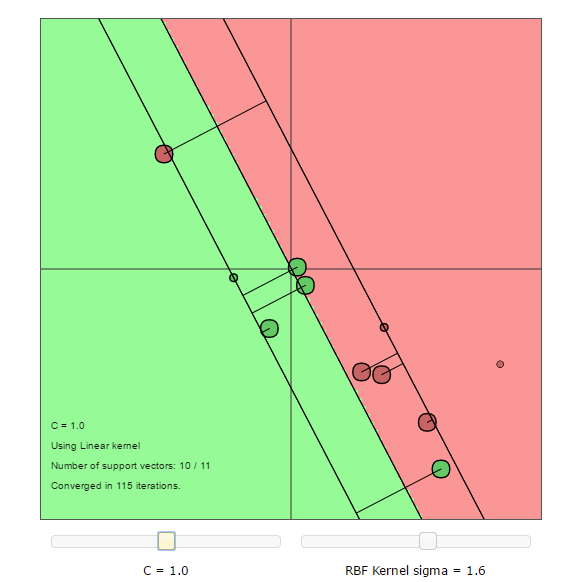
* 1. Two Gaussians

Geometric construction of a classifier line looks complicated given the data points are spread across each other. The classifying line should ideally be a decision boundary between the two classes namely ‘versicolor’ and ‘virginica. In our case, the decision boundary cannot be a straight line considering the nature of the data points which all are pertaining to two different classes. When the decision boundary separates these 2 classes with clearly indicating all the data points pertaining to ‘versicolor’ on one side and all the data points pertaining to‘virginica’ on the other side, then we can consider that the classifier is optimal/valid.

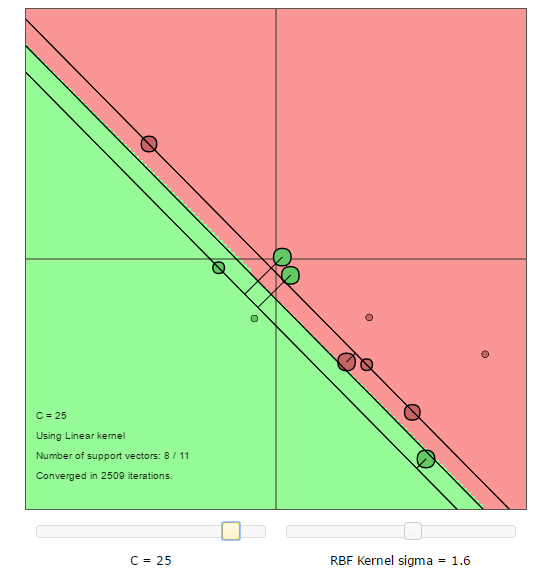
* 1. The Support Vector Machines
     1. When adding the new data points to the vector space the following happens;
* Each additional data point gets added to the dataset
* Decision boundary adjusts itself with respect to the modified dataset and accordingly the number of iterations gets modified with respect the inclusion of data point and re-optimization.
* When the new data point is added to the extreme left or right sides of the decision boundary, the existing decision boundary changes its position drastically (making a larger difference in the positional points in the vector space) considering the new data point.
  + 1. When adding an outlier, the following happens;
* With the same given parameters of ‘C’ and ‘kernel sigma’, the decision boundary changes with respect to the new data point and the new support vector is getting added to the side opposite of the class in which the outlier is present.
  + 1. C is the parameter for the soft margin cost function, which controls the influence of each individual support vector.

Changing the C value to different values the following are the observations;

* The increase in C parameter, makes the classifier to be more accurate as possible by penalising the misclassification heavily.
* The decrease in C parameter, makes the classifier to be leaving some of the data points under the wrong class which makes the classifier less robust.
* For example, with C parameter of ‘1’, the classifier is as below;



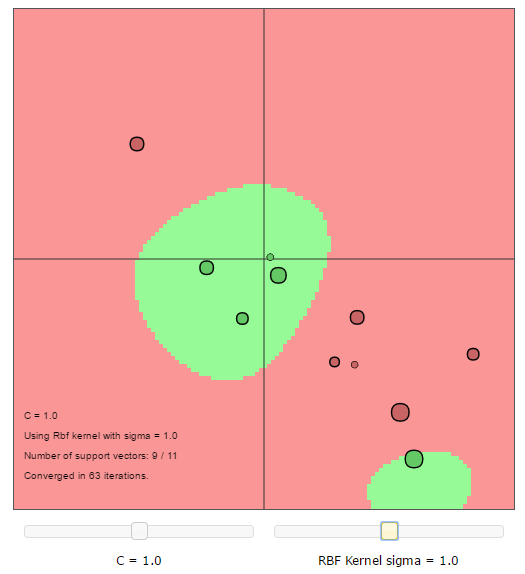
With C parameter of ‘25’, the classifier is as below;



* In Fig-1, with C=1, the classifier is not robust and classifies one or some of the data points on the wrong side of the decision boundary.
* In Fig-2, with C=25, the classifier attempts to be more robust and creates a very closely possible decision boundary especially giving importance to the data points which all are outliers.

***Using RBF kernel:***

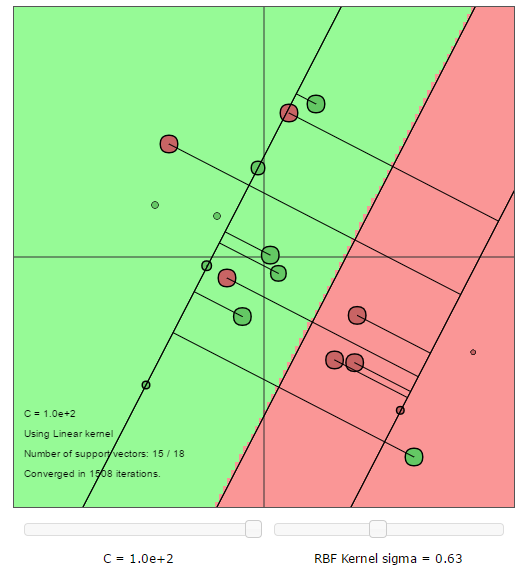
* + 1. Switching over to RBF kernel gives the decision boundary which is not a straight line but closed over the different cluster(s) of data points. This clearly indicates that the RBF can be used for non-linearly separable data points. The classifier space looks as below;



* + 1. Sigma is a parameter of the Gaussian kernel and defines the steepness of the rise around the landmark. The following changes happen with respect to the change in ‘sigma’ value.
* Increase in ‘sigma’ makes the decision boundary with stronger smoothing i.e. high bias with low variance
* Decrease in ‘sigma’ makes the decision boundary with overfitting for the chosen data points

Changing the both the hyperparameters provides the following the observations for a given kernel;

* The C Parameter implies the penalties on the misclassified data and thus higher C value provides better classified data with smooth boundaries.
* High C parameter value and low sigma value provides an optimal decision boundary
  + 1. Role of the Given Kernel, Regularization parameter, and the kernel parameter:
* Linear Kernel, is limited to classifying the data points which all are linearly separable using the straight line. They can’t be used for complex classification applications where as the linearly non-separable data are presented.
* C Parameter in both the kernels, gives an indication about smoothing of the decision boundary. The higher the value, the more penalty it awards for the misclassifying data points. The less they are, the kernel will leave some of the outliers and place the data points on the wrong side of the decision boundary.
* Sigma is applicable to the ‘RBF’ kernel function where the higher value puts the data points on the wrong side of the decision boundary whereas the small values make the decision boundary which overfits.
* In case of Linear kernel, an optimum C Parameter is required.
* In case of RBF kernel, an optimum C and Sigma parameter are required.
* Essentially *both* the kernel and regularization parameters (C) control capacity, and one can get a diagonal trough in the Cross-validation error as a function of the hyper-parameters because their effects are correlated and different combinations of kernel parameter and regularization provide similarly good models.
  + 1. Role of Support Vectors:



* In the above figure, the support vectors are the points which lie along the supporting hyperplanes (the hyperplanes parallel to the dividing hyperplane at the edges of the margin). Regardless of the number of dimensions or size of data set, the number of support vectors could be as little as 2.
* When the margin becomes very wide, there are lots of support vectors.
* The number of support vectors depend upon the how much C parameter is allowed for the given distribution of the data. If large value of C parameter is allowed, then we will have a large number of support vectors and If small value of C parameter is allowed, we will have very few support vectors.
* Accuracy of our classification depends on the trade-off between a high-complexity model which may over-fit the data and a large-margin which will incorrectly classify some of the training data in the interest of better generalization. The number of support vectors can range from very few to every single data point which will completely over-fit your data. This tradeoff is controlled via C parameter and through the choice of kernel and kernel parameters.
* The role of support vectors is …
* Addition of more data points increases the number of support vectors normally as the decision boundary changes and accordingly, the data points on close to the margin changes.
* A Particular data point becomes support vector when ….
  1. Using LS-SVMlab
     1. The performance on the given test set for iris dataset using the linear (kernel) function is as below:

*#misclass = 11, error rate = 55.00%*

When changing the kernel to polynomial with varying degree of polynomials, the misclassifications and the error rates are as below;





The observations when increasing the degree of polynomial have similar impacts as the value of Sigma decreases. The increased polynomial makes the classifier to separate the non-linearly separable data points using the curvature/quadratic based decision boundary.

* 1. Homework Problems
     1. The Ripley Data-set
     2. Breast Cancer Dataset
     3. Diabetes Database