Lab 8 Support Vector Machines

Problem Statement

We are going to study how to generate toy dataset and build different SVM model to predict class labels, estimate test error rates and tune parameters to obtain the best models.

Dataset 1 - linearly separable dataset

x:

i) Generate 100 2-dimensional points by rnorm() with `mean = 0` and `sd = 1`.

$\Longrightarrow$ You will obtain a matrix with 100 rows and 2 columns.

Remember to set seed to be 300 so that we will have the same matrix.

ii) Generate 100 2-dimensional points by rnorm() with `mean = 3` and `sd = 1`.

$\Longrightarrow$ You will obtain a matrix with 100 rows and 2 columns.

Remember to set seed to be 300 so that we will have the same matrix.

iii) Combine the above two matrices so that you obtain a new matrix called `x` with 200 rows and 2 columns.

[**Hint**: use `rbind` function to do this.]

y:

Generate a vector `y` of length 200 where the first 100 numbers are 1 and the last 100 numbers are -1.

Dataset: dat

Combine `x` and `y` to form a data frame and call it `dat`.

Plot the dataset:

Plot `x` such that the points are black (`col = 1`) if `y > 0` and red (`col=2`) if `y <= 0`. Add some legend to the plot if you can.

Generate training and test dataset

Now split the data into a training set (70%) and a test set (30%). Set seed be 400 before calling `sample()` function.

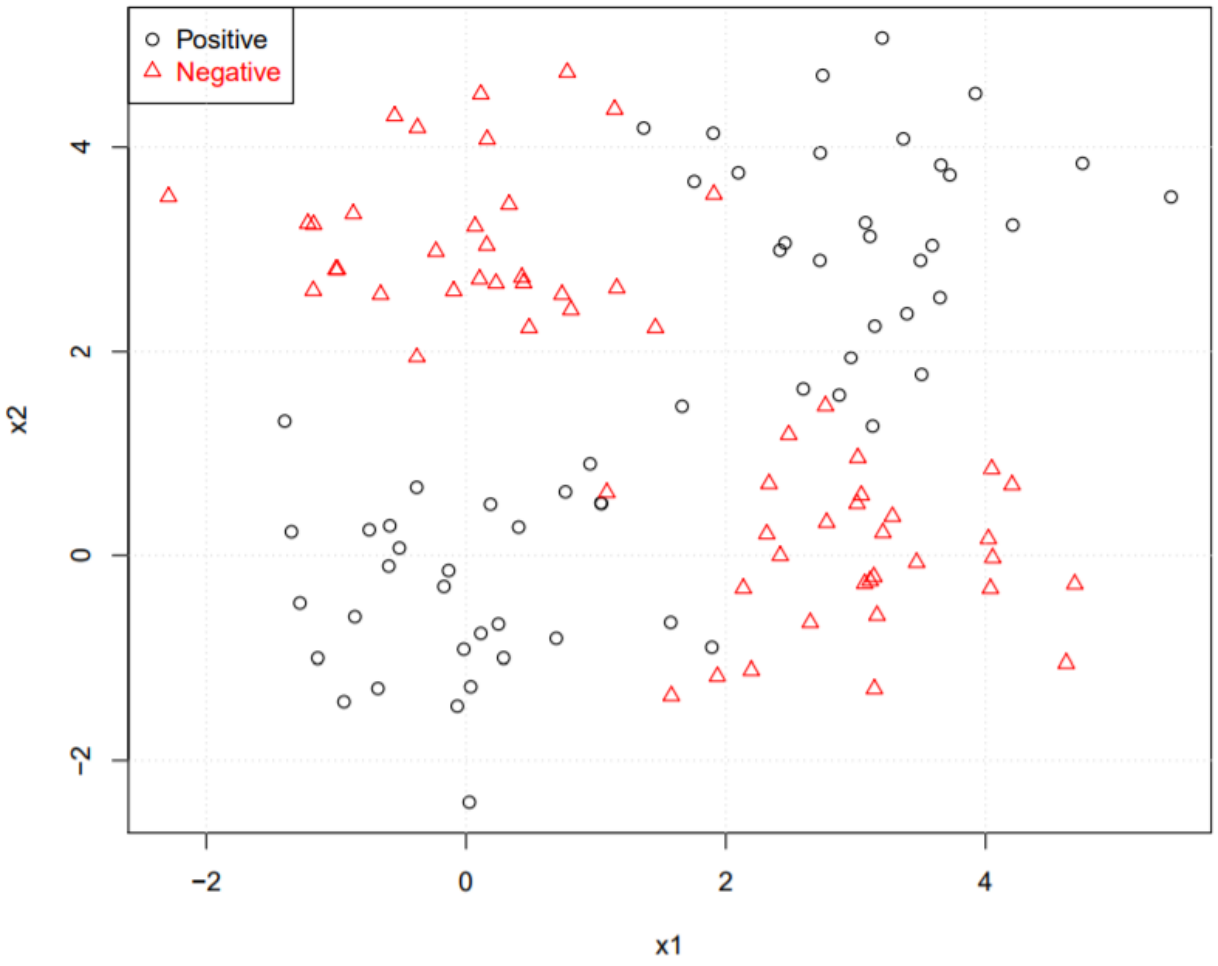
Questions to Dataset 1

1) Install and load the library `e1071`.
2) Build a linear SVM model with `cost = 1`.
3) Report how many support vectors are from each class respectively.
4) Plot the svm model and check how many support vectors are on the wrong side of the boundary and how many data points are very close to the margin.
5) Predict the class label of `y` on the test set and estimate the test error rate.
6) Now try a smaller `cost = 0.01` and a larger `cost = 1e5` and repeat step 2)-5).
7) Use `tune()` function to select the best model (tune the parameter `cost`). Set seed to be 1.

Dataset 2 - linearly inseparable dataset

Make a toy example that look like the following figure and test a linear SVM with different values of `C` or `cost`.

[**Hint**: you may use four `rnorm()` functions to generate this.]



Generate training and test dataset:

Now split the data into a training set (70%) and a test set (30%).

Questions to Dataset 2

8) Build an SVM model with a radial kernel, `gamma = 1` and `cost = 1`. In this model,

a) see the summary; b) plot the SVM; and c) estimate the test error rate.

9) Build an SVM model with a radial kernel, `gamma = 1` and `cost = 1e5`. In this model,

a) see the summary; b) plot the SVM; and c) estimate the test error rate.

10) Choose the best choice of gamma and cost for an SVM with a radial kernel.

11) Use the best model to estimate the test error rate.