ITM 6285 Data Mining Lab - Classification

Support Vector Machine Practice on a Made Up Dataset (expected time - 1 hour)

Learning objective:

1. Generate normal distributed random number

2. Learn about using colors in the plot function

3. Convert data into categorical values

4. Use the SVM function in the e1071 package

5. Learn to interpret the output of the SVM function

Task 1: Generate the made up data set

First initialize the randomization using set.seed(any number). Then try out the following codes:

x<-matrix(rnorm(40), ncol=2)   # rnorm(40) is the generate 40 normal distributed numbers

y<-c(rep(-1,10), rep(1,10))

Take a look at the two variables (x and y) you just generated, short answer what you just did with the two lines of code.

Task 2: Visualize the data

Plot the 2 columns in x and use 2-y as the color.

Hint: first column of a matrix: x[,1]; first row of a matrix: x[1,]. The color index should always be positive, so we have to use 2-y as the color.

Task 3: Combine x and y together as a dataset, with x being the two features and y being the category

We use the data.frame function to merge the two parts of the data. Use the following codes:

dat=data.frame(feature1=x[,1],feature2=x[,2], class=as.factor(y)) # the feature1, feature2 and class are just column names.

Task 4: Load the necessary package

Load the e1071 package and prepare for the svm

Task 5: SVM

Do a classification using the svm algorithm. Use linear kernel function, and let the misclassification cost coefficient be 10.

Task 6: Visualize the Decision Boundary

Plot the decision boundary. Copy the plot to the answer sheet.

Task 7: Prediction with our classifier

Generate the predicted classes based on the x values and our trained classifier. Are they the same as the **actual** classes?

Task 8: Calculate classification accuracy

Generate the confusion table. Copy the table to the answer sheet. And then calculate the accuracy of our svm model.