IST687 – Support Vector Machines Lab

Step 1: Load the data

Let go back and analyze the air quality dataset (if you remember, we used that previously, in the visualization lab). Remember to think about how to deal with the NAs in the data.

Step 2: Create train and test data sets

Using techniques discussed in class, create two datasets – one for training and one for testing.

Step 3: Build a Model using KSVM & visualize the results

1) Build a model (using the 'ksvm' function, trying to predict onzone). You can use all the possible attributes, or select the attributes that you think would be the most helpful.
2) Test the model on the testing dataset, and compute the Root Mean Squared Error
3) Plot the results. Use a scatter plot. Have the x-axis represent temperature, the y-axis represent wind, the point size and color represent the error, as defined by the actual ozone level minus the predicted ozone level).
4) Compute models and plot the results for 'svm' (in the e1071 package) and 'lm'. Generate similar charts for each model
5) Show all three results (charts) in one window, using the grid.arrange function

Step 4: Create a 'goodOzone' variable

This variable should be either 0 or 1. It should be 0 if the ozone is below the average for all the data observations, and 1 if it is equal to or above the average ozone observed.

Step 5: See if we can do a better job predicting 'good' and 'bad' days

1) Build a model (using the 'ksvm' function, trying to predict 'goodOzone'). You can use all the possible attributes, or select the attributes that you think would be the most helpful.
2) Test the model on the testing dataset, and compute the percent of 'goodOzone' that was correctly predicted.
3) Plot the results. Use a scatter plot. Have the x-axis represent temperature, the y-axis represent wind, the shape representing what was predicted (good or bad day), the color representing the actual value of 'goodOzone' (i.e. if the actual ozone level was good) and the size represent if the prediction was correct (larger symbols should be the observations the model got wrong).

4) Compute models and plot the results for ‘svm’ (in the e1071 package) and ‘nb’ (Naive Bayes, also in the e1071 package).
5) Show all three results (charts) in one window, using the grid.arrange function (have two charts in one row).

Step 6: Which are the best Models for this data?

Review what you have done and state which is the best and why.