Machine Learning Lab → Hw12

# Using SVM (and other models) on an Air Quality Dataset (Tutorial) and on an extended MPG cars dataset (Hw12)

Grab the Hw12starter.R file from the class Google Drive and follow these steps as a tutorial. For Hw12, you will use a different data set and modify the code to present your results.

**Step 1: Load the data**

Grab the air quality dataset (built into R). We will need 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 and visualize the results (prediction)**

1. Build a model (using the “ksvm” function, trying to predict Ozone). 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 (classification)**

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 “naiveBayes” (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. Include your explanation as comments in your R code.

**Hw12: Repeat this tutorial with the auto-mpg.csv data set.**

Use the tutorial code (Hw12starter.R) as the starting point for your solution. Edit the code to work with the new dataset. Here we will attempt to predict MPG from the data provided (note this is NOT the mtcars dataset). You will need to experiment a bit with the dataset to get the best fits in each situation. Do you need all the columns? Be sure to summarize your findings in Step 6, which are to be included in your R file as comments. The deliverable is your R code, and I must be able to run it. Send the R file itself -- not a version cut/pasted into word, not a pdf, etc.

Name your source file submission: ***LastNameFirstName*\_DS212Hw12.R** and submit via Google Classroom by the due date/time.