# Homework (HW10) Support Vector Machines (SVM)

### General Instructions

For this homework you will upload 1 R file into blackboard.

Reminder:

* All HW must start with an Identification Block like this sample…

################################################

# IST 387/687, Standard Homework Heading

#

# Student name:

# Homework number:

# Date due:

#

# Attribution statement: (choose the statements that are true)

# 1. I did this work by myself, with help from the book and the professor

# 2. I did this work with help from the book and the professor and these Internet sources: <provide the urls>

# 3. I did this work with coaching from <Name of another student> but did not cut and paste any code

# Run these three functions to get a clean test of homework code

dev.off() # Clear the graph window

cat('\014') # Clear the console

rm(list=ls()) # Clear all user objects from the environment!!!

# Set working directory

# Change to the folder containing your homework data files

setwd("~/MyDesktop/ISTX87/Homework")

This homework builds on our efforts from the Prep Exercise and depends on a careful read of Chapter 18 of *An Introduction to Data Science*. As usual we’ll use the Prep Ex R file that you created as a starting point. We’ll revisit the cleaned and conditioned “diamonds” dataframe that you created in the PE. In this homework we will use SVM to classify the “cut” of a diamond into one of two categories “Premium” or “Ideal”.

### HW10

**Step 1: Re-execute the Code Created in PE10.**

1. Re-run the entire R code you created during the Prep Exercise. You might receive a warning message for both “kernlab” and “ggplot2” that reads “Updating Loaded Packages”, select “No” and continue executing the code.
2. Using the *table()* command, identify the total number of “Premium” and “Ideal” observations within the dataframe. Record these observations in a comment.

**Step 2: Creating training and test data sets.**

1. Generate a set of random indices that will allow you to choose cases for your training and data sets and assign it to a new variable name. The range of your new indices should span from 1 to the final element index of the diamond subset data (35,342 if previous steps have been done correctly).
2. Build a training dataset and test dataset. The training datatset should be two-thirds of the data, and the test dataset should be one third of the data.
3. Use the *dim()* command to demonstrate that the resulting training dataset and test data contain the appropriate number of cases.

**Step 3: Build a support vector model.**

1. Create a support vector model based on the training dataset created in the previous step using the *ksvm()* function (or the *train* method, using the method svmRadial or svmLinear from the caret package). If needed, you can use the same parameters as shown on page 237 of the textbook.

**Parameters: kernal = “rbfdot”, kpar = ”automatic”, C = 5, cross = , prob.model = TRUE**

1. In a block comment, explain the above parameters. Be sure to thoroughly explain *C = 5*.
2. Store the output of the support vector machinefunction into a variable called svmOutput, and echo that variable to the console.

**Step 4: Predicting values in the test data.**

1. Using the *predict()* function, validate the model against the test data. Assuming that your test data set is in a dataframe called testData, the call would be:

**svmPred <- predict(svmOutput, testData)**

1. The object svmPred now contains a list of either “Premium” or “Ideal” cuts.
2. Using the *str()* and *head()* commands, review the contents of svmPred.

**Step 5: Create a confusion matrix.**

1. Create a confusion matrix that compare the second row of svmPred to the contents of testData$cut.
2. Calculate an error rate based on what you see in the confusion matrix. If you are stuck refer to pages 243 – 244 in the textbook.
3. In a block comment, explain how good of a model you have built.

**Step 6: Understanding the reasoning behind the practice.**

1. In a block comment answer the following question. Why is it valuable to have a “test” dataset that is separate from a “training” dataset?

***You must submit all Homework to blackboard prior to the deadline specified for each assignment.***

Late HW assignments will not be accepted for credit.

*Materials provided by the instructors of this course are copyrighted 2018 by Jeffrey Saltz and Jeffrey Stanton, or by the respective instructor that produced them. These materials are provided to enrolled students for their exclusive use during the semester of instruction. Uploading instructor-produced materials to Internet sites (e.g., Coursehero) without permission is considered a breach of professional ethics and will be reported as a possible academic integrity violation.*