Lab

# Lab 10 – SVM: Food Inspection

The chapter “What’s Your Vector Victor” introduces support vector machines – a powerful method of supervised machine learning. Support vector machines have proven value in classification tasks. At the simplest level a classification task is one where the system/algorithm must predict whether an outcome is this or that, true or false, sick or well, or any other dichotomous outcomes. In this final exercise for the semester, you will use a public dataset of restaurant inspections to see if you can predict whether or not a restaurant received an inspection violation.

As with the other recent exercises, this one has no sample R code. Likewise, the instructions are quite skeletal. As we reach the end of the semester, your diagnostic skills should be very sharp. As you run into barriers and problems, you should have all the necessary skills to solve the problems on your own.

1. The dataset is available at the NY State Health Data website:

<https://health.data.ny.gov/Health/Food-Service-Establishment-Last-Inspection/cnih-y5dw>

When you view this URL you will find that the dataset pops right up in your browser. Obviously to work with it in R you will have to download it to your own computer.

2. After you have loaded the data into R, take a close look at the VIOLATIONS variable. Because we want to test a simple dichotomous classification model, you will have to create a new variable that uses the data in VIOLATIONS to encode whether or not the restaurant had one or more violations. For this particular exercise, we don’t care how many violations there were, or what type, but simply whether there were some violations or not.

3. Next, apply all of the techniques you learnt from Chapter 18. This means that you will have to load the kernlab package, run ksvm(), set the parameters correctly, inspect the results, visualize the results, and make sense out of what you find. You should start with a very simple model that only uses one or two predictor variables. Which predictors in this dataset are likely to be most useful? You will have to give some thought to the process of restaurant inspection to consider what factors (among those available in this dataset) are likely to predict a violation.

4. After you have run and interpreted a model with one or two predictors, make sure to run a second model with additional predictors. How good can you make your prediction (i.e., how low can you drive the error rate)? For your final model, make sure to create, display, and interpret a confusion matrix. You should use the confusion matrix to calculate and report at least one error or accuracy rate.

Challenge layer: There is no additional challenge layer for this exercise, because most people will find step 4 sufficiently hard. Note that depending upon the power of the machine you are using to run these models, it might be quite time-consuming to develop the support vectors for each model. How can you get around this problem (short of getting a better computer)? The other challenging thing to pay attention to is this: Chapter 18 walks you through the steps of creating and interpreting an SVM model, but some of the codes in the chapter may not work exactly as you expect on these new data. You must be alert to changes in data structures so that you can modify your R code as needed to get around any problems you encounter.