The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Sukjin Han

Assignment 5

Due: Nov. 29th 11:00am

- Report all the codes and the outputs in answering the questions.

1. Consider the datset "heart.csv" on information about 303 patients who presented with chest pain. The dataset contains a binary outcome HD, which value of Yes indicates the presence of heart disease based on an angiographic test, while No means no heart disease. There are 4 predictors: $Thal$, Ca, $Oldpeak$, and $MaxHR$. For a patient whose $Thal = 1$, suppose the values of the 3 predictors, Ca, $Oldpeak$, and $MaxHR$, are

$$(x_1, x_2, x_3) = (1, 2.5, 150),$$

respectively. Suppose we want to determine whether this patient has heart disease or not. That is, we want to take those observation with $Thal = 1$ as the training observations, based on which we build a classifier and classify the test observation. Among the classification methods, we want to use the Support Vector Machine (SVM).

   (a) First we want to draw some scatter plots using the training set. Since $p = 3$, we cannot draw a scatter plot for all the covariates. Instead, draw scatter plots for pairs of predictors: $(Ca, Oldpeak)$, $(Oldpeak, MaxHR)$, and $(Ca, MaxHR)$. In doing so, indicate the class of each dot (i.e., whether or not the observation has a heart disease) using different colors or different shapes. Describe whether or not there exist a separating hyperplane in each scatter plot.

   (b) Your answer in (a) justifies the use of the SVM (instead of the maximal margin classifier or the support vector classifier) as your classifier. (i) Conduct the classification of the test observation given above by using the SVM. In doing so, you want to choose two tuning parameters, C and γ (using the notation in the ISL). Choose the optimal tuning parameters using the CV; the procedure is built in the package you are using, so please use it instead of manualling writing the code for the CV. (ii) Also, discuss how C and γ will change the bias and variance of the resulting classification, i.e., discuss the bias-variance trade-off of the choice of C and γ.