For this assignment, we were given two folders of data sets called ‘a’ and ‘b’, and there are 5 datasets within each of the folders. Each of these datasets contains a training subset and a testing subset. Our job was to build a regression model using the training subset and apply the model to the testing subset.

Firstly, I imported all the necessary libraries and packages including the model I plan to use. I did not end up using all the packages, but they are there just in case. Then, I wanted to get a good sense of the dataset by reading them in as panda data frames. I read each of the CSV files and named them either “train” or “test” with their directory names.

Then, I looked at train\_1a, which is the training subset of the dataset in ‘a/1’. We can see that for each of them, we have 13 features in total, and then an output label is also provided to train our models and select them. There are 177 rows and 14 columns in total.

Then, we began our model selection process. My goal was to select the best fit model there is using cross-validation. I first used SVR in sklearn.svm and the validation results were not ideal. Plus, even after dropping the points with high correlation, SVR gave me a1’s prediction to be all around -0.2, which clearly could not be the case. My next model choice was Bayesian ARD Regression, which fit the weights of a regression model using an ARD prior, and the weights of the regression model are assumed to be in Gaussian distributions. This method provided me with good cross-validation results, so I selected the Bayesian ARD regression model to be the final model I will use for all the data sets.

Then, I applied the model to all 10 data sets and saved the results as we went through each model process. Finally, we save the results to a csv file named “hw5\_results.csv”

For the future, I think it would be best if I can run cross-validation on each of the datasets and apply more models to them. The purpose of doing that is that the best fit model might be different on different datasets. So by doing that, we can personalize the model used on each data set, and achieve better results.