Question 7 – overfitting

I have decided to discuss about the following types of overfitting: Traditional Overfitting, Parameter Tweak Overfitting and Old Datasets.

# 1. Traditional overfitting

In the case of using too few examples for training a complex predictor, it may happen that it will not offer a good generalization and the error for the test sets will be higher.

For example, by using our dataset with all the 2500 patterns to build the linear regression model I obtained the root-mean squared error 0.0565 but if I create the model by using only the first 500 examples I get higher RMS: 0.0732. This happen because the fewer examples and a high number of parameters (11 in our situation) will predict a wrong redshift. If the redshift was depending only on 5 parameters and the examples in dataset offer a good generalization for the regression then the training model was going to be more accurate at predicting the values in test set.

The same thing could happen for using other machine learning algorithms to solve this problem: if creating a complex neural network and train it to offer a good accuracy on train set will result into overfitting if we use a very low number of examples and the behavior of model for test set will result into very high RMS because the weights into the nodes of neural network that will result from training will not offer a good generalization.

# 2. Parameter Tweak Overfitting

As it can be seen at the performance of random forests that I implemented for Question 2, the value of RMS for the model built on the train set is lower than the RMS obtained for applying the test set on that model. Still, the accuracy of non-linear model was better than for using regression because the random forests offer good generalization. It is certain that by also including the test set into creating the random forest I would have decided to use a larger number of trees to obtain better accuracy and tune the height of the tree and the number of parameters considered per split at internal nodes of the trees, then the accuracy of the model would have been higher for the test set but it does not offer good generalization and the performance is not relevant for using the model into real applications.

# 3. Old datasets

As it is described by John Langford, the old datasets have very well optimized model for obtaining good accuracy but it does not mean that they are very effective into calculating redshift for new data obtained through spectroscopy. The best way to make a good model is to also consider new datasets into training the model because it will make the machine learning algorithm to provide better generalization.

Another reason would be that in many domains the techniques for obtaining the features evolve, by using more accurate devices and algorithms and perhaps the photometric data about the galaxies is more accurate and calibrated differently. By using the newer data and offering it more weight into building the model will provide better accuracy into predicting the redshift for new patters that are obtained.