

HW5: Programming for Model Comparison

Mike Maccia

Task 1: Conceptual Questions

- What is the purpose of using cross-validation when fitting a random forest model?

Cross-validation is used when fitting a random forest model in an effect to prevent overfitting. Since random forest models are ensembles of decision trees, where each decision tree model should be created on different versions of the training set. We still want to only train our model using training data (preventing the model from seeing the test data), so using cross-validation allows us to create multiple folds of the data so each tree can be created on different versions of the training data.

These individual trees are combined into a random forest.

- Describe the bagged tree algorithm.

The bagged tree algorithm is a method of using bootstrapping to get multiple samples to fit on the same model. The results of these models would then be averaged together across multiple trees to create the final model. This allows for decreased variance when compared to an individual tree fit model.

Bootstrapping is a method of using resampling of data to create many samples. You would treat each sample as the population and this would then be used to fit the regression tree. The individual sample “y”s are then combined together to create the final model/prediction. In resampling of the data, normally it is done with replacement, so the same observation could be pulled into each individual sample multiple times.

- What is meant by a general linear model?

General linear models (GLM) are a form of modeling in which the response variable could be binary (success or failure, yes/no, etc). The GLM models are considered more flexible since they allow for this flexibility of binary variable. One example of a GLM is logistic regression in which the predicted probabilities of the response variable are a log function and would fall between 0 and 1. When graphing, this takes on a sigmoid shape. Even though the response variable is binary, continuous variables are able to be part of the prediction formula.

- When fitting a multiple linear regression model, what does adding an interaction term do? That is, what does it allow the model to do differently as compared to when it is not included in the model?

Adding an interaction effect allows for a more flexible surface when modeling a multiple linear regression. As opposed to a line of linear regression (plotted as x vs y), a multiple linear regression model would be plotted on more axes (in 3-dimensions). The model would instead be plotted as a “plane” and could take on a flat or saddle-like shape. An example of an interaction could be seen using this equation:

$$E[Y \mid x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \cdot x_2)$$

In this instance the effect of x_1 on Y differs based on the value of x_2 .

- Why do we split our data into a training and test set?

When creating predictive models, our goal is for our model to predict well for observations the model has never seen before. If we train our model on all of our data, there is a risk of overfitting and the model would not perform well on data it has not seen before.

In order to prevent overfitting, we create a training and test set by randomly splitting the data. Normally the data is split into either 80/20 or 70/30 training / test sets. The training data set is what we use to train the model. After, we can then predict using the test set and use a metric to determine how well our model worked.