Lab comments

Decision Trees

rpart model print

node), split, n, loss, yval, (yprob)

Node sub-section of the root population determined by the splitting decisions made by the model

split The attribute and value that the split happens on e.g CHK_ACCT < 1.5

How is the split determined?

For each node... the entropy of the population is determined, $H(Y) = -\sum P(y) \log P(y)$. Then, the entropy of the population given each independent variable is determined, H(Y|X). Joint information gain, I(X, Y), is determined by the following formula, H(Y) - H(Y|X). The IV, X, which gives us the largest information gain is chosen as the next split in building the decision tree. This process is repeated until either the target variables classes are all separated into their own buckets or some other factor such as maximum tree height is reached.

Intuitively the formula I(X, Y) = H(Y) - H(Y|X) makes sense as a means of choosing which variable to split on. This is because entropy is a measure of uncertainty and if H(Y|X) is *less* uncertain, than the variable X is a good predictor of the data. It provides information about Y.

If the predictor is continous then one could decide to take every 10% value, i.e the the value 10% of the way into the data, 20% and so on...

It's common to perform binary classification, if you have a multiclass predictor with greater than 2 classes, you determine the gain achieved from a single class against all other classes. Do this for each class and choose the class with the greatest gain as the split.

n number of entries in the population

loss The number of NON-majority class entries

yVal Your majority class in the given population sample

(yprob) The percentage of the population that the classes make up.

CP-Table

CP Complexity Parameter is used to control the size of the decision tree. It is a cost associated with adding another node to the tree. If the cost of adding another node to the tree is greater than the complexity parameter, than the tree stops growing.

Rel error Relative error is the error for predictions made by the model on the training data. The number of samples in the training data that are misclassified by the model. As the number of branches in the decision tree grows the relative error should always improve as the granularity of the models ability to sort data improves.

Relative error is obtained by the following calculation, $1 - R^2$ where R^2 is the root mean square error.

xerror This is the cross validated error of a model.

Cross validation is a method used for model selection/building whereby the dataset is split into training and testing data. In k-fold cross validation you split the data in to k subsets. You then train a model on k-1 of the subsets. The remaining subset of data is used to validate the models performance. At the end you can choose the model which performs best. In the models you can vary the IVs used, parameters chosen and any number of other variables which could impact the generated model.

xstd Is the standard deviation of the nodes population

Lab 3

apply function (lapply etc.)

apply is essentially the same as the map function in any other language. pass in your matrix, 1 or 2 (row or column), and the function to execute on each piece of data

lapply is executed on a list objects and returns a list object of the same length as the original set.

Random forrest output

Growing multiple models from randomly selecting subsets of the dataset (bagging) as well as randomly selecting subsets of the features (feature bagging) to use for the model training. These are often referred to as the *weak* models. The goal is to build a strong model from these many weaker models. To do this, the final model is selected using a method of model aggregation, such as taking the modal class at each split, combining the weak models into a stronger one.

```
Call:
 randomForest(formula = as.factor(Survived) ~ Pclass + Sex + Age +      SibSp + Parch + Fare + Embarke
d, data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 18.09%
Confusion matrix:
    0   1 class.error
0 391  42  0.09699769
1  87 193  0.31071429
```

OOB stands for *out of bag*

OOB estimate of the error rate: What is happening here is some data is being maintained for validation, the out of bag data. the amount of miss-classified data is the error rate. Looking at the table above the row indicated the class the data should have been and the column indicated what class the data was predicted to belong to.

Importance Determines the value of splitting on a variable, determined by averaging the decrease in impurity of the model by splitting on that variable over all the models in the random forest. The Predictors/Variables which reduce the amount of impurity/heterogeneity more are deemed to be of a higher importance. For classification, node impurity is measured by the gini index and for regression by the residual sum of squares.

Lab 4 regression tree

Will output numeric values instead of discrete categorical values like classification trees.

the constant to predict is based on the average of all values that fall into a subgroup.

for each subset of data, every distinct value of every predictor is used to determine the possible split that could occur that would split the data into two regions (R_1, R_2) such that the overall sums of squared error are minimized:

$minimize\{SSE = \sum_{i \in R_1}(y_i - \bar{c}_1)^2 + \sum_{i \in R_2}(y_i - \bar{c}_2)^2\}$

The lower the SSE the better the a split is at fitting the trend in the data. It has a lower variance. The opposite is true with a higher SSE

where y_i is the actual value of the target variable and c_x is the mean of the predicted values in a region

A possible split is the average between two consecutive data points.

Lab 5 data imputation with mean, median and regression

the cbind function takes the columns, the variables/predictors, and combines them into a new data frame. It is short for column bind

the lm function stands for linear model and is used to create simple regression models

Acronyms