Bagging, Boosting, Random Forests

Exercises

1. Mark the following statements as *true/false*. If false, explain why.

   a. Bootstrap samples are just simple random samples of a dataset where we take 63% of the observations for training.

   b. Some observations are likely to be repeated in a bootstrap sample.

   c. Bagging involves training classifiers on several different bootstrap samples of the same training data, creating several models whose results can be ensembled into a final result.

   d. If you implement bagging on a model with high variance, the resulting model is going to have even higher variance.

   e. Random Forests work best when the predictions of the underlying trees are not correlated.

   f. Random Forests provide you with a few simple, interpretable rules with which to classify new observations.

   g. Boosting is a technique that helps a model focus on observations for which the target is easy to predict.

   h. Gradient Boosted Trees are typically slower to train than random forests.

   i. The essence of gradient boosting is the iterative prediction of residuals from the previous round of modeling.

2. In addition to bagging, what additional protocol is enacted to create a random forest? In other words, do we simply train a collection of trees on bootstrap samples or is there some other step that takes place in a random forest?

3. To create a bootstrap sample in R, the following code can be implemented on a data frame or vector, x, which has n observations/elements:

```
sample(x,n, replace=T)
```

Let's try a simulation experiment. The following code creates 100 sample datasets from a vector with n observations/elements. It then calculates the average proportion of the original observations/elements that are contained in each sample.

```
n=10
x=1:n
samples = matrix(NA,100,n)
for (i in 1:100) {
  samples[i, ]=sample(x,n,replace=T)
}
numObs = apply(samples, 1, function(x) {length(unique(x))})
mean(numObs)/n
```

Calculate this proportion for n=10, n=100, and n=1000. Consider graphing the result as the size of the original vector n grows. The theory says that as the size of the original dataset grows, the proportion of the original observations contained in a bootstrap sample approaches 63%. (The theoretical value it approaches is actually $1 - \frac{1}{e} \approx 0.63212$.) Do you see anything like this through experimentation?

List of Key Terms

Bootstrap Sample

Bagging

Boosting

Random Forest

AdaBoost

Gradient Boosting

Regularization