Statistical Machine Learning 4 — Trees: Bagging and Random Forests 10/16/2014

Lecturer: Prof. Homrighausen Scribe: Zach Weller

1 Aaron Nielsen: the QUEST of Find Decent TAs

To start the class, Aaron gave a brief exposition on the QUEST tree algorithm and its use on a data set of TA ratings from the University of Wisconsin-Madison. A general outline follows:

- Review of advantages and disadvantages of trees
- FACT tree algorithm and an application of FACT to the "iris" data set (trees created by FACT are not necessarily binary)
- QUEST (quick, unbiased and efficient statistical tree) algorithm: Loh & Shih, 1997
  1. Uses Bonferroni corrected p-values from ANOVA or Chi-sq test (for independence) to select variables to split on
  2. Application: predict instructor rating (poor, average, good) based on covariates
  3. Some of the covariates used: summer vs. regular class, English speaker: native vs. non-native, class size
- Take away: QUEST algorithm was much faster than an exhaustive search

2 Bagging-bootstrapping Review

- Goal: keeping the same amount of bias, we want to reduce the variance in our predictions by averaging over many bootstrap trees
- Bagging for regression: never increasing prediction MSE
- Bagging for classification:
  1. no easy decomposition of bias + variance
  2. predictions can actually be made worse
- Pro: lower variance in predictions
- Con: lose some interpretability

3 Bagging Trees and its Benefits

Some benefits of bagging trees:

1. Mean decrease variable importance: gives us an indication of the importance of a variable[1]

[1] A very important caveat to this interpretation will be coming up when we talk about adding random, uncorrelated predictors

   (a) For each of the B trees and each of the p variables, we record the amount that the Gini index is reduced by the addition of that variable

   (b) Report the average reduction over all B trees.

2. Out-of-Bag (OOB) error estimation:

   (a) One can show that, on average, drawing n samples from n observations with replacement results in about 2/3 of the observations being selected.

   (b) The remaining one-third of the observations not used are referred to as out-of-bag (OOB)

   (c) We can think of it as a for-free "cross-validation"

3. Permutation variable importance: consider the b^{th} tree T_b

   (a) The OOB prediction accuracy of T_b is recorded

   (b) Then, the j^{th} variable is randomly permuted in the OOB samples ie: If $\mathcal{D}_{OOB,b} = (Z_{t_1}, \ldots, Z_{t_{n/3}})$ then draw a new $t_1^*, \ldots, t_{n/3}^*$ without replacement and predict $\mathcal{D}^*_{OOB,b} = (Z^*_{t_1}, \ldots, Z^*_{t_{n/3}})$, where $Z^*_{t_i} = (Y_{t_i}, X_{t_i,1}, \ldots, X_{t_i^*,j}, \ldots, X_{t_i,p})$

   (c) The prediction error is recomputed and the change in prediction error is recorded

4. Proximity plot

   (a) For the b^{th} tree, we can examine which OOB observations are assigned to the same terminal node

   (b) Form an $n \times n$ matrix P and increment $P[i,i'] \leftarrow P[i,i'] + 1$ if Z_i and $Z_{i'}$ are assigned to the same terminal node

   (c) Now, use some sort of dimension reduction technique to visualize the data in 2-3 dimensions. Multidimensional scaling is most commonly used (between observation distances are preserved)

   (d) The idea is that even if the data have combinations of qualitative/quantitative variables and/or have high dimension, we can view their similarity through the forward operator of the bagged estimator

4 Random Forests

Idea: because bagging imposes no restrictions on which predictors are used to split the trees, trees could potentially be highly correlated. We wish to reduce variance by averaging over bootstrap trees that are decorrelated.

Procedure:

1. Draw a bootstrap sample and start to build a tree.

2. At each split, we randomly select m of the possible p predictors as candidates for the split.

3. A new sample of size m of the predictors is taken at each split. Usually, we use about $m = \sqrt{p}$ ($p/3$ for regression)$^{\#}$

   $^{\#}$ Dr. Leo Breiman from the Department of Statistics at UC-Berkeley, one of the developers of the random forest algorithm, has a nice paper on the choice of the number of features. Brieman actually suggests using $m = \lfloor log_2 p + 1 \rfloor$. While I didn't find any specific reference to the $\sqrt{p}$ quantity, I felt the important take away from Brieman was the following. When choosing m, it is important to choose in such a way that:

(a) It is small enough such that the trees are decorrelated (large $m \implies$ high correlation)

(b) It is large enough that the trees have strength, i.e. the individual trees are good classifiers

Idea: (continued)

1. Suppose there is 1 really strong predictor and many mediocre ones.
   - Then each tree will have this one predictor in it,
   - Therefore, each tree will look very similar (i.e. highly correlated).
   - Averaging highly correlated things leads to much less variance reduction than if they were uncorrelated.
2. If we don't allow some trees/splits to use this important variable, each of the trees will be much less similar and hence much less correlated.
3. Bagging is Random Forest when $m = p$, that is, when we can consider all the variables at each split.
4. An average of B i.i.d random variables has variance
$$\frac{\sigma^2}{B}$$
5. An average of B i.d random variables with correlation ρ has variance (not independent)$^{\#}$
$$\rho\sigma^2 + \frac{(1-\rho)\sigma^2}{B}$$
Proof. Let $Var(X_i) = \sigma^2$, and $Corr(X_i, X_j) = \rho, i \neq j$. Then $Cov(X_i, X_j) = \rho\sigma^2, i \neq j$.
$$\begin{aligned} Var\left(\frac{\sum_{i=1}^{B} X_i}{B}\right) &= \frac{1}{B^2}\left[\sum_{i=1}^{B} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)\right] && (1)\\ &= \frac{1}{B^2}\left[B\sigma^2 + 2\sum_{i<j}\rho\sigma^2\right] && (2)\\ &= \frac{1}{B^2}\left[B\sigma^2 + B(B-1)\rho\sigma^2\right] && (3)\\ &= \frac{\sigma^2}{B} + \rho\sigma^2 - \frac{\rho\sigma^2}{B} && (4)\\ &= \rho\sigma^2 + \frac{\sigma^2(1-\rho)}{B} && (5)\end{aligned}$$
■
6. As $B \to \infty$, the second term goes to zero, but the first term remains. Hence, correlation of the trees limits the benefit of averaging.

Trees, Bagging, and Random Forests in R:

- tree package: used to fit binary classification or regression trees
- randomForest package: this can be used for both random forests and bagging (set $mtry = p$ in the "randomForest" function)

5 Model Evaluation: Sensitivity & Specificity

Consider a medical test used to identify whether or not a patient has a disease. Then,

1. Sensitivity: probability of a positive test, given the patient is ill; $P(+|ill)$
2. Specificity: probability of a negative, given the patient is healthy; $P(-|healthy)$

We can think of this in terms of hypothesis testing. If we have H_0 : patient is healthy, then

Sensitivity:	P(reject $H_0 \| H_0$ is false), that is: 1 - P(Type II error)
Specificity:	P(accept $H_0 \| H_0$ is true), that is: 1 - P(Type I error)

Reporting our results in the confusion matrix:

		Truth	
		Healthy	Ill
Our	Healthy	(A)	(B)
Predictions	Ill	(C)	(D)

For each observation in the test set, we compare our prediction to the truth. The total number of each combination is recorded in the table.

The sensitivity is given by:

$$\frac{\text{(D)}}{\text{(B)} + \text{(D)}}$$

The specificity is given by:

$$\frac{\text{(A)}}{\text{(A)} + \text{(C)}}$$

The overall miss-classification rate is

$$\frac{\text{(B)} + \text{(C)}}{\text{(A)} + \text{(B)} + \text{(C)} + \text{(D)}} = \frac{\text{(B)} + \text{(C)}}{n_{\text{test}}}$$