Notes on Statistical and Machine Learning

Random Forest

Chapter: *15* **Prepared by:** *Chenxi Zhou*

This note is prepared based on

- *Chapter 15, Random Forests* in Hastie, Tibshirani, and Friedman (2009), and
- *Chapter 14, Committee Machines* in Izenman (2009).

I. Definition of Random Forests

1. **General Idea of Random Forests:** *Random forest* is a modification of bagging that builds a large collection of *de-correlated* trees, and then averages them.

2. **Review of Relevant Methods:**

   (a) *Trees:*

   - Trees can capture complex interaction structures in the data and have relatively small bias, if it is sufficiently large;
   - Trees are noisy and have large variance.

   (b) *Bootstap Aggregating (Bagging):*

   - Has relatively low variance comparing to trees;
   - Each tree in bagging is identically distributed, the expectation of an average of B such trees is the same as the expectation of any one of them. This means the bias of bagged trees is the *same* as that of the individual trees.

   (c) *Boosting:*

   - Boosting combines the outputs of many weak learners to produce a powerful committee;
   - Trees in boosting are grown in an adaptive fashion to remove bias.

3. **Facts:**

   (a) Let $X_1, \cdots, X_B$ be i.i.d random variables and $\mathrm{Var}[X_b] = \sigma^2$ for all $b = 1, \cdots, B$. Then, the average of all X_b's has the variance $\frac{1}{B}\sigma^2$.

   (b) If X_b's are simply identically distributed with positive pairwise correlation ρ, the variance of the average is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2. \tag{1}$$

Proof. (a) Note the following

$$\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B} X_b\right] = \frac{1}{B^2}\mathrm{Var}\left[\sum_{b=1}^{B} X_b\right] = \frac{1}{B^2}\sum_{b=1}^{B}\mathrm{Var}[X_b] = \frac{1}{B}\sigma^2.$$

(b) Since now $\mathrm{Cor}(X_b, X_{b'}) = \rho > 0$ for all $b \neq b'$, we have

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B} X_b\right] &= \frac{1}{B^2}\mathrm{Var}\left[\sum_{b=1}^{B} X_b\right] \\
&= \frac{1}{B^2}\left[B \cdot \mathrm{Var}[X_b] + 2 \cdot \sum_{b<b'} \mathrm{Cov}(X_b, X_{b'})\right] \\
&= \frac{1}{B}\sigma^2 + \frac{2}{B^2}\frac{B(B-1)}{2}\rho\sigma^2 \\
&= \frac{1}{B}\sigma^2 + \left(1 - \frac{1}{B}\right)\rho\sigma^2 \\
&= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.
\end{aligned}$$

■

Remark. As B increases, the second term in (1) disappears, but the first remains. Hence, the *size of the correlation* of pairs of bagged trees limits the benefits of averaging.

4. **Algorithm of Random Forest:** The main idea behind *random forests* is to *improve the variance reduction* of bagging by *reducing the correlation* between the trees through random selection of the input variables, without increasing the variance too much.

Algorithm 1 Random Forest for Classification or Regression

```
For $b = 1$ to $B$:
    (a) Draw a bootstrap sample of size $n$ from the training data;
    (b) Grow a random-forest tree $T_b$ to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size $n_{\min}$ is reached.
        i. Select $m$ variables at random from the $p$ variables;
        ii. Pick the best variable/split-point among the $m$ selected variables;
        iii. Split the node into two daughter nodes.
Output the ensemble of trees $\{T_b\}_{b=1}^{B}$.
To make a prediction at a new point $\mathbf{x}$:
    • Regression: $\hat{f}_{\mathrm{rf}}^{B}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x})$;
    • Classification: Let $\widehat{C}_b$ be the class prediction of the $b$-th random-forest tree. Then $\widehat{C}_{\mathrm{rf}}^{B}(\mathbf{x})$ is the majority vote of $\{\widehat{C}_b(\mathbf{x})\}_{b=1}^{B}$.
```

Remark 1. When growing a tree on a bootstrapped dataset, before *each split*, select $m \le p$ of the input variables at random as candidates for splitting. Typical choice for m is $\sqrt{p}$ or as low as 1.

Remark 2. Intuitively, reducing m will reduce the correlation between any pair of trees in the ensemble, and, hence, by (1) reduce the variance of the average.

5. **On the Mean, Variance and Correlation of Bootstrap Estimates:** Suppose $X_1, X_2, \cdots, X_n$ are i.i.d with mean μ and variance σ^2. Let $\bar{X}_1^*$ and $\bar{X}_2^*$ be two bootstrap realizations of the sample mean. We show that the sampling correlation is

$$\mathrm{Cor}(\bar{X}_1^*, \bar{X}_2^*) = \frac{n}{2n-1}.$$

Proof. Let X^* be a generic bootstrap sample, and we have the following distribution table:

x^*	$\mathbb{P}(X^* = x^*)$
X_1	$1/n$
X_2	$1/n$
$\cdots$	$\cdots$
X_n	$1/n$

Note that, *conditional on* $X_1, \cdots, X_n$, bootstrap samples $X_1^*, \cdots, X_n^*$ are i.i.d. We first show that

$$\mathbb{E}[X^* \,|\, X_1, \cdots, X_n] = \frac{1}{n}\sum_{i=1}^n X_i,$$

and then $\mathbb{E}[X^*] = \mu$. Note that, conditional on $X_1, \cdots, X_n$,

$$\mathbb{E}[X^* \,|\, X_1, \cdots, X_n] = \sum_{i=1}^n X_i \cdot \mathbb{P}(X^* = X_i) = \frac{1}{n}\sum_{i=1}^n X_i,$$

therefore,

$$\mathbb{E}[X^*] = \mathbb{E}\big[\mathbb{E}[X^* \mid X_1, \cdots, X_n]\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \mu.$$

As a consequence, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^*\right] = \mu.$$

Next, we show

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i^*\right] = \frac{\sigma^2}{n}\left(2 - \frac{1}{n}\right).$$

By the law of total variance

$$\mathrm{Var}[X^*] = \mathbb{E}\big[\mathrm{Var}[X^* \,|\, X_1, \cdots, X_n]\big] + \mathrm{Var}\big[\mathbb{E}[X^* \,|\, X_1, \cdots, X_n]\big],$$

we have

$$\begin{aligned}
\mathrm{Var}[X^* \,|\, X_1, \cdots, X_n] &= \mathbb{E}[(X^*)^2 \,|\, X_1, \cdots, X_n] - \big(\mathbb{E}[X^* \,|\, X_1, \cdots, X_n]\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2, \\
\mathbb{E}\big[\mathrm{Var}[X^* \mid X_1, \cdots, X_n]\big] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2\right] \\
&= \left(1 - \frac{1}{n}\right)\sigma^2,
\end{aligned}$$

and

$$\mathrm{Var}\big[\mathbb{E}[X^* \mid X_1, \cdots, X_n]\big] = \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{\sigma^2}{n}.$$

Therefore,

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^*\right] &= \mathbb{E}\left[\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^* \,\middle|\, X_1, \cdots, X_n\right]\right] + \mathrm{Var}\left[\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^* \,\middle|\, X_1, \cdots, X_n\right]\right] \\ &= \frac{\sigma^2}{n}\left(1 - \frac{1}{n}\right) + \frac{\sigma^2}{n} \\ &= \frac{\sigma^2}{n}\left(2 - \frac{1}{n}\right).\end{aligned}$$

Finally, we show that

$$\mathrm{Cov}(\bar{X}_1^*, \bar{X}_2^*) = \frac{\sigma^2}{n}.$$

By the definition $\mathrm{Cov}(\bar{X}_1^*, \bar{X}_2^*) = \mathbb{E}[\bar{X}_1^* \cdot \bar{X}_2^*] - \mathbb{E}[\bar{X}_1^*] \cdot \mathbb{E}[\bar{X}_2^*]$, we only need to find $\mathbb{E}[\bar{X}_1^* \cdot \bar{X}_2^*]$ and note that $\mathbb{E}[\bar{X}_1^*] = \mathbb{E}[\bar{X}_2^*] = \mu$. Note the following

$$\begin{aligned}\mathbb{E}\big[\bar{X}_1^* \cdot \bar{X}_2^*\big] &= \mathbb{E}\Big[\mathbb{E}\big[\bar{X}_1^* \cdot \bar{X}_2^* \mid X_1, \cdots, X_n\big]\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[\bar{X}_1^* \mid X_1, \cdots, X_n\big] \cdot \mathbb{E}\big[\bar{X}_2^* \mid X_1, \cdots, X_n\big]\Big] \\ &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] \\ &= \mu^2 + \frac{\sigma^2}{n}.\end{aligned}$$

It follows that $\mathrm{Cov}(\bar{X}_1^*, \bar{X}_2^*) = \frac{\sigma^2}{n}$. Therefore,

$$\mathrm{Cor}(\bar{X}_1^*, \bar{X}_2^*) = \frac{\mathrm{Cov}(\bar{X}_1^*, \bar{X}_2^*)}{\sqrt{\mathrm{Var}[\bar{X}_1^*]\,\mathrm{Var}[\bar{X}_2^*]}} = \frac{\sigma^2/n}{(\sigma^2/n)(2-1/n)} = \frac{n}{2n-1}.$$

■

Remark. Not all estimators can be improved by shaking up the data like this. It seems that highly nonlinear estimators, such as trees, benefit the most.

On the other hand, bagging does *not* change linear estimates, such as the sample mean; the pairwise correlation between bootstrapped means is about 50%, as indicated by the results above.

II. Details of Random Forests

1. Some Recommendations on the Choice of the Value of m:

(a) For *classification*, the default value for m is $\lfloor\sqrt{p}\rfloor$ and the minimum node size is one;

(b) For *regression*, the default value for m is $\lfloor p/3 \rfloor$ and the minimum node size is five.

2. **Out-of-Bag (OOB) Estimates:** For each observation $\mathbf{z}_i := (\mathbf{x}_i, y_i)$, construct its random forest prediction by averaging *only* those trees corresponding to bootstrap samples in which $\mathbf{z}_i$ did *not* appear.
   We can fit random forests in one sequence, with cross-validation being performed along the way. Once the OOB error stabilizes, the training can be terminated.

3. **Variable Importance:**

   (a) At each split in each tree, the *improvement* in the split-criterion is the importance measure attributed to the splitting variable, and is accumulated over all the trees in the forest separately for each variable.

   (b) Comparing to boosting, the candidate split-variable selection *increases* the chance that any single variable gets included in a random forest. Boosting ignores some variables completely, but the random forest does *not*.

4. **Variable Prediction Importance:** Random forests also use the *OOB samples* to construct a different variable importance measure in order to measure the *prediction strength* of each variable:

   (a) For a single tree: let $T_b(\,\cdot\,; \Theta_b)$ be the tree constructed using the b-th bootstrapped sample.

      i. Pass OOB samples down to $T_b(\,\cdot\,; \Theta_b)$, compute the OOB error (which depends on the regression or classification problem at hand and the metric used), denoted by $\mathrm{PE}_b(\mathrm{OOB})$;

      ii. Randomly permute the OOB values on the j-th variable while leaving the data on all other variables unchanged;

      iii. Pass OOB samples down to $T_b(\,\cdot\,; \Theta_b)$, compute the OOB error, denoted by $\mathrm{PE}_b(\mathrm{OOB}_j)$, which should be larger than the error computed from the unaltered data;

      iv. Compute the difference between these two OOB errors

      $$S_{b,j} := \mathrm{PE}_b(\mathrm{OOB}_j) - \mathrm{PE}_b(\mathrm{OOB});$$

   (b) The final (predictive) importance of variable j in the random forest is computed by averaging the importances of variable j from all B trees

   $$\mathrm{Im}_j = \frac{1}{B} \sum_{b=1}^{B} S_{b,j}.$$

   Remark 1. The rational of the procedure above is that, if X_j is important, permuting its observed values will reduce our ability to classify successfully each of the OOB observations.

   Remark 2. The value of $\mathrm{PE}_b(\mathrm{OOB}_j)$ should be larger than that of $\mathrm{PE}_b(\mathrm{OOB})$, so that $S_{b,j} \geq 0$.

5. **Proximities:** One can use the random forest to compute the proximities between pairs of observations, which can then be used in imputing missing values and identifying multivariate outliers.

   (a) *Goal:* We want to define a similarity measure $\mathrm{prox}(\,\cdot\,,\,\cdot\,)$ between pairs of observations, say $\mathbf{x}_i$ and $\mathbf{x}_j$, so that the closer $\mathbf{x}_i$ and $\mathbf{x}_j$ are to each other, the larger the value of $\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_j)$ is.

   (b) *Computation of* $\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_j)$*:*

      i. Set `counter` $= 0$;

      ii. Pass down the observations $\mathbf{x}_i$ and $\mathbf{x}_j$ to the b-th tree in the random forest, $T(\,\cdot\,; \Theta_b)$;

      iii. If the observations $\mathbf{x}_i$ and $\mathbf{x}_j$ end up at the same terminal node in $T(\,\cdot\,; \Theta_b)$, we increase `counter` by one;

      iv. We repeat the preceding two steps over all B trees in the random forest, and then divide the frequency totals of pairwise proximities by the number, B, of trees in the forest; this gives us the proportion of all trees for which each pair of observations end up at the same terminal nodes.

   Remark. Note that each tree is *unpruned* and typically is terminated to further grow by certain criteria. Hence, each terminal node in the tree will contain a few observations.

   (c) *Computation of Dissimilarity:* The dissimilarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be obtained by subtracting $\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_j)$ from 1, i.e.,

   $$\delta_{i,j} := 1 - \mathrm{prox}(\mathbf{x}_i, \mathbf{x}_j), \qquad \text{for all } i, j = 1, 2, \cdots, n.$$

   We collect these pairwise dissimilarities into an $n \times n$ proximity matrix $\Delta := (\delta_{i,j})_{i,j=1,\cdots,n}$, which is symmetric, positive-definite, with diagonal entries equal to zero.

6. **Identifying Multivariate Outliers:** The proximity matrix constructed above can be used to identify multivariate outliers. We focus on the classification problem.

   (a) *Main Idea:* We identify an outlier by how far away it is from all other observations belonging to *its class* in the learning set.

   Suppose $\mathbf{x}_i$ and $\mathbf{x}_j$ both belong to Class w. Then, they are far apart from each other if and only if $\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_j)$ is small. If $\mathbf{x}_i$ is far away from *all* the other observations in Class w in the training set, then all the proximities, $\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_\ell)$, of $\mathbf{x}_i$ with $\mathbf{x}_\ell$, where $i \neq \ell$, will be small.

   (b) *Raw Outlier Measurement:* A *raw outlier measure* for the i-th observation, $\mathbf{x}_i$, in Class w is given by

   $$u_{i,w} := \frac{1}{\frac{1}{n} \sum_{\{\ell | \mathbf{x}_\ell \text{ belongs to Class } w, i \neq \ell\}} [\mathrm{prox}(\mathbf{x}_i, \mathbf{x}_\ell)]^2}, \qquad \text{for all } i = 1, \cdots, n, \quad (2)$$

   where $w = 1, 2, \cdots, W$.

Thus, if $\mathbf{x}_i$ is really an outlier for Class w, the denominator of (2) will be small, so that $u_{i,w}$ will be large.

(c) *Standardized Outlier Measurement:* Let

$$m_w := \text{median}_{\{\ell \,|\, \mathbf{x}_\ell \text{ belongs to Class } w\}} u_{\ell,w}$$

be the median of the raw outlier measures over all the observations in Class w. Then, for $w = 1, 2, \cdots, W$, a standardized version of $u_{i,w}$ is given by

$$\tilde{u}_{i,w} := \frac{u_{i,w} - m_w}{\sum_{\{\ell \,|\, \mathbf{x}_\ell \text{ belongs to Class } w\}} |u_{\ell,w} - m_w|}, \qquad \text{for all } i = 1, 2, \cdots, n. \tag{3}$$

Values of (3) in excess of 10 can be considered as a sign of outlier.

7. **Issue of Small Proportion of Relevant Variables:** When the number of variables is large, but the fraction of relevant variables is small, random forests are likely to perform poorly with small m. This is because at each split the chance can be small that the relevant variables will be selected.

8. **Issue of Overfitting:** Increasing the number of trees B does *not* cause the random forest to overfit. However, the average of fully grown trees can result in too rich a model and leads to overfitting and incur unnecessary variance.

III. Analysis of Random Forests

1. **Predictor in Regression Using Random Forest:** With B trees $\{T(\,\cdot\,; \Theta_b)\}_{b=1}^B$ grown, the random forest predictor is

$$\hat{f}_{\text{rf}}^B(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B T(\mathbf{x}; \Theta_b).$$

As $B \to \infty$, the random forest regression estimator is

$$\hat{f}_{\text{rf}}(\mathbf{x}) = \mathbb{E}_{\Theta|\mathbf{Z}}[T(\mathbf{x}; \Theta(\mathbf{Z}))],$$

which depends on the training data $\mathbf{Z}$.

2. **Variance of Limiting Random Forest Regression Estimator:** By (1), we have

$$\text{Var}[\hat{f}_{\text{rf}}(\mathbf{x})] = \rho(\mathbf{x})\sigma^2(\mathbf{x}),$$

where

- $\rho(\mathbf{x}) := \text{Cor}(T(\mathbf{x}; \Theta_1(\mathbf{Z})), T(\mathbf{x}; \Theta_2(\mathbf{Z})))$ is the sampling correlation between any pair of trees used in the averaging, and $\Theta_1(\mathbf{Z})$ and $\Theta_2(\mathbf{Z})$ are a randomly drawn pair of random forest trees grown to the randomly sampled $\mathbf{Z}$;

- $\sigma^2(\mathbf{x}) := \mathrm{Var}[T(\mathbf{x};\Theta(\mathbf{Z}))]$ is the sampling variance of any *single* randomly drawn tree.

3. **More on $\rho(\mathbf{x})$:** $\rho(\mathbf{x})$ is the theoretical correlation between a pair of random-forest trees evaluated at $\mathbf{x}$, induced by repeatedly making training sample draws $\mathbf{Z}$ from the population, and then drawing a pair of random forest trees. This is the correlation induced by the sampling distribution of $\mathbf{Z}$ and Θ.

4. **Variability in $\rho(\mathbf{x})$ and $\sigma^2(\mathbf{x})$:** The variability in $\rho(\mathbf{x})$ and $\sigma^2(\mathbf{x})$ is both

   - conditional on $\mathbf{Z}$: due to the *bootstrap sampling* and *feature sampling* at each split, and
   - a result of the sampling variability of $\mathbf{Z}$ itself.

5. **Decomposition of Variance:** Note that

$$\begin{aligned}\mathrm{Var}_{\Theta,\mathbf{Z}}[T(\mathbf{x};\Theta(\mathbf{Z}))] &= \mathrm{Var}_{\mathbf{Z}}[\mathbb{E}_{\Theta|\mathbf{Z}}[T(\mathbf{x};\Theta(\mathbf{Z}))]] &+& \quad \mathbb{E}_{\mathbf{Z}}[\mathrm{Var}_{\Theta|\mathbf{Z}}[T(\mathbf{x};\Theta(\mathbf{Z}))]] \\ \text{Total Variance} &= \mathrm{Var}_{\mathbf{Z}}[\hat{f}_{\mathrm{rf}}(\mathbf{x})] &+& \quad \text{within-}\mathbf{Z}\text{ Variance.}\end{aligned}$$

   - The first term is the sampling variance of the random forest ensemble, and decreases as m decreases;
   - The second term is the *within-$\mathbf{Z}$ variance*, and is a result of the randomization, and increases as m decreases.

6. **Bias in Random Forest:** The bias in random forest is the same as the bias of any individual samples trees $T(\,\cdot\,;\Theta(\mathbf{Z}))$, and

$$\begin{aligned}\mathrm{Bias}(\mathbf{x}) &= \mu(\mathbf{x}) - \mathbb{E}_{\mathbf{Z}}[\hat{f}_{\mathrm{rf}}(\mathbf{x})] \\ &= \mu(\mathbf{x}) - \mathbb{E}_{\mathbf{Z}}\big[\mathbb{E}_{\Theta|\mathbf{Z}}[T(\mathbf{x};\Theta(\mathbf{Z}))]\big].\end{aligned}$$

   It is typically *greater* than the bias of an unpruned tree grown to $\mathbf{Z}$. The improvements in prediction of bagging or random forest are solely a result of *variance reduction*.

   Typically, as m increases, the bias decreases.

7. **Similarity between Random Forest and Ridge Regression:**

   - In *ridge regression*, when one has a large number of variables with similarly sized coefficients, ridge regression shrinks their coefficients toward zero, and those of strongly correlated variables toward each other. This regularization via ridge stabilizes the model.
   - In *random forests*, with small m, each of the relevant variables get their turn to be the primary split, and the ensemble averaging reduces the contribution of any individual variable.

8. **Connection between Random Forest and k-Nearest Neighbors:** When each tree is grown to maximal size, for a particular Θ^*, $T(\mathbf{x};\Theta^*(\mathbf{Z}))$ is the response value of one of the training samples. The averaging process assigns weights to these training responses. Via the random-forest voting mechanism, those observations close to the target point get assigned weights to produce the final prediction.

References

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Izenman, Alan J (Mar. 2009). *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. en. Springer Science & Business Media.