MATH 748: Weekly Report

Nov 15-21, 2021

Nayeong Kim (nkim10@sfsu.edu)

Review

Tree method is segmenting the feature space into smaller pieces having similar features.

14 Classification and Regression Trees

14.6 Classification Trees

(1) Criterion 1:
The classification error is *not sufficiently sensitive* for tree-growing. Hence there are more criterion.

(2) Criterion 2: Gini Index
The Gini index is defined as below:

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}).$$

It is related to the concept *purity*. High purity means the samples are distributed in a smaller number of classes. For example, if all the nodes are in *class 1*, the purity is very high. (And G is small for larger purity.) The value of G is in $[0, \frac{K-1}{K}]$. G has smaller value when p is either close to 1 or 0. In practice, Gini Index is preferred and the classification error is more sensitive to the change of the probabilities.

(3) Criterion 3: Cross-Entropy:
The definition of cross-entropy is

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} log(\hat{p}_{mk}).$$

It is also related to node purity: more sensitive to classification error. It is quite similar to Gini-index numerically. I.e. the two functions in the range $[0, 1]$, the graph shows similar plots.

(4) Summary of the Measures of 3 Criterion

- Misclassification Error: $1 - max(p, 1-p)$
- Gini-index: $2p * (1-p)$
- Cross-Entropy: $-p * \log p - (1-p) * \log(1-p)$

14.7 Build and Prune: Regression and Classification Trees

(1) Regression Tree

i. Build: RSS

ii. Prune: RSS + Penalty Term ($\alpha|T|$, α can be decided using CV)

(2) Classification Tree

i. Build: Gini-index or Cross-entropy

ii. Prune: Misclassification Rate + Penalty Term ($\alpha|T|$)

14.8 Advantages and Limitations

(1) Trees vs Linear Models
The performances can differ for different cases. If the sample space has a simple planner boundary, the linear model can perform better. But the trees shows more flexible results by segmenting the space into more pieces.

(2) Advantages
It is more intuitive than linear models. (Some people think that it is more close to human decision.) It is a kind of automatic step-wise variable selection. Also, it is robust to outliers: which is far from other samples and may cause a wrong bias to the model.

(3) Limitations
It doesn't have the same level of predictive accuracy as previous models we covered. It is too sensitive to the change in data. Small change in data can cause a very different series of splits. Since the tree has a hierarchical structure, the error in the beginning steps can cause a large error.

15 Bagging and Boosting

15.1 Introduction

The motivation of this can come from the problems of decision trees. In decision trees, the variance can be high. A little change in the training set can cause a huge difference in the result. To solve this problem, *bagging*(**b**ootstrap **agg**regat**ing**) is proposed.
The main idea of bagging is *Averaging and Bootstrapping*. Recalling that the average of n samples has $\frac{1}{n}$ of the variance of the original distribution. ($\bar{Z} = \frac{1}{n}(Z_1 + \cdots + Z_n)$ has the variance $\frac{1}{n}\sigma^2$ where Z_js are n i.i.d. variables having the variance σ^2.)

15.2 How Bagging works

(1) Bootstrapping
The idea is pretending the samples as the total population. It is a resampling of the observed data set (allowing repetition) from the original dataset.

$$\text{Population} \overset{sample}{\longrightarrow} \text{Sample} \overset{sample}{\longrightarrow} B \text{ Bootstrapped Datasets}$$

Since the bootstrapped data set is from the sample space, not from the population space, it can differ from the distribution of the average of some samples. However, with several repetitions of bootstrapping, we can get a confidence interval.

(2) Averaging
With the B different bootstrapped datasets, train the statistical learning method and obtain the prediction.
For regression case, we can average all predictions. On the other hand, for classification case, we can get the most popular prediction or average the probabilities (similar to regression case) if the classifier produces probabilities before it classifies the given sample.

(3) Remarks
Using this method, we don't prune. Recall that a single big tree has a small bias and a large variance (due to more flexible boundaries). With pruned ones, it has larger bias and smaller variance than before pruning. In this method, we can lower the variance but can't lower the bias. We want lower bias in this case, too. Hence pruning is not applied when we use bagging.

15.3 Out-of-Bag Observation

It has a very natual way to estimate the test error of a bagged model without CV. For a tree from a bootstrapped data set, there are some non-selected samples. It can provide the test error. Since $P[Z_1 \neq X_1, \cdots Z_n \neq X_1] = (\frac{n-1}{n})^n \simeq \frac{1}{e}$, we can use about $\frac{1}{e}$ of the observations for testing. It is referred to as the out-of-bag(OOB) observations.
OOB error is the overall OOB MSE or OOB classification error. It is convenient when performing bagging on large data sets(which guarantee a portion of observations not used for fitting model and which CV would be computationally expensive).

15.4 Variable Importance Measure

Bagging improves the accuracy over the prediction but it is hard to interpret the model. Hence variable important plot can help interpretation. Variable important plot shows scores of variables which represents the importance of the variable.
For bagged regression trees, the average in B trees of the total amount of the decrease of the RSS due to splits over predictor can be recorded. The higher value implies the more importance. For bagged classification trees, the average in B trees of the total amount of the decrease of the Gini-index due to splits over a given predictor can be uses as a measure.

15.5 Random Forest

(1) Random Forest
The idea comes from bagging. The further idea of Random Forest(RF) is that it de-correlates the trees. Highly correlated trees have similar structures. RF overcome this by forcing each split to consider only a subset of the predictors (size m).

(2) Procedure of RF

i. Grow a forest of B trees.

ii. Grow each tree on an independent bootstrap samples which have only m features of p features. Typically choose $m \simeq \sqrt{p}$.

15.6 Boosting

Like bagging, boosting is also a general approach to various statistical learning methods. Boosting works similarly to bagging but it has a difference that it grows trees sequentially. Each tree is grown using information from previously grown trees. It is for correcting the problem that the mistake in the nodes near the root can cause a huge error.