Data Analysis and Machine-Learning

Chapter 9:

Random Forest



*Alone we can do so little;*

*Together we can do so much.*

*- Helen Keller*

1. Introduction

As illustrated in chapter 8, decision tree algorithm is well-known for its model accuracy and prediction performance. However, there are few critical limitations which calls for an alternative solution.

First, the peculiar characteristic of decision trees, namely the hierarchical structure starting from root nodes to terminal nodes result in venerability to errors in the intermediate stage. Since all nodes are connected to the root node while trickling down to the terminal node, errors in the intermediate stage is emitted to the next level, without corrections. Second, in a related context, individual decision tree model is affected sensitively by even few numbers of noises and minor fluctuations of learning data. Third and finally, as I have illustrated earlier, overfitting possibilities increase accordingly to the increasing number of depth due to smaller samples being included for each subset region.

These problems call for an alternative solution, namely, the ***Random Forest*** Method, which as implied from its name per se uses multiple number of individual tree models to generate a summative (i.e., ensemble) model.

2. Random Forest Model

Random forest model can be classified as one type of methodology from ***Ensemble***, which is a method of boosting prediction accuracy via taking majority voting or average calculation on multiple ***base model*** results, as follows:

Base1(x)=1

Base2(x)=0

Base3(x)=1

Base4(x)=1

Ensemble(x)=1

In order for the resulting ensemble model to have better prediction performance than the base models, two prerequisite conditions should be satisfied: (1) base models should be independent to each other. (2) each base model should have at least better performance than the random prediction model (e.g., random model’s error rates of 0.5 for binary classification problems). In other words when the error rate of the ensemble model is defined as:

When e is defined as the error rates of the base model and,

N is defined as the number of base models,

Ensemble model’s error rate is minimized at the point where consisting base models’ error rates are lower than the random model. Thus, for instance in binary classifications, base models with error rates lower than 0.5 (random model’s error rate) should be selected to construct an ensemble model.

Decision tree models covered in chapter 8 are particularly known to be useful as the base model for an ensemble model, and the reason is twofold. First, decision tree models have low computation al complexity, which enables constructing a model rapidly even in cases for enormous data sizes. Second, decision tree models are non-parametric, which means no prerequisites regarding the distribution of data (e.g., normal distribution) are required in prior to generating a model. In this context, Random Forest model refers to the particular type of ensemble model that utilizes decision tree model as the base model. Generally, random forest model depicts higher prediction accuracy than the individual tree model, and is widely used as the feature selection method for high-dimensional data.

The general sequence of random forest model is as follows. First, multiple numbers of training data are generated via ***bootstrapping*** technique, which we will cover in depth soon. Second, decision tree models are constructed based on the generated training data (using random variables). Third and finally, prediction results are ensembled together to generate a final prediction result.

Two core ideas are attached to the concept of random forest model: **diversity** and **randomness**. First, multiple numbers of training data are generated, and individual decision tree models are constructed for each dataset. This process is referred to as **bagging (Bootstrap Aggregating)**. Second, in the process of constructing a decision tree model, variables are selected randomly. This process is referred to as **random-subspacing**.

3. Bootstrapping

Bagging refers to the process of aggregating each model generated from bootstrap samples. In order to understand the concept of bagging, therefore, prior understanding on bootstrap is necessary. Bootstrapping refers to the sampling technique of generating as many data as the original data by ***sampling with replacement***. Each individual dataset generated under this procedure is called bootstrap set. Consider following illustration for better understanding.

In the illustration above, first notice that each bootstrap set has an **equal** **total** **number** **of data** to the original dataset. In bootstrap set 1, x2 is selected twice, while x5 is not selected. In bootstrap set 2, x3 is selected twice and x4 is not selected. This happens due to **sampling with replacement**. Each datum can be either selected multiple times, or not selected at all. Theoretically, the probability of a certain datum not being selected at all for a bootstrap can be calculated as:

which is around 0.367 (36.7%).

4. Bootstrap Aggregating (Bagging)

Now, suppose that we acquired 20 bootstrap sets for a binary classification problem (0 or 1). Using these sets, we can create 20 individual models, with P(Y=1) for each test instance, and final predicted class label (which will be either 0 or 1). When aggregating the results (bagging) for a classification problem, the ensemble model takes majority voting on the predicted class labels, as follows:

For instance, if N is 20, and we got:

Alternatively, it would be possible to take a weighted voting method instead of simple majority voting to incorporate the training accuracies of each individual model, as follows:

Or, it is possible to focus on the predicted probability of the test instances (P(y=1)), which would be:

5. Random Subspace

Now that we have satisfied diversity condition via bagging process, next is to meet the requirements for randomness condition. In order for individual decision tree model to be ***independent*** to each other, the generating process of each model should be based on randomness, which can be done via random subspacing.

Throughout bootstrapping, we have acquired N number of training data from the original training data; and for each training data, suppose we have generated N number of individual tree models. Normally in order to make a decision tree model, we find dividing variables and dividing points by searching all variables (X1~Xn).

Instead in random subspacing, we randomly select limited number of input variables among the total number of input variables to construct a model, and select a dividing variable and point among that limited pool of input variables. The procedure of selecting a dividing variable and point per se is the same with the procedures for making an individual tree model. Full-grown tree can be generated by repeating this process trickling down to each node.

The core point is that we consider only the **subset of variables** for each step instead of considering all variables. In other words, we randomly select smaller number of subset variables from the total number of variables in terms of exploring branch point of the decision tree model, and only consider those variables in terms of making the model.

6. Generalization Error

The two prerequisites, i.e., diversity via bagging and randomness via random subspaces, work as the very legitimacy of random forest model. Although each individual tree can be vulnerable to overfitting problems, strong **law of large numbers** achieved from the process of satisfying the two preconditions prevents overfitting, and the errors converge into limiting values. This is to say, generalization error of a random forest model can be defined as:

where p refers to the average correlation between each decision tree and s refers to the average of gaps between the number of correctly predicted trees and wrongly predicted trees.

As p shows the correlation rate between each individual decision tree, it can be inferred from the equation that the lower value of p (which means that each tree is independent to each other) would be better. Also, s would increase accordingly to the individual tree’s prediction accuracy, which means larger s and thus smaller [(1-s\*\*2)/s\*\*2] would be better.

In sum, two insights can be extracted from the generalization error equation. In order to minimize the error of a random forest model and to maximize model performance, we should either (1) minimize the value of p by minimizing the correlation rate between each individual tree model, and/or (2) maximize the value of s by concentrating on the prediction accuracies of each individual model.

7. Feature Selection

Looking back on linear regression or logistic regression that were covered in the earlier chapters, each model supposes a certain type of probability distribution (e.g., normal distribution for regression models, and Bernoulli distribution for logistic regression). Thus it is possible to get an information regarding the statistical significance of each individual variable, which comes in the form of coefficients or weights. Random forest model, on the other hand, does not provide specific information regarding the statistical significance of a variable, which is because, as we covered earlier, the model is basically non-parametric (i.e., the model does not suppose a certain probability distribution).

Thus, the significance of a variable is determined indirectly using the concept of Out of Bag Errors (OOB Errors). Suppose that we want to figure out the significance of X(i) variable. Out of Bag data refer to the set of data that have not been included in the bootstrap sets. These data are then applied to the made tree models, from Tree(1), Tree(2), to Tree(t), to calculate OOB errors r(i), which can be denoted as:

since there are t number of generated tree models. OOB errors are then calculated on the data subset with randomly mixed values of X(i) variable, and thus we get:

In most cases, therefore, we can assume that e(i) will have larger values than r(i), and thus the difference of the two values can be denoted as:

And the average of such differences can be denoted as:

and the variance of differences can be denoted as:

Then, let us consider the meaning of ‘d’ that we just calculated. When the value of d (in other words, e – r) for a certain variable X(i) is large, this would mean that the error rates will increase greatly if that variable is replaced by other values, meaning that X(i) is significant by the extent of d. Thus, the variable significance of X(i) can be denoted as:

where ‘d’ becomes the main criteria for determining the significance of a variable and the variance S of d works as the penalty for scaling reasons.

8. Demonstrations

Now that we have covered the main concepts of the random forest model, let me demonstrate the procedures ***manually*** using samples. Rather than importing the random forest library directly, the following manual steps will help reviewing the procedures and principles fundamentally.

#Import Essential Libraries

import numpy as np

import matplotlib.pyplot as plt

from sklearn import datasets

from sklearn.tree import DecisionTreeClassifier

from collections import Counter

from sklearn.metrics import accuracy\_score

from sklearn.ensemble import BaggingClassifier, RandomForestClassifier

import sys

sys.path.append('C:\\Users\\Master\\Desktop\\dataTools')

import dataTools as dt

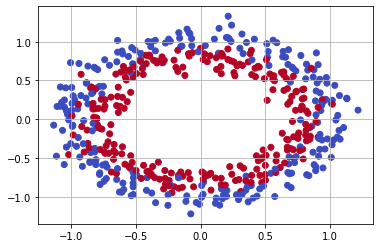
#Sample Generation

x, y = datasets.make\_circles(n\_samples=500, noise=0.1, random\_state=123)

plt.scatter(x[:, 0], x[:, 1], c=y, cmap='coolwarm')

plt.grid()

plt.show()



#Bootstrapping (Manually) - sampling with replacement

subsample = []

for i in range(11):

    index = [x for x in range(x.shape[0])]

    sampling = np.random.choice(index, x.shape[0]//2, replace=True)

    subsample.append(sampling)

subsample

OUTPUT:

[array([304, 186, 460, 123, 320, 388, 423, 328, 17, 248, 226, 22, 95,

386, 8, 447, 147, 72, 122, 338, 497, 430, 367, 8, 335, 309,

91, 71, 396, 159, 452, 127, 300, 3, 276, 161, 318, 189, 182,

230, 200, 421, 405, 394, 236, 89, 286, 326, 0, 189, 90, 370,

299, 154, 196, 89, 266, 417, 493, 338, 145, 34, 220, 51, 309,

390, 284, 114, 72, 352, 322, 192, 479, 283, 180, 181, 302, 405,

195, 280, 52, 265, 175, 57, 151, 101, 226, 308, 416, 118, 488,

364, 110, 373, 376, 280, 253, 336, 360, 207, 22, 182, 224, 462,

182, 267, 431, 252, 0, 83, 57, 4, 116, 131, 43, 494, 369,

251, 68, 341, 23, 136, 486, 468, 71, 205, 6, 474, 359, 81,

45, 455, 106, 212, 203, 152, 419, 124, 23, 334, 255, 62, 281,

221, 399, 490, 432, 93, 491, 428, 309, 157, 112, 346, 143, 302,

16, 151, 334, 140, 451, 317, 443, 340, 410, 302, 216, 99, 492,

369, 97, 98, 498, 173, 466, 295, 227, 382, 362, 223, 226, 231,

198, 450, 262, 236, 151, 52, 204, 27, 497, 153, 286, 416, 419,

113, 64, 14, 107, 89, 365, 48, 456, 422, 19, 226, 482, 332,

99, 462, 136, 389, 467, 438, 140, 29, 36, 358, 96, 407, 97,

376, 458, 38, 494, 372, 275, 395, 37, 105, 352, 155, 351, 309,

485, 298, 55, 60, 328, 327, 134, 495, 79, 300, 412, 483, 65,

382, 124, 154]),

array([161, 422, 278, 265, 439, 318, 73, 238, 109, 180, 413, 72, 175,

184, 165, 46, 454, 237, 129, 497, 487, 429, 87, 5, 255, 211,

116, 101, 297, 168, 485, 293, 1, 92, 402, 198, 155, 335, 378,

283, 249, 455, 1, 55, 386, 313, 471, 454, 355, 193, 47, 457,

168, 135, 295, 483, 83, 441, 409, 348, 14, 284, 93, 27, 300,

[show more (open the raw output data in a text editor) ...](vscode-file://vscode-app/c:/Users/Master/AppData/Local/Programs/Microsoft%20VS%20Code/resources/app/out/vs/code/electron-browser/workbench/workbench.html)

372, 401, 172, 418, 62, 55, 131, 168, 231, 357, 493, 64, 146, 239, 398, 47, 195, 156, 418, 165, 113, 4, 63, 201, 180, 219, 433, 89, 190, 143, 73, 298, 452, 86, 301, 452, 22, 260, 339, 216, 135, 80, 14, 488, 200, 268, 49, 379, 174, 36, 377, 133, 394, 119, 366, 471, 424, 4, 213, 386, 246, 477, 17, 57, 67, 263, 271, 398])]

#Bagging (manually)

bagging = []

for index in subsample:

    x\_train = x[index]

    y\_train = y[index]

    model = DecisionTreeClassifier().fit(x\_train, y\_train)

    bagging.append(model.predict(x))

display(len(bagging))

display(len(bagging[0]))

OUTPUT:

11

500

bagging\_prediction = []

for i in range(len(bagging[0])):

    total = []

    for j in bagging:

        total.append(j[i])

    cnt = Counter(total)

    bagging\_prediction.append(cnt.most\_common(1)[0][0])

print(bagging\_prediction)

OUTPUT:

[1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0]

#Evaluation

yhat = np.array(bagging\_prediction)

accuracy\_score(y, yhat)

OUTPUT:

0.934

#Decision Tree + Bagging

base\_model = DecisionTreeClassifier()

baggingmodel = BaggingClassifier(

    base\_estimator=base\_model,

    n\_estimators=11

)

baggingmodel.fit(x,y)

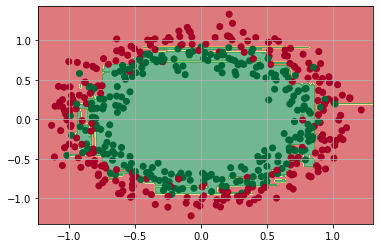
#Result Visualization

dt.dimensionchange(baggingmodel, x[:,0], x[:,1], cmap='RdYlGn', alpha=0.6)

plt.scatter(x[:, 0], x[:, 1], c=y, cmap='RdYlGn')

plt.grid()

plt.show()



#Random Forest (Library)

model = RandomForestClassifier(

    n\_estimators=1000,

    random\_state=12

)

model.fit(x,y)

dt.dimensionchange(model, x[:,0],x[:,1],cmap='RdYlGn',alpha=0.6)

plt.scatter(x[:,0],x[:,1],c=y, cmap='RdYlGn')

plt.grid()

plt.show()

