FLIPoo Final Assessment

Cong Ma

[1] QingDao Technological University, China

One-hot Coder

- One-hot coding is a process of transforming class variables into machine learning algorithms.

In my opinion,one-hot can process the data,and turn them into binary vector. If an attribute has n values, attributes. Only one of the N attributes of each sample can be 1, which means that the attribute of the sample belongs to this category, and the other extended attributes are 0.

One-hot Coder Shortcoming

- It doesn't work in *logistic regression*. Because logistic regression requires variables to be independent of each other. If you have only one attribute that needs one hot coding, if a sample are 1 at the same time, the two attributes will be completely related, which will inevitably lead to singular Error.
- That is, nonsingular matrix can't solve the unique solution and get the unique model, but you can't delete a certain one hot extension variable of the same attribute.
- For example, if we want to code ""China", "America", "Japan", "America"" one hot, how can we do it? First we let "China" = 0 , "America" = 1, "Japan" = 2;

Country	China	America	Japan
China	1	0	0
America	0	1	0
Japan	0	0	1

Now we can get the one-hot code as:
""China", "America", "Japan", "America"" = [[1,0,0], [0,1,0], [0,0,1], [0,1,0]]

Bagging Algorithm

Based on bootstrap sampling, bagging algorithm can be constructed. In this method, the train set is sampled several times by bootstrap, and a weak learner model is trained with the data set formed by each sampling, and several independent weak learners are obtained. Finally, the combination of thus to predict. The training process is as follows:

- Cycle, for I = 1,..., t.
- The training sample set is obtained by bootstrap sampling.
- A sample set of H is used to train the model.
- End cycle output model combination.

Ensemble learning and Bootstrap sampling

- Before we talk about the random forest,there are two ideas we need to know.

Ensemble learning combines multiple models to form a more accurate model. The model involved in the combination is called weak learner. These weak learner models are used to predict jointly

Bootstrap sampling is to take n samples back in the n samples set to form a data set. By the way,If the sample size is large, there is a 0.368 probability that each sample will not be selected in the whole sampling process.
But why? let's make a experience.

- If there are 10 samples, bootstrap sampling randomly takes 10 samples from them. The following two situations are possible:

1 1 1 1 1 1 1 1 1 1
1 2 3 4 5 6 7 8 9 10

- Suppose there are n samples in the sample set, and the probability of any one sample in each sampling is 1 / N, that is equal probability. The probability that a sample is not selected in each sampling is 1-1 / n.

$$\lim_{n\to\infty}(1-1/n)^n = 1/e = 0.368.$$

Random Forest

First, a random forest is composed of multiple decision trees. For the classification pro sample will be sent to each decision tree for prediction, and then vote. The class with the most votes is the final classification result. For regression problems, the prediction output of random forest is the mean value of all decision tree outputs.

That is to say, multiple random variables are added to get the mean value, and the variance will be reduced. If the output value of each decision tree is regarded as a random variable, the mean variance of the output value of multiple trees will be smaller than that of a single tree, so the variance of the model can be reduced.

The training samples of each tree are obtained by bootstrap sampling from the original training set. The features used in training each node of decision tree are also obtained by random sampling, that is, some features are randomly extracted from the feature vector to participate in the training.

- There is no standard answer to determine the number of decision trees and the number of features to be selected for each split. The first problem depends on the size of the training set and the characteristics of the problem. There is no exact theoretical answer to the second question, which can be determined by experiments.

TULIP *Team for Universal Learning and Intelligent Processing*

Thanks for watching 2020/10/16