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, iff 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.

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 sdfdsfdsafasfafais 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.

Bagging Algorithm

Based on bootstrap sampling, bagging algorithm can be constructed. In this method, the trainine 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 thused 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

Random Forest

First, a random forest is composed of multiple decision trees. For the classification problet 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 regressadasdasdassion problems, the prediction output of random forest is the mean value of all decision tree outputs.

TULIP *Team for Universal Learning and Intelligent Processing*

Thanks for watching!