MACHINE LEARNING

Random Forest

Trainer : Sujata Mohite
Email: sujata.mohite@sunbeaminfo.com

Overview

- Random forest classifier creates a **set of decision trees** from randomly selected subset of training set
- It then aggregates the votes from different decision trees to decide the final class of the test object
- This works well because a single decision tree may be prone to a noise, but aggregate of many decision trees reduce the effect of noise giving more accurate results

How does it work?

- Suppose training set is given as : [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4], random forest may create three decision trees taking input of subset for example,

  [X1, X2, X3] -> tree1
  [X1, X2, X4] -> tree2
  [X2, X3, X4] -> tree3

- So finally, it predicts based on the majority of votes from each of the decision trees made

Case Study

- A certain country has a population of 118 million
- Salary data is collected with following attributes
  - Age, Gender, Education, Residence, Industry
- Salary bands :
  - Band 1 : Below $40,000
  - Band 2: $40,000 – 150,000
  - Band 3: More than $150,000

Case Study

- Following are the outputs of the 5 different CART model.

DT1

	Salary Band	1	2	3
Age	Below 18	90%	10%	0%
	19-27	85%	14%	1%
	28-40	70%	23%	7%
	40-55	60%	35%	5%
	More than 55	70%	25%	5%

DT2

	Salary Band	1	2	3
Gender	Male	70%	27%	3%
	Female	75%	24%	1%

DT3

	Salary Band	1	2	3
Education	<=High School	85%	10%	5%
	Diploma	80%	14%	6%
	Bachelors	77%	23%	0%
	Master	62%	35%	3%

DT4

	Salary Band	1	2	3
Residence	Metro	70%	20%	10%
	Non-Metro	65%	20%	15%

DT5

	Salary Band	1	2	3
Industry	Finance	65%	30%	5%
	Manufacturing	60%	35%	5%
	Others	75%	20%	5%

Case Study

- Using these 5 CART models, we need to come up with singe set of probability to belong to each of the salary classes
- For simplicity, we will just take a mean of probabilities in this case study. Other than simple mean, we also consider vote method to come up with the final prediction.
- To come up with the final prediction let's locate the following profile in each CART model :
  - 1. Age : 35 years
  - 2. Gender : Male
  - 3. Highest Educational Qualification : Diploma holder
  - 4. Industry : Manufacturing
  - 5. Residence : Metro

Case Study

- For each of these CART model, following is the distribution across salary bands :

Age , Salary
Gender , Salary
Education , Salary
Industry , Salary
Residence , Salary

CART	Band	1	2	3
Age	28-40	70%	23%	7%
Gender	Male	70%	27%	3%
Education	Diploma	80%	14%	6%
Industry	Manufacturing	60%	35%	5%
Residence	Metro	70%	20%	10%
Final probability		70%	24%	6%

- The final probability is simply the average of the probability in the same salary bands in different CART models
- As you can see from this analysis, that there is 70% chance of this individual falling in class 1 (less than $40,000) and around 24% chance of the individual falling in class 2

Advantages of Random Forest algorithm

- For applications in classification problems, Random Forest algorithm will avoid the overfitting problem(tree is very complex →too many levels)
- For both classification and regression task, the same random forest algorithm can be used
- The Random Forest algorithm can be used for identifying the most important features from the training dataset, in other words, feature engineering.

Applications of Random Forest

- For the application in banking, Random Forest algorithm is used to find loyal customers, which means customers who can take out plenty of loans and pay interest to the bank properly, and fraud customers, which means customers who have bad records like failure to pay back a loan on time or have dangerous actions.
- For the application in medicine, Random Forest algorithm can be used to both identify the correct combination of components in medicine, and to identify diseases by analyzing the patient's medical records.
- For the application in the stock market, Random Forest algorithm can be used to identify a stock's behavior and the expected loss or profit.
- For the application in e-commerce, Random Forest algorithm can be used for predicting whether the customer will like the recommend products, based on the experience of similar customers.
- For all types of classification

Thank You!!