Implementing the Naive Bayes classifier in Mahout

Dhruva Gajjar
Student(Masters of Engineering)
Birla Vishwakarama Mahavidhyalaya, V.V.Nagar

Abstract **- The Mahout binaries contain ready-to-use scripts for using and understanding the classical Mahout Dataset. We will use this dataset for testing or coding. Using mahout Naive Bayes classification should be done. To implement this we also need Hadoop and java on machine. Mahout classification classify data into given set of category. Naïve Bayes is a probabilistic data mining classifier which fits nicely into the Map Reduce model and gives pretty good predictive performance for its simplicity. The Hadoop implementation uses a single map/reduce operation to calculate the mean and standard deviation of each attribute/class combination, as well as the global class distribution of the training dataset.**

Keywords **- Hadoop, JDK, Maven, Subversion, Mahout**

I. INTRODUCTION

Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware. Once big data is stored on the Hadoop Distributed File System (HDFS), Mahout provides the data science tools to automatically find meaningful patterns in those big data sets. The Apache Mahout project aims to make it faster and easier to turn big data into big information.

Hadoop

The Apache Hadoop framework is composed of the following modules:

1. *Hadoop Common* – contains libraries and utilities needed by other Hadoop modules
2. *Hadoop Distributed File System (HDFS)* – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster.
3. *Hadoop YARN* – a resource-management platform responsible for managing compute resources in clusters and using them for scheduling of users' applications.
4. *Hadoop MapReduce* – a programming model for large scale data processing.

The HDFS File System is an optimized file system for distributed processing of very large datasets on commodity hardware.The map reduces framework works in two main phases to process the data. Which are the Map phase and the Reduce phase.

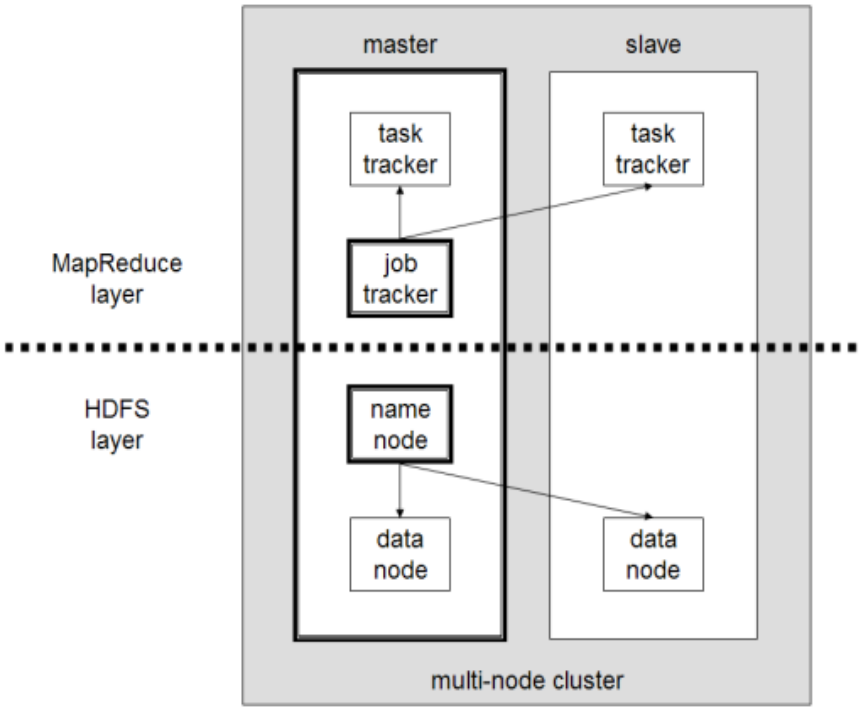

Fig 1

Maven

Maven is a build automation tool used primarily for Java projects. Maven addresses two aspects of building software: First, it describes how software is built, and second, it describes its dependencies.

Mahout

Mahout provides for types of processing:

1. *Collaborative filtering* – mines user behavior and makes product recommendations (e.g. Amazon recommendations)
2. *Clustering* – takes items in a particular class (such as web pages or newspaper articles) and organizes them into naturally occurring groups, such that items belonging to the same group are similar to each other
3. *Classification* – learns from existing categorizations and then assigns unclassified items to the best category
4. Frequent itemset mining – analyzes items in a group (e.g. items in a shopping cart or terms in a query session) and then identifies which items typically appear together.

Naive Bayes Classifier

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. Using Bayes' theorem, this can be written

$$p(C|F_1,\ldots,F_n) = \frac{p(C)\ p(F_1,\ldots,F_n|C)}{p(F_1,\ldots,F_n)}.$$

II. HOW IT WORKS

Berofe go for Mahout we ensure that Hadoop is already install on that machine.

Install Mahout (Hadoop in distributed mode on Ubuntu)

[1] Install Java RE 1.7

[2] Install Maven 3.0.4
sudo apt-get install maven
(To verify, mvn -version)

[3] Add a dedicated hadoop user
sudo addgroup hadoop
sudo adduser --ingroup hadoop hduser
(A prompt will be shown to set the password)

Set up passphraseless ssh
su - hduser
ssh-keygen -t dsa -P "" -f ~/.ssh/id_dsa
cat ~/.ssh/id_dsa.pub >> ~/.ssh/authorized_keys

(For CMU Kerberos users, the "kinit" would still pop-up to stop the ssh.
In ~/.cmuscs_settings (create if not exist), add "no_kinit=true".)

[4] Disable IPV6
/etc/sysctl.conf
net.ipv6.conf.all.disable_ipv6 = 1
net.ipv6.conf.default.disable_ipv6 = 1
net.ipv6.conf.lo.disable_ipv6 = 1

[5] Install Hadoop
Hadoop home folder: usr0/home/hduser/hadoop
sudo tar xzf hadoop-*.*.*.tar.gz
sudo mv hadoop-*.*.* hadoop
sudo chown -R hduser:hadoop hadoop

[6] Configurations

Update /usr0/home/hduser/.bashrc
export HADOOP_HOME=/usr/local/hadoop
export JAVA_HOME=/usr/lib/jvm/java-7-oracle

Update HADOOP_HOME/conf/hadoop-env.sh
export JAVA_HOME=/usr/lib/jvm/java-7-oracle

Create tmp directory for hadoop
sudo mkdir HADOOP_HOME/hadoop-tmp
sudo chown hduser:hadoop HADOOP_HOME/hadoop-tmp

[7] Configure Pseudo-Distributed Mode

Update HADOOP_HOME/conf/

conf/core-site.xml:
```
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/usr0/home/hduser/hadoop/hadoop-tmp</value>
</property>
    <property>
      <name>fs.default.name</name>
      <value>hdfs://localhost:9000</value>
    </property>
</configuration>
```

conf/hdfs-site.xml:
```
<configuration>
    <property>
      <name>dfs.replication</name>
      <value>1</value>
    </property>
</configuration>
```

conf/mapred-site.xml:
```
<configuration>
    <property>
      <name>mapred.job.tracker</name>
      <value>localhost:9001</value>
    </property>
</configuration>
```

[8] Format the HDFS
bin/hadoop namenode -format

[9] Start hadoop
bin/start-all.sh

[10] Run a test program
Find a large text file, say a.txt, and set up an output folder, say bfolder.
bin/hadoop dfs -copyFromLocal a.txt /user/hduser/testin/a.txt
bin/hadoop jar hadoop-examples-1.1.2.jar wordcount /user/hduser/testin /user/hduser/testout
bin/hadoop dfs -copyToLocal /user/hduser/testout/* bfolder

[11] Download mahout
svn co http://svn.apache.org/repos/asf/mahout/trunk
copy the mahout folder to hduser's own directories

[12] Set up environment variables (~/.bashrc)
$HADOOP_HOME
$HADOOP_CONF_DIR
$JAVA_HOME
$MAHOUT_HOME
run "source ~/.bashrc" to make them effective immediately.

Naive Bayes using Bayesian Probability terminology, the above equation can be written as

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}.$$

In practice, there is interest only in the numerator of that fraction, because the denominator does not depend on C and the values of the features F_i are given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$p(C, F_1, \ldots, F_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
p(C, F_1, \ldots, F_n) &= p(C)\ p(F_1, \ldots, F_n|C) \\
&= p(C)\ p(F_1|C)\ p(F_2, \ldots, F_n|C, F_1) \\
&= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3, \ldots, F_n|C, F_1, F_2) \\
&= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ p(F_3|C, F_1, F_2)\ p(F_4, \ldots, F_n|C, F_1, F_2, F_3) \\
&= p(C)\ p(F_1|C)\ p(F_2|C, F_1)\ \ldots p(F_n|C, F_1, F_2, F_3, \ldots, F_{n-1})
\end{aligned}$$

Now the "naive" conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$ given the category C. This means that

$$p(F_i|C, F_j) = p(F_i|C),$$
$$p(F_i|C, F_j, F_k) = p(F_i|C), p(F_i|C, F_j, F_k, F_l) = p(F_i|C),$$ and so on, for $i \neq j, k, l$.

Thus, the joint model can be expressed as

$$\begin{aligned}
p(C|F_1, \ldots, F_n) &\propto p(C, F_1, \ldots, F_n) \\
&\propto p(C)\ p(F_1|C)\ p(F_2|C)\ p(F_3|C)\ \cdots \\
&\propto p(C) \prod_{i=1}^{n} p(F_i|C).
\end{aligned}$$

$$\begin{aligned}
p(C|F_1, \ldots, F_n) &\propto p(C, F_1, \ldots, F_n) \\
&\propto p(C)\ p(F_1|C)\ p(F_2|C)\ p(F_3|C)\ \cdots \\
&\propto p(C) \prod_{i=1}^{n} p(F_i|C).
\end{aligned}$$

This means that under the above independence assumptions, the conditional distribution over the class variable C is:

$$p(C|F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

where the evidence $Z = p(F_1, \ldots, F_n)$ is a scaling factor dependent only on $F_1, \ldots, F_n$, that is, a constant if the values of the feature variables are known.

- Best described through an example

Three features

Class label (male or female)

sex	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Training dataset

- For each feature in each label
  - Compute the *mean* and *variance*

male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

sex	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033e-02	176.25	1.2292e+02	11.25	9.1667e-01
female	5.4175	9.7225e-02	132.5	5.5833e+02	7.5	1.6667e+00

That is the model (classifier)

Male or female?

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

- For each label ➔ Compute *posterior* value
- The label with the largest posterior is the suggested label

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\,p(height|female)\,p(weight|female)\,p(footsize|female)}{evidence}$$

Male or female?

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

$P(\text{female}) = 0.5$
$p(\text{height}|\text{female}) = 2.2346e-1$
$p(\text{weight}|\text{female}) = 1.6789e-2$
$p(\text{foot size}|\text{female}) = 2.8669e-1$
$\text{posterior numerator (female)} = \text{their product} = 5.3778e-04$

The sample is predicted to be female

Male or female?

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

$$posterior(female) = \frac{P(female)\,p(height|female)\,p(weight|female)\,p(footsize|female)}{evidence}$$

>> **evidence:** Can be ignored since it is the same constant for all labels

>> **P(label):** % of training points with this label

>> **p(feature|label) =** $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\text{f} - \mu)^2}{2\sigma^2}\right)$ eature value in sample

Male or female?

sex	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$posterior(male) = \frac{P(male)\,p(height|male)\,p(weight|male)\,p(footsize|male)}{evidence}$$

$$P(male) = 0.5$$

$$p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

$$p(\text{weight}|\text{male}) = 5.9881e-06$$

$$p(\text{foot size}|\text{male}) = 1.3112e-3$$

$$\text{posterior numerator (male)} = \text{their product} = 6.1984e-09$$

III. CONCLUSION

While 2012 has been the year of Big Data, 2013-14 is becoming the year of Big Data analytics. Gathering and maintaining large collections of data is one thing, but extracting useful information from these collections is even more challenging. We discussed in this paper some insights about the topic, and what we consider are the main concerns and the main challenges for the future. At the edge of statistics, computer science and emerging applications in industry, this research community focuses on the development of fast and efficient algorithms for real-time processing of data with as a main goal to deliver accurate predictions of various kinds. Machine learning techniques can solve such applications using a set of generic methods that differ from more traditional statistical techniques.

REFERENCES

[1] Maksudul Alam, S M Arifuzzaman, Md Hasanuzzaman Bhuiyan,*Text Classification using Mahout* ,2012.
[2] http://en.wikipedia.org/wiki/Naive_Bayes_classifier
[3] http://korolevbin.blogspot.in/2013/05/install-hadoop-in-pseudo-distributed.html
[4] http://hortonworks.com/hadoop/mahout/
[5] Jainendra Singh,Big Data Analytic and Mining with Machine Learning Algorithm