**Hadoop Implementation**

CS 675: Big Data Management and Analytics

Monroe College

Professor Syed Farid

**Group 1**

Gita Gurung

Madeline (Zhiyang) Lin

Priyanka Chaudhari

Sukriya Gurung

**Chapter 7: Machine Learning with Hadoop (12 pages) - Madeline**

This Chapter focuses on machine learning in the context of big data. The chapter presents an historical overview of the field, a classification of machine learning tasks, a motivation for using big data, and commonly used tools and platforms. Lastly, it discusses the future of machine learning and artificial intelligence.

Machine learning is a combination of statistics, computer science, and applied mathematics. Deep learning is based on neural networks. Deep learning algorithms consist of multiple layers of neural networks.

In terms of terminology, the chapter talks about terms including observation, feature, and target. Observation is a representation of some object or entity data such as an email message, a customer, a piece of equipment, etc. Feature is part of an observation. Observation is represented as a vector of features. Target is the attribute that the model predicts.

Task types in machine learning include supervised learning and unsupervised learning. Supervised learning is to learn an association between the inputs (features) and output (target variable) by using the examples provided. Unsupervised learning refers to a feature mix of observations without a target variable, and thus is often used for exploratory analysis or classification to gain insight or as a step before supervised learning. Unsupervised learning includes clustering, anomaly detection, recommender systems, and market basket analysis.

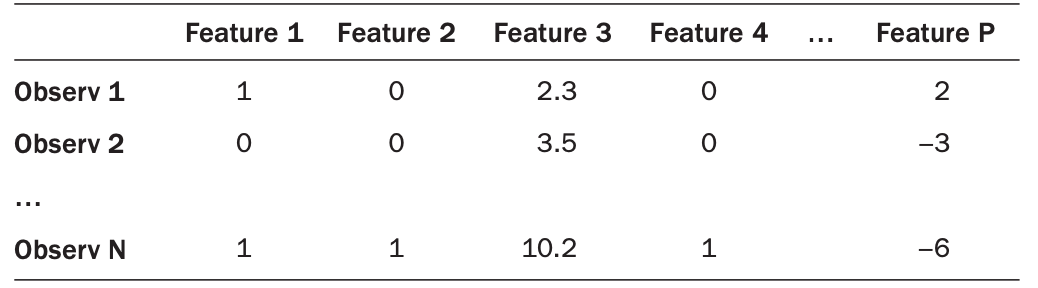
Data sets are growing faster than Moore's law. Scalable tools like Hadoop are one way to mitigate this issue that data comes from "everywhere". Big data with new and accurate features provides for better training of machine learning techniques. However, more instances do not always make machine-learning techniques perform better. The performance of a machine learning technique depends on the quality of the training set and how well the model generalizes to unseen data points.

**Applying Data Modeling with Hadoop - Machine Learning (26 pages) - Madeline (Machine Learning part till page 8)**

Machine learning is a result of successful research in fields such as statistics, computer science, and applied mathematics. Historically, machine learning developed as part of artificial intelligence (AI) in the late 1950s, with the goal of building machines that can mimic human minds.

In machine learning, some terminologies should be understood. They are observation, feature and target.

* Observation: Each observation is a representation, as data, of some object or entity such as an email message, a customer, a piece of equipment, etc.
* Feature: Each observation is represented as a vector of features (also called variables or attributes) of the observation.
* Target: Target represents our desire to predict the feature value.

Below is an illustration of the above.The quality of machine learning models is strongly dependent on the quality and size of data given to them. Tools for machine learning include R, Python, Java, and Vowpal Wabbit.

In terms of the future of AI and ML, since the inception of artificial intelligence as a research topic, it has always fascinated academics and practitioners alike, with promises of autonomous robots and machines that can talk and learn, think and behave like humans or even better than humans.

The era of big data is a new and exciting step in that direction. For the first time, technology platforms enable the rapid and efficient collection and storage of massive datasets and their utilization to drive research and application of machine learning at scale, with Hadoop at the center of it.

**Applying Data Modeling with Hadoop - Predictive Modeling (starting from page 9) (Sukriya)**

**Predictive Modeling**

Predictive Modeling is a statistical technique that uses machine learning and data mining to predict and forecast outcomes with the help of historical and existing data. It is used in every conceivable domain.

**Classification Versus Regression**

Classification is the task of predicting a discrete class label, while regression is the task of predicting a continuous quantity output. Let’s consider an example of a dataset that contains student information from a particular university. A regression model can be used to predict the height of any student based on other categories like weight, gender, diet, or subject major. On the other hand, classification can be used to analyze whether an email is spam or not. This kind of classification is called binary. When the classification is based on predefined categories, it is called a multi-class categorical classification.

**Evaluating Classifiers**

A confusion matrix is an NxN matrix. N is the number of classes or predicted categories. Each row reflects instances in the actual class, and each column represents instances in the predicted class. The data in the confusion matrix reflects the correct or incorrect classification decision by the model in the form of True positives, True Negatives, False Negatives and False positives. And the metrics of accuracy can be highlighted by accuracy, precision, recall and specificity.

**Evaluating Regression Models**

Validation and evaluation of a model provides more color to our hypothesis and helps evaluate different models that would provide better results against the data. There are three main errors (metrics) used to evaluate models, Mean absolute error, Mean Squared error and R2 score.

**Cross Validation**

Cross validation is a technique for evaluating machine learning models to ensure that the estimated accuracy metrics are robust and helps prevent overfitting of the model. K-Fold cross validation is the most popular method. It guarantees that the score of the model does not depend on the way the train and test sets are picked. The data set is divided into k numbers of subsets and the holdout method is repeated k number of times.

**Supervised Learning Algorithms**

Supervised learning algorithm consists of an outcome or dependent variable which is to be predicted from a given set of predictors or independent variables. Some of the supervised learning algorithms are highlighted below:

* K-Nearest neighbors: It is a non-parametric algorithm that classifies data points based on their proximity and association to other available data. It assumes that similar data points can be found near each other. It calculates the distance between data points, and then it assigns a category based on the most frequent category or average.
* Neural Networks: It processes training data by mimicking the interconnectivity of the human brain through layers of nodes. Each node is made up of inputs, weights, a bias, and an output.
* Generalized linear models: It is widely used, and recent extensions such as least absolute shrinkage and selection operator and elastic net make them very robust and effective.
* Decision trees: It is the hierarchical tree-structured algorithm that is used for deriving a meaningful output from a variety of inputs.
* Tree ensembles: Random forest is a powerful technique where a set of decision trees is constructed, each using a random subset of the training examples and features. Each tree is used to make a single decision for both classification and regression.

**Hadoop-Hive-Pig-MapReduce-ebook (22 pages)**

**Hadoop Installation & Download: (PRIYANKA)**

Hadoop installation requires java because it is built in java language that's why it requires all java libraries that are present in JDK and you will only get JDK when you install java.

**JAVA Versions:**

**Hadoop Configuration:**

**BASIC CLI Commands:**

**Running a Mapreduce Job:**

**The Hadoop ecosystem: (PRIYANKA)**

Hadoop Ecosystem is a part of Apache Hadoop, which is an open source, scalable and fault tolerant framework that processes and stores data which is collected from various sources. Hadoop is an open source framework which is written in Java. It is scalable(as it allows us to add more nodes on the fly) and a fault tolerant tool(even if one node goes down, data won’t stop processing rather is processed by another node).

It is not only a storage system but also a platform for large data storage as well as processing the large data on a cluster of commodity hardware.

Being an Open source tool, Hadoop can be coded in any language and is not limited up to Java only. One can code in C, C++, Perl, Python, etc. Most of the codes of Hadoop are written by *Yahoo, IBM, social media giant-Facebook, and Cloudera.* Hadoop also provides you parallel processing of data as it works on multiple clusters simultaneously.

**Hadoop contains three key parts-**

* Hadoop Distributed File system(HDFS)
* MapReduce
* YARN

**How does Hadoop work? (PRIYANKA)**

The following steps show the working of Hadoop:

* The data which is collected is broken into blocks of 128MB size and then those are moved to different nodes.
* Once the blocks are stored on datanodes, a user can then process the data.
* After that, a master schedules the program(which is submitted by the user) on individual nodes.
* Lastly, once all the nodes process the data then the obtained output is written back to HDFS.

Adding more to it, there are **several components in the Hadoop Ecosystem**. These are-

* **HDFS**- Primary storage system of Hadoop.
* **MapReduce**- The data processing layer of Hadoop.
* **YARN**- resource management provider of Hadoop.
* **Hive**- It is an Open source data warehouse system for querying and analyzing large datasets.
* **Pig**- It is a High level language platform for analyzing and querying huge data sets.
* **HBase**- A component designed to store structured data in tables that could have billions of rows and columns.
* **Flume**- It efficiently collects, and moves a large amount of data from its origin and sends it back to HDFS.
* **Sqoop**- It is the importer and exporter of data from HDFS to external sources *vice-versa.*
* **Oozie**- It manages Hadoop jobs. It acts as a scheduler system.
* **Zookeeper**- large clusters of machines are managed and coordinated by Zookeeper.
* **Mahout**- an open source framework for creating scalable machine learning algorithms and data mining libraries.

**Core Hadoop Components: (PRIYANKA)**

* **HDFC:** HDFS supports the rapid transfer of data between compute nodes. At its outset, it was closely coupled with MapReduce, a programmatic framework for data processing.

When HDFS takes in data, it breaks the information down into separate blocks and distributes them to different nodes in a cluster, thus enabling highly efficient parallel processing.The default size of block is 128MB.

Moreover, the Hadoop Distributed File System is specially designed to be highly fault-tolerant. The file system replicates, or copies, each piece of data multiple times and distributes the copies to individual nodes, placing at least one copy on a different server rack than the others. As a result, the data on nodes that crash can be found elsewhere within a cluster. This ensures that processing can continue while data is recovered.

HDFS uses master/slave architecture. In its initial incarnation, each Hadoop cluster consisted of a single NameNode that managed file system operations and supporting DataNodes that managed data storage on individual compute nodes. The HDFS elements combine to support applications with large data sets.

* **Mapreduce:** The MapReduce algorithm contains two important tasks, namely Map and Reduce.

1. The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

2. The Reduce task takes the output from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

Many companies use Hadoop to compute Big Data. It has become the requirement for each and every company which is responsible for data management to manage various types of structured and unstructured data. Hadoop was initially designed to collect information from various sites. It also includes predictive analytics,data mining and machine learning applications. Hadoop comes in handy when dealing with huge amounts of data containing a variety of different data.

**Companies who deals in Hadoop: (PRIYANKA)**

* Amazon Web Services.
* Cloudera.
* Hortonworks.
* MapR Technologies.
* Intel.
* Microsoft.
* Teradata.
* Pivotal Software.