Apache Mahout

Apache Mahout is an open source library of scalable machine learning algorithms that focuses on clustering, classification, and recommendations.

**Supervised VS Unsupervised**

**Supervised learning**: Take a collection of mixed pebble, and categorize (label) them as small, medium, and large pebbles. Examples of supervised learning are regression and classification.

**Unsupervised learning**: Here, just group them based on similar sizes but don't label them. An example of unsupervised learning is clustering.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Machine learning library** | **Open source or commercial** | | **Scalable?** | | **Language used** | | **Algorithm support** |
| MATLAB | Commercial | | No | | Mostly C | | High |
| R packages | Open source | | No | | R | | High |
| Weka | Open source | | No | | Java | | High |
| Sci-Kit Learn | | Open source | | No | | Python | |
| Apache Mahout | Open source | | Yes | | Java | | Medium |
| Spark MLib | Open source | | Yes | | Scala | | Low |
| Samoa | | Open source | | Yes | | Java | |

features of Mahout:

Here, machine learning algorithms can be executed in sequential (in-memory mode) or distributed mode (MapReduce is enabled)

Data is stored in HDFS (data storage) or in memory

It runs on top of the Hadoop framework for scaling

It is not a domain-specific but a general purpose library

Even though MapReduce provides a suitable programming model for batch data processing, it does not perform well with real-time data processing.

Apache Spark is a large-scale scalable data processing framework, which claims to be 100 times faster than Hadoop MapReduce when in memory and 10 times faster in disk, has a distributed memory-based architecture. H2O is an open source, parallel processing engine for machine learning by 0xdata.

Apache Mahout is working on integrating Apache Spark and H2O as the backend integration (with the Mahout Math library).

Further, programming with Spark is easier than programming with MapReduce because Spark decouples the machine learning logic from the distributed backend. Accordingly, the distribution is hidden from the machine learning API users.

Apache Mahout is a scalable machine learning library that runs on top of the Hadoop framework. In v0.10, Apache Mahout is shifting toward Apache Spark and H20 to address performance and usability issues that occur due to the MapReduce programming paradigm



**Unsupervised learning and clustering**

Information is a key driver for any type of organization. However, with the rapid growth in the volume of data, valuable information may be hidden and go unnoticed due to the lack of effective data processing and analyzing mechanisms.

Clustering is an unsupervised learning mechanism that can find the hidden patterns and structures in data by finding data points that are similar to each other.

No pre-labeling is required. So, you can organize data using clustering with little or no human intervention.

For example, let's say you are given a collection of balls of different sizes without any category labels, such as big and small, attached to them; you should be able to categorize them using clustering by considering their attributes, such as radius and weight, for similarity.