11/07/2013 Hadoop Demo

Jian Zhang

Mahout is the ML platform.

- Actively contributed by the community
- Provide Java ML math library equivalent to matlab/R
- Transparently integrate with Hadoop and map-reduce paradigm

Demo Mahout KMeans algorithm on Hadoop mapred and hdfs

- KMeans is an unsupervised learning clustering algorithm. KMeans algorithm as follows

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

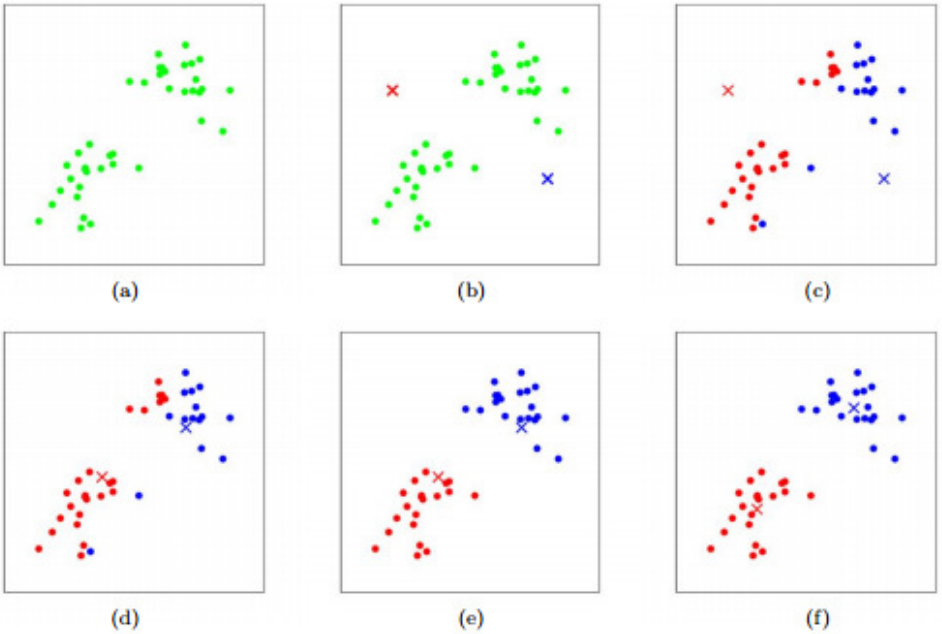
   For every i, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set

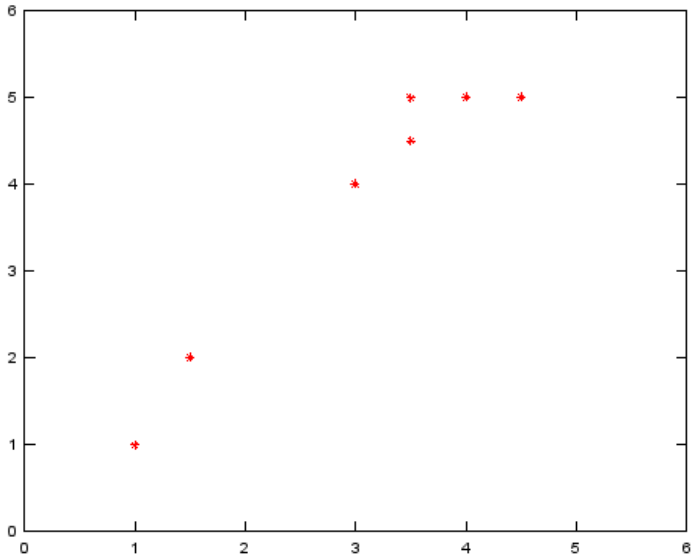
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }



- Run Mahout Kmeans job

  Input data -



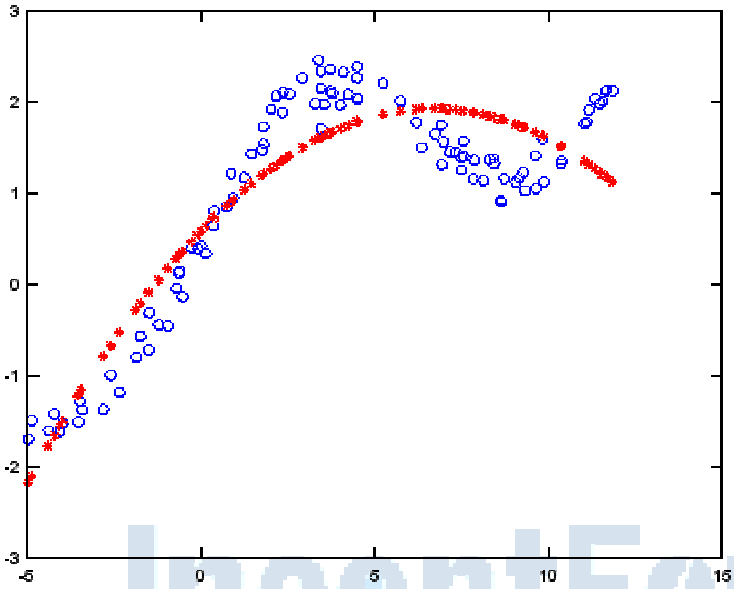
  Output - 2 clusters


```
[jzhang@jak-linux04 ~]$ start-all.sh
[jzhang@jak-linux04 ~]$ jps
26049 TaskTracker
25919 JobTracker
25544 NameNode
25821 SecondaryNameNode
25669 DataNode
28180 Jps
[jzhang@jak-linux04 ~]$ hadoop fs -put test.data /user/jzhang/testdata
[jzhang@jak-linux04 ~]$ mahout org.apache.mahout.clustering.syntheticcontrol.kmeans.Job
14/02/22 13:10:01 INFO kmeans.Job:141  Dumping out clusters from clusters: output/clusters-*-final a
oints
VL-0{n=2 c=[1.250, 1.500] r=[0.250, 0.500]}
        Weight : [props - optional]:  Point:
        1.0: [1.000, 1.000]
        1.0: [1.500, 2.000]
VL-2{n=5 c=[3.900, 5.100] r=[0.735, 1.020]}
        Weight : [props - optional]:  Point:
        1.0: [3.000, 4.000]
        1.0: [5.000, 7.000]
        1.0: [3.500, 5.000]
        1.0: [4.500, 5.000]
        1.0: [3.500, 4.500]
14/02/22 13:10:01 INFO clustering.ClusterDumper:217  Wrote 2 clusters
14/02/22 13:10:01 INFO driver.MahoutDriver:197  Program took 73237 ms (Minutes: 1.2206166666666667)
[jzhang@jak-linux04 ~]$ hadoop fs -ls output
Found 7 items
-rw-r--r--   1 jzhang supergroup        194 2014-02-22 13:09 /user/jzhang/output/_policy
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:10 /user/jzhang/output/clusteredPoints
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:09 /user/jzhang/output/clusters-0
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:09 /user/jzhang/output/clusters-1
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:09 /user/jzhang/output/clusters-2-final
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:09 /user/jzhang/output/data
drwxr-xr-x   - jzhang supergroup          0 2014-02-22 13:09 /user/jzhang/output/random-seeds
```

Demo regression algorithms

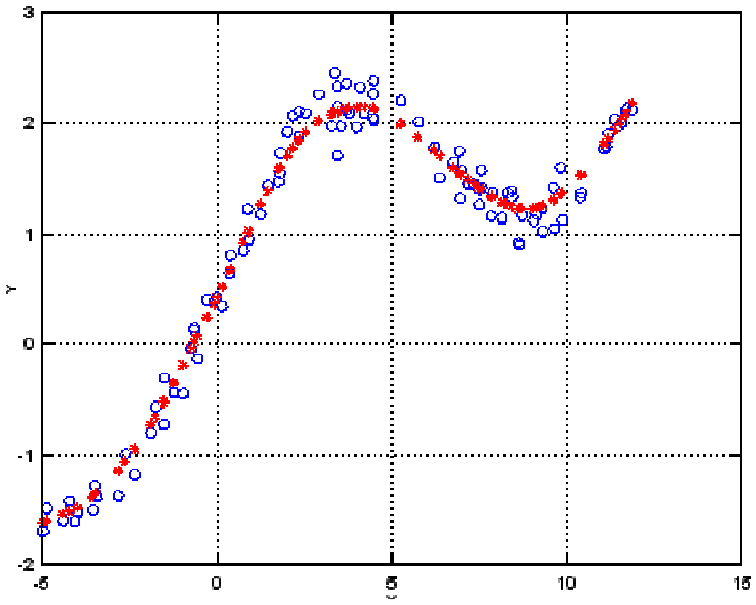
- Linear regression including multiple features and polynomial order

Linear regression is an approach to model the relationship between a scalar dependent variable y and one or more explanatory variables denoted X



- Locally weighted linear regression

Fit parameters to minimize a weighted L2 distance of original y and hypothesis y. The weight is cosmetically similar to Gaussian density function



- Logistic regression

Logistic regression is a type of probabilistic classification model used for predicting the outcome of a categorical dependent variable based on one or more feature variables. Hypothesis function is a perceptron or sigmoid function