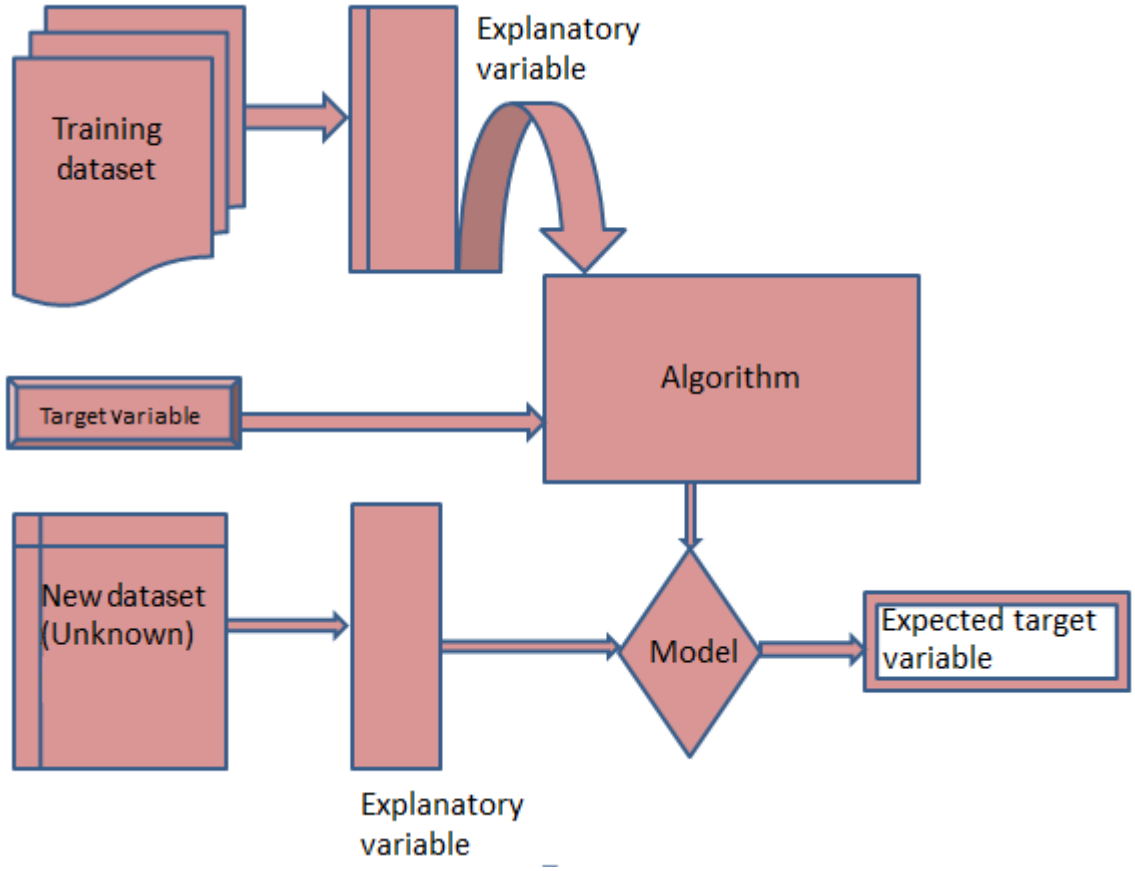
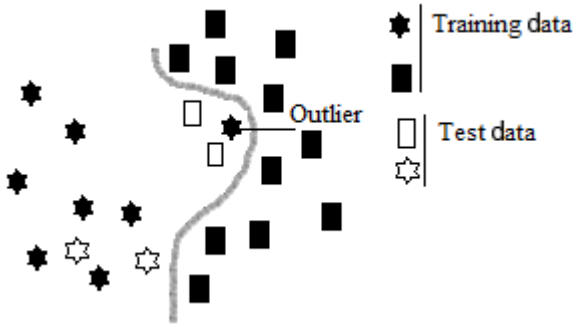
Chapter 1

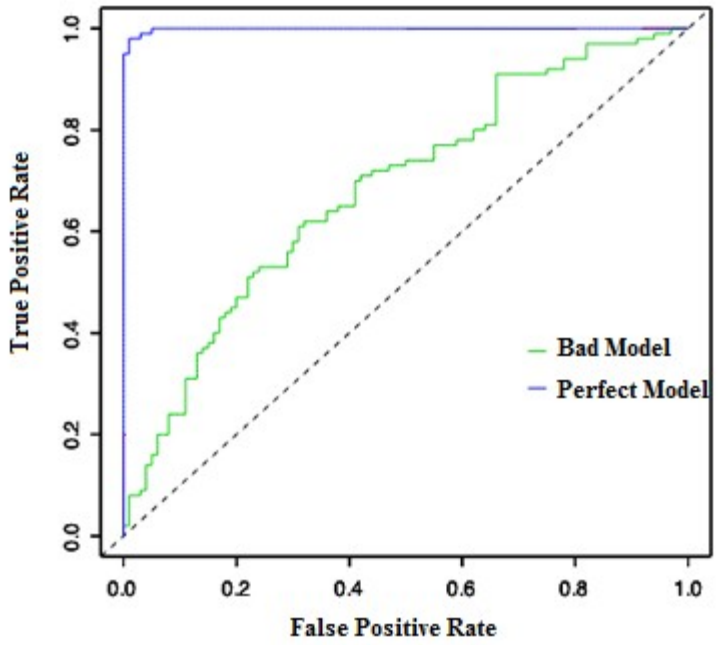
Explanatory Variables | **Target Variable (Class label)**

Customer Age	Customer Income (PA)	Customer Account Balance	Loan Granted
35	$145000	$50000	Yes
24	$50000	$500	No

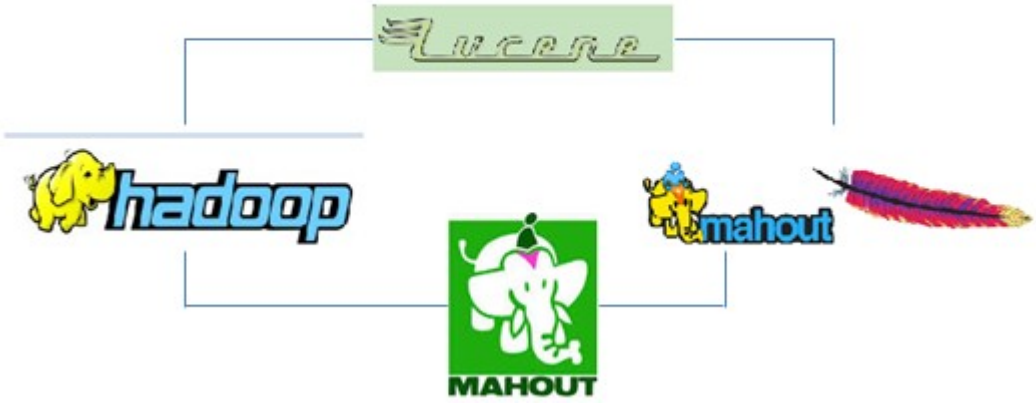
(Figure 1)



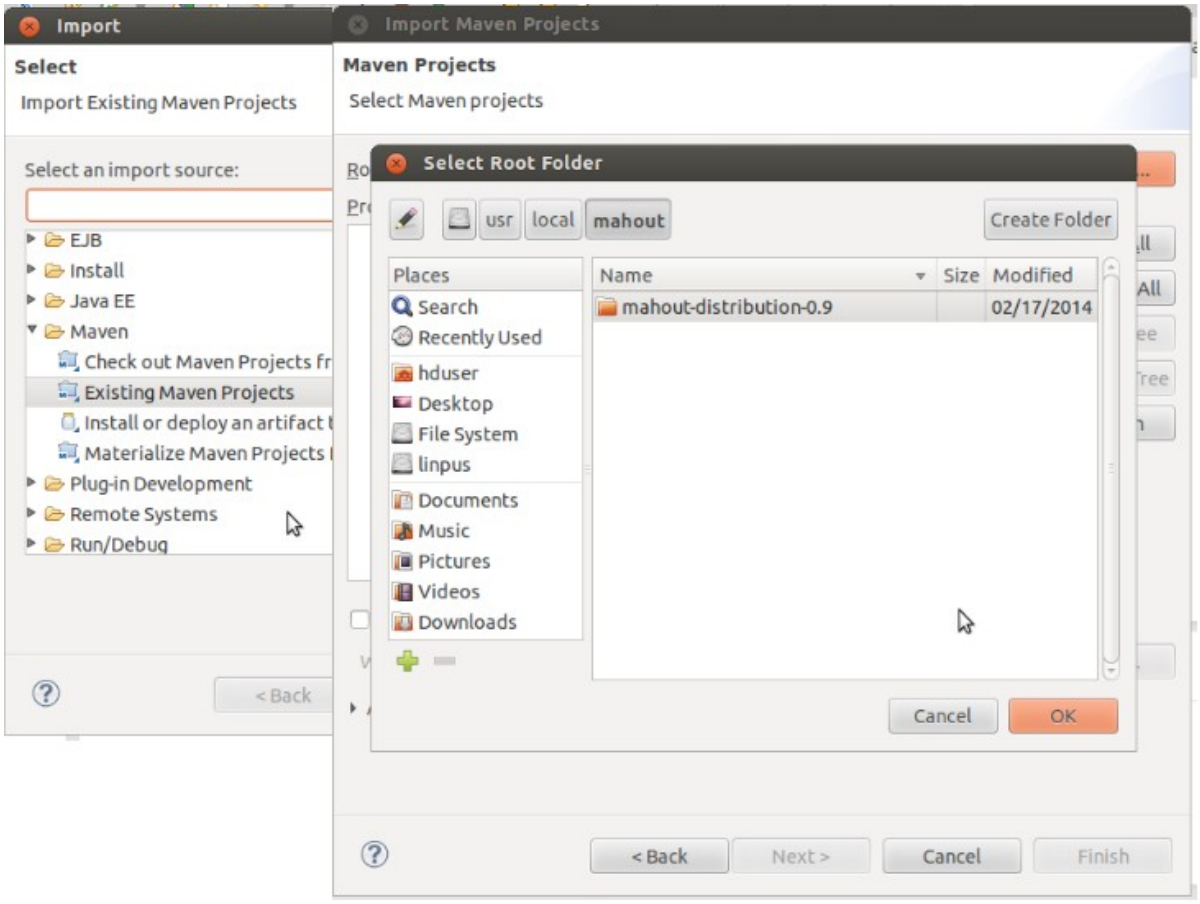

Training data
Outlier
Test data

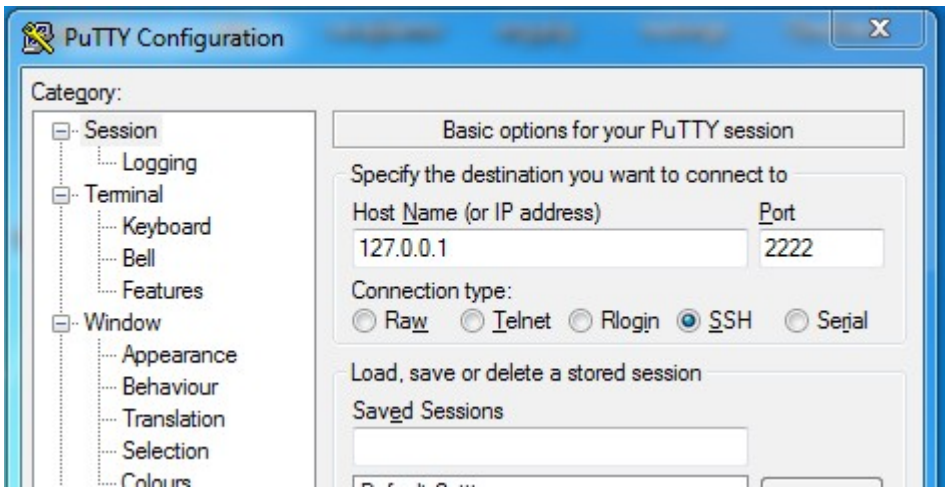

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Bad Model
Perfect Model

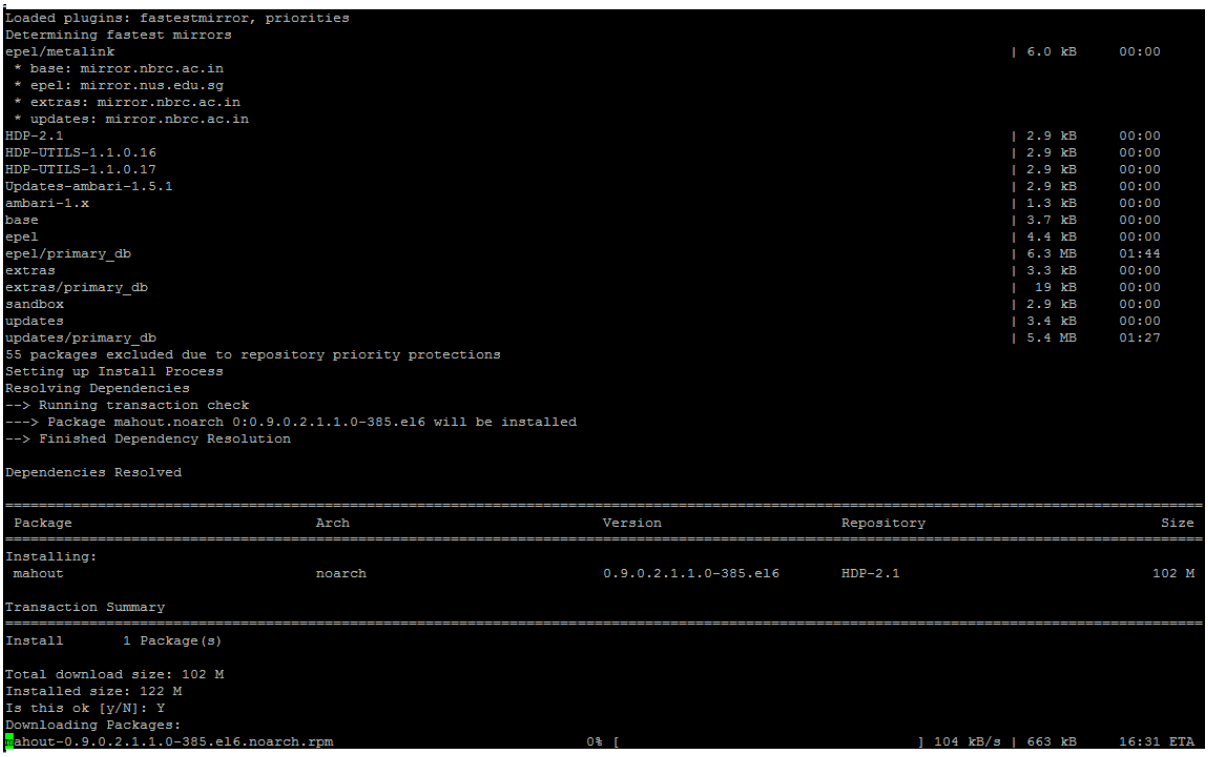
Chapter 2




```
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
An example program must be given as the first argument.
Valid program names are:
  arff.vector: : Generate Vectors from an ARFF file or directory
  baumwelch: : Baum-Welch algorithm for unsupervised HMM training
  canopy: : Canopy clustering
  cat: : Print a file or resource as the logistic regression models would see it
  cleansvd: : Cleanup and verification of SVD output
  clusterdump: : Dump cluster output to text
  clusterpp: : Groups Clustering Output In Clusters
  cmdump: : Dump confusion matrix in HTML or text formats
  concatmatrices: : Concatenates 2 matrices of same cardinality into a single matrix
  cvb: : LDA via Collapsed Variation Bayes (0th deriv. approx)
  cvb0_local: : LDA via Collapsed Variation Bayes, in memory locally.
  evaluateFactorization: : compute RMSE and MAE of a rating matrix factorization against probes
  fkmeans: : Fuzzy K-means clustering
  hmmpredict: : Generate random sequence of observations by given HMM
  itemsimilarity: : Compute the item-item-similarities for item-based collaborative filtering
  kmeans: : K-means clustering
  lucene.vector: : Generate Vectors from a Lucene index
  lucene2seq: : Generate Text SequenceFiles from a Lucene index
  matrixdump: : Dump matrix in CSV format
  matrixmult: : Take the product of two matrices
  parallelALS: : ALS-WR factorization of a rating matrix
  qualcluster: : Runs clustering experiments and summarizes results in a CSV
  recommendfactorized: : Compute recommendations using the factorization of a rating matrix
  recommenditembased: : Compute recommendations using item-based collaborative filtering
  regexconverter: : Convert text files on a per line basis based on regular expressions
  resplit: : Splits a set of SequenceFiles into a number of equal splits
```

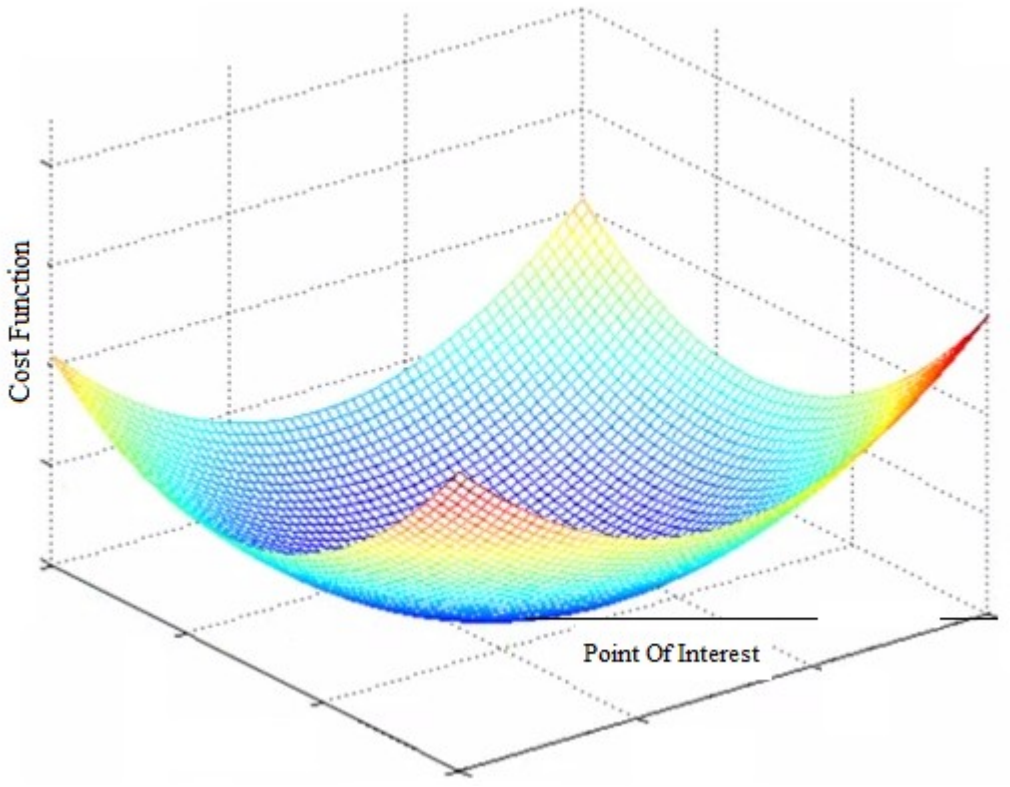

Import
Select
Import Existing Maven Projects
Select an import source:
EJB
Install
Java EE
Maven
Check out Maven Projects fr
Existing Maven Projects
Install or deploy an artifact
Materialize Maven Projects
Plug-in Development
Remote Systems
Run/Debug
< Back
Import Maven Projects
Maven Projects
Select Maven projects
Select Root Folder
usr
local
mahout
Create Folder
Places
Search
Recently Used
hduser
Desktop
File System
linpus
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
mahout-distribution-0.9
02/17/2014
Cancel
OK
< Back
Next >
Cancel
Finish

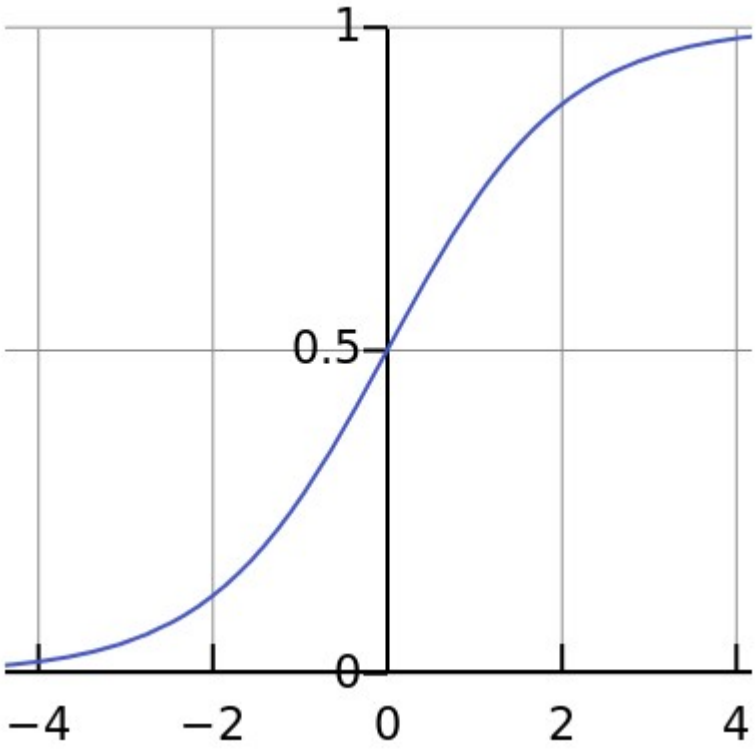

PuTTY Configuration
Category:
Session
Logging
Terminal
Keyboard
Bell
Features
Window
Appearance
Behaviour
Translation
Selection
Colours
Basic options for your PuTTY session
Specify the destination you want to connect to
Host Name (or IP address)
Port
127.0.0.1
2222
Connection type:
Raw
Telnet
Rlogin
SSH
Serial
Load, save or delete a stored session
Saved Sessions



Chapter 3

$$\text{Cost Function}\left(p_0, p_1\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\left(p_0 + p_1 R_i\right) - C_i\right)^2$$


Cost Function
Point Of Interest


1
0.5
0
−4
−2
0
2
4

```
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
30
Diagnosis ~
43553.785*Area + -9462.918*AreaStdError + 9.282*Compactness + 9.282*CompactnessStdError + 2.084*ConcavePointStdError + 4534.083*ConcavePoints + -36.930*Concavity + 1.79
0*ConcavityStdError + 54.933*FractalDimensionStderror + 32.428*Fractal_Dimension + 483.880*Intercept Term + 27723.588*Perimeter + 21.351*PerimeterStdError + 4534.083*Ra
dius + -9462.918*RadiusStdError + 40.122*Smoothness + 3.178*SmoothnessStdError + 84.224*Symmetry + 8.053*Symmetrystderror + 7237.525*Texture + 410.626*TextureStdError +
 4216.168*WorstRadius + -41451.961*worstarea + -15.680*worstcompactness + 43553.785*worstconcavepoints + 483.880*worstconcavity + 25139.478*worstfractaldimensions + 251
39.478*worstperimeter + 54.933*worstsmoothness + 105.327*worstsymmentry + 8167.715*worsttexture
              Area 43553.78472
       AreaStdError -9462.91766
        Compactness 9.28179
 CompactnessStdError 9.28179
ConcavePointStdError 2.08434
      ConcavePoints 4534.08299
          Concavity -36.92966
   ConcavityStdError 1.79025
FractalDimensionStderror 54.93294
   Fractal_Dimension 32.42769
      Intercept Term 483.87988
           Perimeter 27723.58802
   PerimeterStdError 21.35148
              Radius 4534.08299
      RadiusStdError -9462.91766
          Smoothness 40.12217
  SmoothnessStdError 3.17821
            Symmetry 84.22359
    Symmetrystderror 8.05271
             Texture 7237.52501
     TextureStdError 410.62574
         WorstRadius 4216.16816
           worstarea -41451.96054
```

```
mahout runlogistic --input /tmp/wdbcTrain.csv --model /tmp//model --auc --confusion
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
AUC = 0.88
confusion: [[151.0, 34.0], [21.0, 264.0]]
entropy: [[NaN, NaN], [-38.5, -4.3]]
14/10/25 02:50:22 INFO driver.MahoutDriver: Program took 2785 ms (Minutes: 0.04641666666666667)
```

```
mahout runlogistic --input /tmp/wdbcTest.csv --model /tmp//model --auc --confusion
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
AUC = 0.87
confusion: [[34.0, 7.0], [6.0, 52.0]]
entropy: [[NaN, NaN], [-42.4, -5.2]]
14/10/25 03:15:29 INFO driver.MahoutDriver: Program took 2248 ms (Minutes: 0.03746666666666667)
```

Chapter 4

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

$$\frac{P(A \cap B)}{P(A)}$$

$$P(A \cap B) = P(B \mid A)P(A)$$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

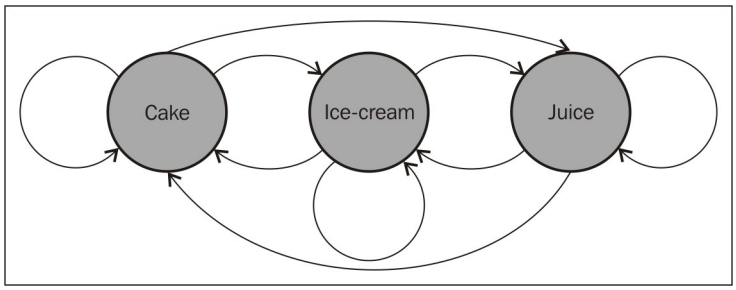
```
mahout seqdirectory -i /user/hue/20newsdata/20newsdataall -o /user/hue/20newsdataseq-out
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/11/01 09:33:07 INFO common.AbstractJob: Command line arguments: {--charset=[UTF-8], --chunkSize=[64], --endPhase=[2147483647], --fileFilterClass=
ext.PrefixAdditionFilter], --input=[/user/hue/20newsdata/20newsdataall], --keyPrefix=[], --method=[mapreduce], --output=[/user/hue/20newsdataseq-out
 --tempDir=[temp]}
14/11/01 09:33:08 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
14/11/01 09:33:08 INFO Configuration.deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
14/11/01 09:33:08 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
14/11/01 09:33:12 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/11/01 09:33:22 INFO input.FileInputFormat: Total input paths to process : 18846
14/11/01 09:33:24 INFO input.CombineFileInputFormat: DEBUG: Terminated node allocation with : CompletedNodes: 1, size left: 35855003
14/11/01 09:33:25 INFO mapreduce.JobSubmitter: number of splits:1
14/11/01 09:33:25 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1414852457629_0001
14/11/01 09:33:27 INFO impl.YarnClientImpl: Submitted application application_1414852457629_0001
14/11/01 09:33:27 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1414852457629_0001/
14/11/01 09:33:27 INFO mapreduce.Job: Running job: job_1414852457629_0001
14/11/01 09:34:01 INFO mapreduce.Job: Job job_1414852457629_0001 running in uber mode : false
14/11/01 09:34:01 INFO mapreduce.Job:  map 0% reduce 0%
14/11/01 09:34:30 INFO mapreduce.Job:  map 1% reduce 0%
```

```
                CPU time spent (ms)=9450
                Physical memory (bytes) snapshot=352317440
                Virtual memory (bytes) snapshot=1801748480
                Total committed heap usage (bytes)=191889408
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=28689283
        File Output Format Counters
                Bytes Written=28689283
14/11/01 10:44:04 INFO common.HadoopUtil: Deleting /user/hue/20newsdatavec/tf-vectors-partial
14/11/01 10:44:04 INFO common.HadoopUtil: Deleting /user/hue/20newsdatavec/tf-vectors-toprune
14/11/01 10:44:05 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/11/01 10:44:08 INFO input.FileInputFormat: Total input paths to process : 1
14/11/01 10:44:08 INFO mapreduce.JobSubmitter: number of splits:1
14/11/01 10:44:08 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1414852457629_0009
14/11/01 10:44:08 INFO impl.YarnClientImpl: Submitted application application_1414852457629_0009
14/11/01 10:44:08 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1414852457629_0009/
14/11/01 10:44:08 INFO mapreduce.Job: Running job: job_1414852457629_0009
14/11/01 10:44:29 INFO mapreduce.Job: Job job_1414852457629_0009 running in uber mode : false
14/11/01 10:44:29 INFO mapreduce.Job:  map 0% reduce 0%
14/11/01 10:44:51 INFO mapreduce.Job:  map 100% reduce 0%
14/11/01 10:45:23 INFO mapreduce.Job:  map 100% reduce 79%
14/11/01 10:45:25 INFO mapreduce.Job:  map 100% reduce 100%
14/11/01 10:45:27 INFO mapreduce.Job: Job job_1414852457629_0009 completed successfully
14/11/01 10:45:27 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=30327803
                FILE: Number of bytes written=57077985
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=28689430
                HDFS: Number of bytes written=28689283
                HDFS: Number of read operations=7
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
```

```
mahout split -i /user/hue/20newsdatavec/tfidf-vectors --trainingOutput /user/hue/20newsdatatrain --testOutput /user/hue/20newsdatatest --randomSelectionPct 40 -overwrite
 --sequenceFiles -xm sequential
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/11/01 11:12:47 WARN driver.MahoutDriver: No split.props found on classpath, will use command-line arguments only
14/11/01 11:12:48 INFO common.AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/user/hue/20newsdatavec/tfidf-vectors], --method=[sequential], --o
verwrite=null, --randomSelectionPct=[40], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --testOutput=[/user/hue/20newsdatatest], --trainingOutput=[/user/hue
/20newsdatatrain]}
14/11/01 11:12:56 INFO utils.SplitInput: part-r-00000 has 162419 lines
14/11/01 11:12:56 INFO utils.SplitInput: part-r-00000 test split size is 64968 based on random selection percentage 40
14/11/01 11:12:56 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
14/11/01 11:12:56 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/11/01 11:12:56 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/11/01 11:13:02 INFO utils.SplitInput: file: part-r-00000, input: 162419 train: 11311, test: 7535 starting at 0
14/11/01 11:13:02 INFO driver.MahoutDriver: Program took 14552 ms (Minutes: 0.24253333333333332)
```

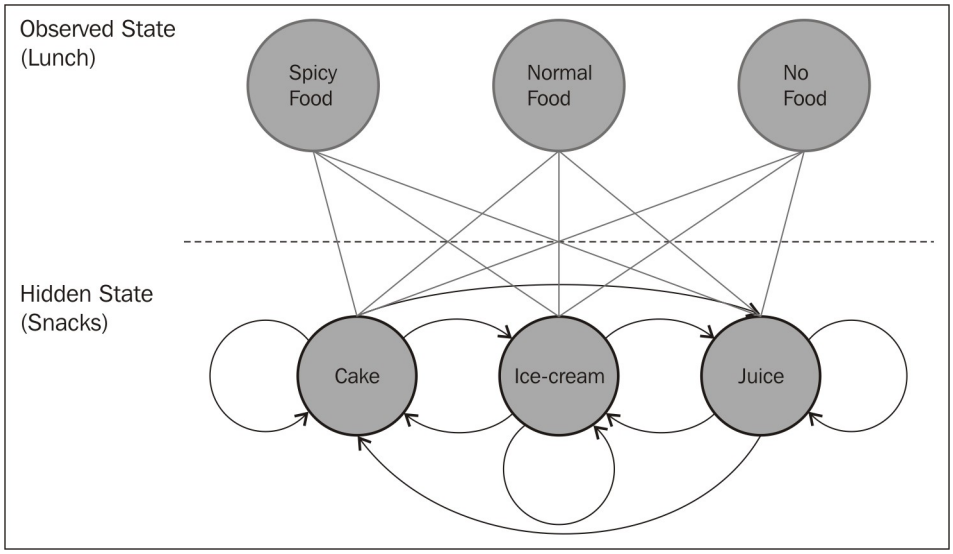
```
=====================================================
Summary
-----------------------------------------------------
Correctly Classified Instances          :       6860       91.0418%
Incorrectly Classified Instances        :        675        8.9582%
Total Classified Instances              :       7535

=====================================================
Confusion Matrix
-----------------------------------------------------
a      b      c      d      e      f      g      h      i      j      k      l      m      n      o      p      q      r      s      t      <--Class
ified as
293    0      0      0      0      0      0      0      0      0      0      1      0      1      1      1      0      0      8      2      |  307
       a     = alt.atheism
2      329    3      10     6      12     10     1      0      0      0      1      6      1      1      0      1      0      0      0      |  383
       b     = comp.graphics
0      31     280    49     16     16     10     0      0      0      0      3      1      1      2      0      0      0      0      2      |  411
       c     = comp.os.ms-windows.misc
0      14     3      353    18     5      12     1      0      0      0      1      10     0      0      0      0      0      0      0      |  417
       d     = comp.sys.ibm.pc.hardware
0      2      3      7      365    1      3      1      0      0      1      0      6      1      1      0      0      0      0      0      |  391
       e     = comp.sys.mac.hardware
0      24     2      3      2      359    5      0      1      0      0      0      0      0      1      0      0      1      0      0      |  398
       f     = comp.windows.x
0      4      0      20     10     0      330    8      3      2      2      0      12     1      1      1      0      0      1      1      |  396
       g     = misc.forsale
0      1      0      1      0      0      8      391    3      0      0      1      7      0      0      0      0      1      0      2      |  415
       h     = rec.autos
0      0      0      0      0      0      4      6      392    0      0      0      2      1      0      0      0      2      0      0      |  407
       i     = rec.motorcycles
0      0      0      0      0      0      4      1      2      365    0      1      2      0      0      0      0      0      1      0      |  376
       j     = rec.sport.baseball
0      2      0      2      1      0      0      0      0      3      389    0      1      0      0      0      0      0      0      1      |  399
       k     = rec.sport.hockey
0      3      0      0      0      2      2      0      0      0      0      375    0      0      0      0      1      2      0      2      |  387
       l     = sci.crypt
0      6      0      9      11     1      4      5      0      0      0      2      326    1      5      0      0      0      0      2      |  372
       m     = sci.electronics
1      2      0      0      1      0      2      2      1      0      1      0      4      393    3      0      0      0      0      2      |  412
       n     = sci.med
0      8      0      0      3      1      1      0      0      1      0      0      0      0      373    0      0      0      3      2      |  392
       o     = sci.space
4      2      0      1      0      0      1      1      0      0      0      0      0      3      0      394    1      1      5      1      |  414
       p     = soc.religion.christian
0      0      0      0      0      0      0      0      0      0      0      0      0      0      0      2      348    1      0      2      |  353
       q     = talk.politics.mideast
0      0      0      0      0      0      1      0      0      1      0      2      0      1      1      1      1      322    0      11     |  341
       r     = talk.politics.guns
33     0      0      1      0      0      0      0      0      0      0      0      0      0      0      9      1      5      182    4      |  235
       s     = talk.religion.misc
1      0      0      0      0      0      0      0      2      2      0      3      0      1      1      0      4      10     4      301    |  329
       t     = talk.politics.misc

=====================================================
```

Chapter 5



Yesterday \ Today	State 1	State 2	State 3
State 1	0.1	0.8	0.2
State 2	0.3	0.1	0.1
State 3	0.6	0.1	0.7

State 1	State 2	State 3
1	0	0



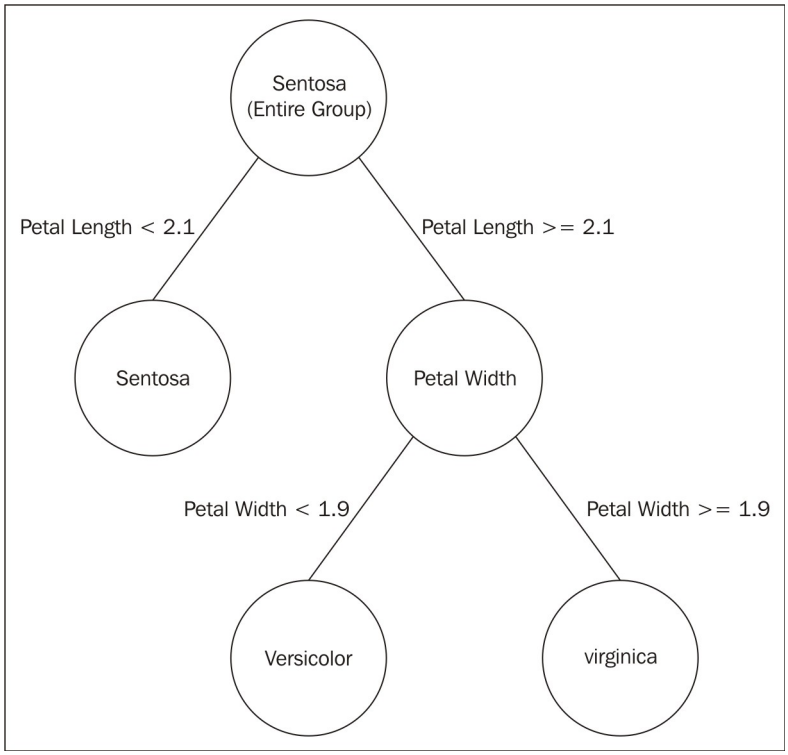

```
mahout baumwelch -i /tmp/hmm/hmm-input -o /tmp/hmm/hmm-model -nh 3 -no 3 -e .0001 -m 1000
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/11/06 11:02:51 WARN driver.MahoutDriver: No baumwelch.props found on classpath, will use command-line arguments only
Initial probabilities:
0 1 2
1.0 1.309026126805178E-89 1.7587674488030393E-78
Transition matrix:
  0 1 2
0 4.993504845479359E-13 0.5227259091584807 0.4772740908410199
1 0.9998785172353453 3.084831951926291E-9 1.2147967982263436E-4
2 2.0416849285597051E-4 0.10221844628449037 0.8975773852226537
Emission matrix:
  0 1 2
0 0.999999999999999 1.1282788180575074E-24 9.881045303840455E-16
1 0.7514502891771425 0.2485497108228551 2.32229480690411E-15
2 0.08635564236932171 0.4108528050151504 0.502791552615528
```

```
mahout viterbi --input /tmp/hmm/hmm-viterbi-input --output /tmp/hmm/hmm-viterbi-output --model /tmp/hmm/hmm-model --likelihood
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/11/06 11:42:22 WARN driver.MahoutDriver: No viterbi.props found on classpath, will use command-line arguments only
Likelihood: 1.2864746327281034E-6
14/11/06 11:42:24 INFO driver.MahoutDriver: Program took 1540 ms (Minutes: 0.025666666666666667)
```

```
cat /tmp/hmm/hmm-viterbi-output
0 2 2 2 2 2 2 1 0 2 2 2
```

```
cat hmm-predictions
0 0 0 1 1 2 2 2 2 2
```

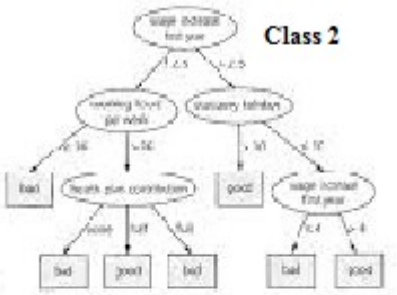
Chapter 6

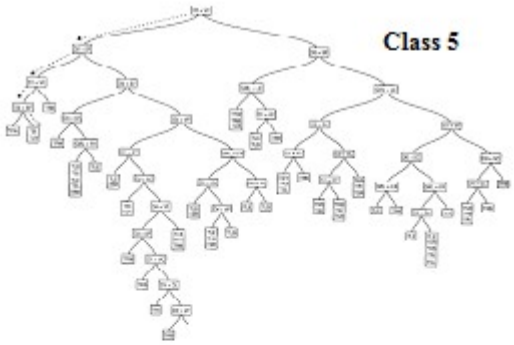


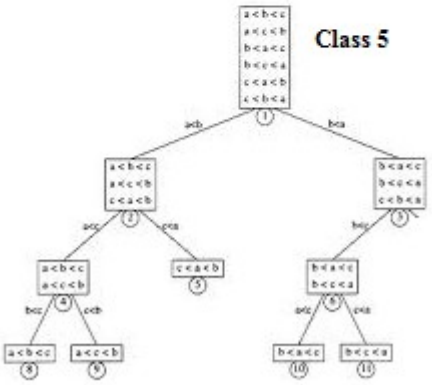
$$\begin{bmatrix} f_1 & f_2 & f_3 & \ldots\ldots & T_1 \\ f_1, & f_2, & f_3, & \ldots\ldots & T_1, \\ & & \vdots & & \\ f_{n1} & f_{n2} & f_{n3} & \ldots\ldots & T_n \end{bmatrix}$$

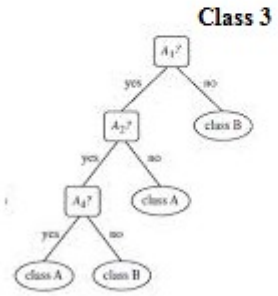
A sample training set with features f's and target classes T's

$$\begin{bmatrix} f_1, f_2, f_3, \ldots\ldots & T_1, \\ \vdots & \\ f_{14}, f_{15}, f_{16}, \ldots & T_{14}, \end{bmatrix} \quad \begin{bmatrix} f_{61}, f_{62}, f_{63}, \ldots. & T_1, \\ \vdots & \\ f_{84}, f_{85}, f_{86}, \ldots & T_{14}, \end{bmatrix}$$









The NSL-KDD Dataset

Abstract

NSL-KDD is a data set suggested to solve some of the inherent problems of the KDD'99 data set which are mentioned in [1]. Although, this new version of the KDD data set still suffers from some of the problems discussed by McHugh [2] and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, we believe it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. Furthermore, the number of records in the NSL-KDD train and test sets are reasonable. This advantage makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research work will be consistent and comparable.

Data Files

KDDTrain+.ARFF	The full NSL-KDD train set with binary labels in ARFF format
KDDTrain+.TXT	The full NSL-KDD train set including attack-type labels and difficulty level in CSV format
KDDTrain+_20Percent.ARFF	A 20% subset of the KDDTrain+.arff file
KDDTrain+_20Percent.TXT	A 20% subset of the KDDTrain+.txt file
KDDTest+.ARFF	The full NSL-KDD test set with binary labels in ARFF format
KDDTest+.TXT	The full NSL-KDD test set including attack-type labels and difficulty level in CSV format
KDDTest-21.ARFF	A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21
KDDTest-21.TXT	A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21

```
@relation 'KDDTrain-20Percent'
@attribute 'duration' real
@attribute 'protocol_type' {'tcp','udp', 'icmp'}
@attribute 'service' {'aol', 'auth', 'bgp', 'courier', 'csnet_ns', 'ctf', 'daytime', 'discard', 'domain', 'domain_u', 'echo', 'eco
'finger', 'ftp', 'ftp_data', 'gopher', 'harvest', 'hostnames', 'http', 'http_2784', 'http_443', 'http_8001', 'imap4', 'IRC', 'iso_
'ldap', 'link', 'login', 'mtp', 'name', 'netbios_dgm', 'netbios_ns', 'netbios_ssn', 'netstat', 'nnsp', 'nntp', 'ntp_u', 'other', '
'printer', 'private', 'red_i', 'remote_job', 'rje', 'shell', 'smtp', 'sql_net', 'ssh', 'sunrpc', 'supdup', 'systat', 'telnet', 'tf
'urh_i', 'urp_i', 'uucp', 'uucp_path', 'vmnet', 'whois', 'X11', 'Z39_50'}
@attribute 'flag' { 'OTH', 'REJ', 'RSTO', 'RSTOS0', 'RSTR', 'S0', 'S1', 'S2', 'S3', 'SF', 'SH' }
@attribute 'src_bytes' real
@attribute 'dst_bytes' real
@attribute 'land' {'0', '1'}
@attribute 'wrong_fragment' real
@attribute 'urgent' real
@attribute 'hot' real
@attribute 'num_failed_logins' real
@attribute 'logged_in' {'0', '1'}
@attribute 'num_compromised' real
@attribute 'root_shell' real
@attribute 'su_attempted' real
@attribute 'num_root' real
@attribute 'num_file_creations' real
@attribute 'num_shells' real
@attribute 'num_access_files' real
@attribute 'num_outbound_cmds' real
@attribute 'is_host_login' {'0', '1'}
@attribute 'is_guest_login' {'0', '1'}
@attribute 'count' real
@attribute 'srv_count' real
@attribute 'serror_rate' real
@attribute 'srv_serror_rate' real
@attribute 'rerror_rate' real
@attribute 'srv_rerror_rate' real
```

```
hadoop jar /usr/lib/mahout/mahout-core-0.9.0.2.1.1.0-385-job.jar org.apache.mahout.classifier.df.tools.Describe -p /user/hue/KDDTrain/KDDTrain+_20Percent.arff -f
/user/hue/KDDTrain/KDDTrain+.info -d N 3 C 2 N C 4 N C 8 N 2 C 19 N L
14/11/22 18:43:21 INFO tools.Describe: Generating the descriptor...
14/11/22 18:43:38 INFO tools.Describe: generating the dataset...
14/11/22 18:43:42 INFO tools.Describe: storing the dataset description
```

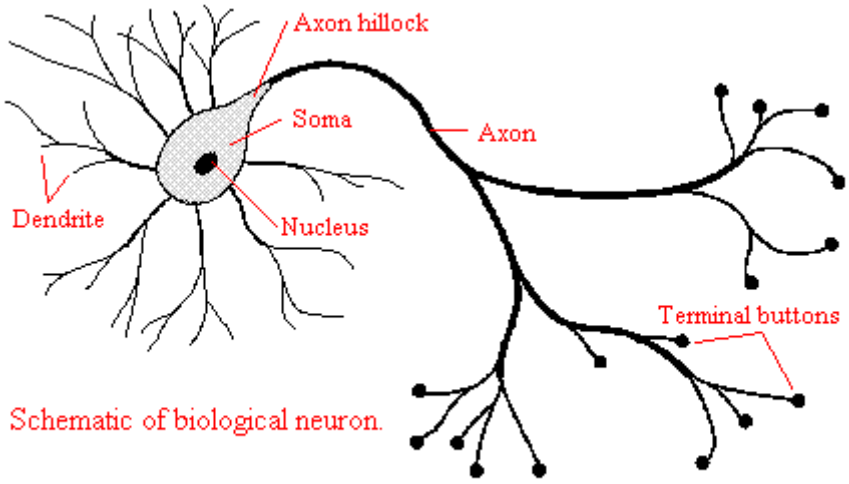
```
hadoop jar /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar org.apache.mahout.classifier.df.mapreduce.BuildForest -Dmapred.max.split.size=1874231 -d /user/hue
/KDDTrain/KDDTrain+_20Percent.arff -ds /user/hue/KDDTrain/KDDTrain+.info -sl 5 -p -t 100 -o /user/hue/nsl-forest
14/11/22 19:27:14 INFO mapreduce.BuildForest: Partial Mapred implementation
14/11/22 19:27:14 INFO mapreduce.BuildForest: Building the forest...
14/11/22 19:27:16 INFO input.FileInputFormat: Total input paths to process : 1
14/11/22 19:27:16 INFO partial.PartialBuilder: Setting mapred.map.tasks = 2
14/11/22 19:27:16 INFO Configuration.deprecation: mapred.map.tasks is deprecated. Instead, use mapreduce.job.maps
14/11/22 19:27:17 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/11/22 19:27:25 INFO input.FileInputFormat: Total input paths to process : 1
14/11/22 19:27:25 INFO mapreduce.JobSubmitter: number of splits:2
14/11/22 19:27:25 INFO Configuration.deprecation: mapred.max.split.size is deprecated. Instead, use mapreduce.input.fileinputformat.split.maxsize
14/11/22 19:27:26 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1416704446466_0001
14/11/22 19:27:27 INFO impl.YarnClientImpl: Submitted application application_1416704446466_0001
14/11/22 19:27:27 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1416704446466_0001/
14/11/22 19:27:27 INFO mapreduce.Job: Running job: job_1416704446466_0001
14/11/22 19:28:31 INFO mapreduce.Job: Job job_1416704446466_0001 running in uber mode : false
14/11/22 19:28:32 INFO mapreduce.Job:  map 0% reduce 0%
14/11/22 19:30:31 INFO mapreduce.Job:  map 100% reduce 0%
```

```
14/11/22 19:30:46 INFO mapreduce.BuildForest: Build Time: 0h 3m 31s 378
14/11/22 19:30:46 INFO mapreduce.BuildForest: Forest num Nodes: 49947
14/11/22 19:30:46 INFO mapreduce.BuildForest: Forest mean num Nodes: 499
14/11/22 19:30:46 INFO mapreduce.BuildForest: Forest mean max Depth: 13
14/11/22 19:30:46 INFO mapreduce.BuildForest: Storing the forest in: /user/hue/nsl-forest/forest.seq
```

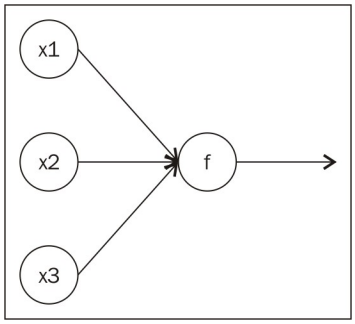
```
hadoop jar /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar org.apache.mahout.classifier.df.mapreduce.TestForest -i /user/hue/KDDTest/KDDTest+.arff -ds /user/
hue/KDDTrain/KDDTrain+.info -m /user/hue/nsl-forest -a -mr -o /user/hue/predictions
14/11/22 19:57:07 INFO mapreduce.Classifier: Adding the dataset to the DistributedCache
14/11/22 19:57:07 INFO mapreduce.Classifier: Adding the decision forest to the DistributedCache
14/11/22 19:57:07 INFO mapreduce.Classifier: Configuring the job...
14/11/22 19:57:07 INFO mapreduce.Classifier: Running the job...
14/11/22 19:57:07 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/11/22 19:57:13 INFO input.FileInputFormat: Total input paths to process : 1
14/11/22 19:57:13 INFO mapreduce.JobSubmitter: number of splits:1
14/11/22 19:57:14 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1416704446466_0002
14/11/22 19:57:15 INFO impl.YarnClientImpl: Submitted application application_1416704446466_0002
14/11/22 19:57:15 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1416704446466_0002/
14/11/22 19:57:15 INFO mapreduce.Job: Running job: job_1416704446466_0002
14/11/22 19:58:00 INFO mapreduce.Job: Job job_1416704446466_0002 running in uber mode : false
14/11/22 19:58:00 INFO mapreduce.Job:  map 0% reduce 0%
14/11/22 19:58:29 INFO mapreduce.Job:  map 100% reduce 0%
```

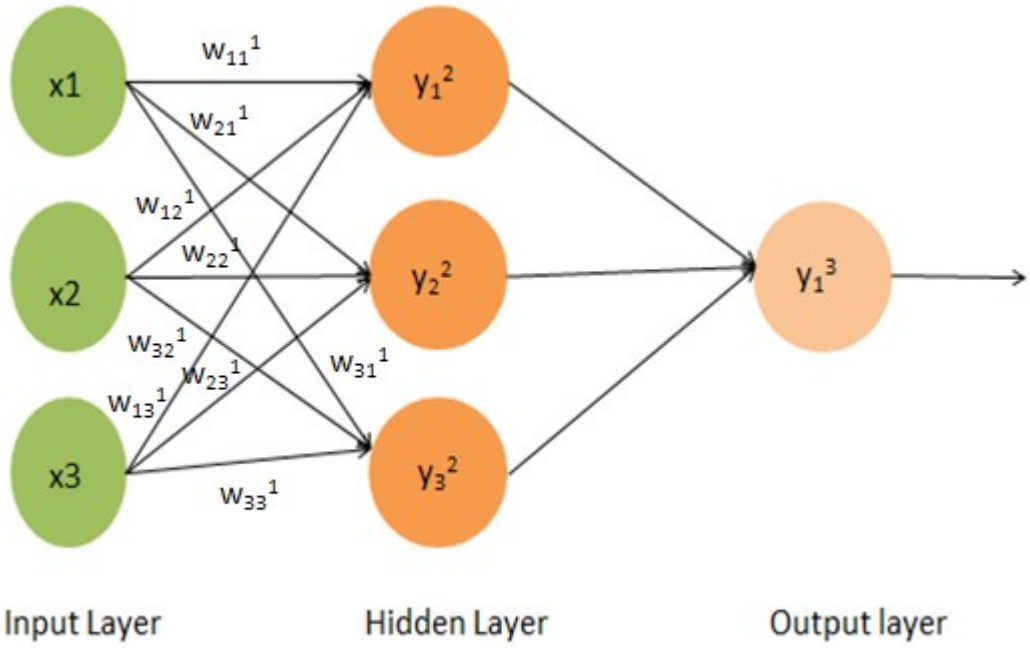
```
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :      17531       77.7635%
Incorrectly Classified Instances        :       5013       22.2365%
Total Classified Instances              :      22544

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       <--Classified as
9396    315      |  9711        a     = normal
4698    8135     |  12833       b     = anomaly

=======================================================
Statistics
-------------------------------------------------------
Kappa                                           0.57
Accuracy                                     77.7635%
Reliability                                  53.3825%
Reliability (standard deviation)              0.4915
```

Chapter 7



Schematic of biological neuron.


x1
x2
x3
f


x1
x2
x3
w_{11}^1
w_{21}^1
w_{12}^1
w_{22}^1
w_{32}^1
w_{23}^1
w_{13}^1
w_{31}^1
w_{33}^1
y_1^2
y_2^2
y_3^2
y_1^3
Input Layer
Hidden Layer
Output layer

Index of /ml/machine-learning-databases/iris

Name	Last modified	Size	Description
Parent Directory		-	
Index	03-Dec-1996 04:01	105	
bezdekIris.data	14-Dec-1999 12:12	4.4K	
iris.data	08-Mar-1993 16:27	4.4K	
iris.names	11-Jul-2000 21:30	2.9K	

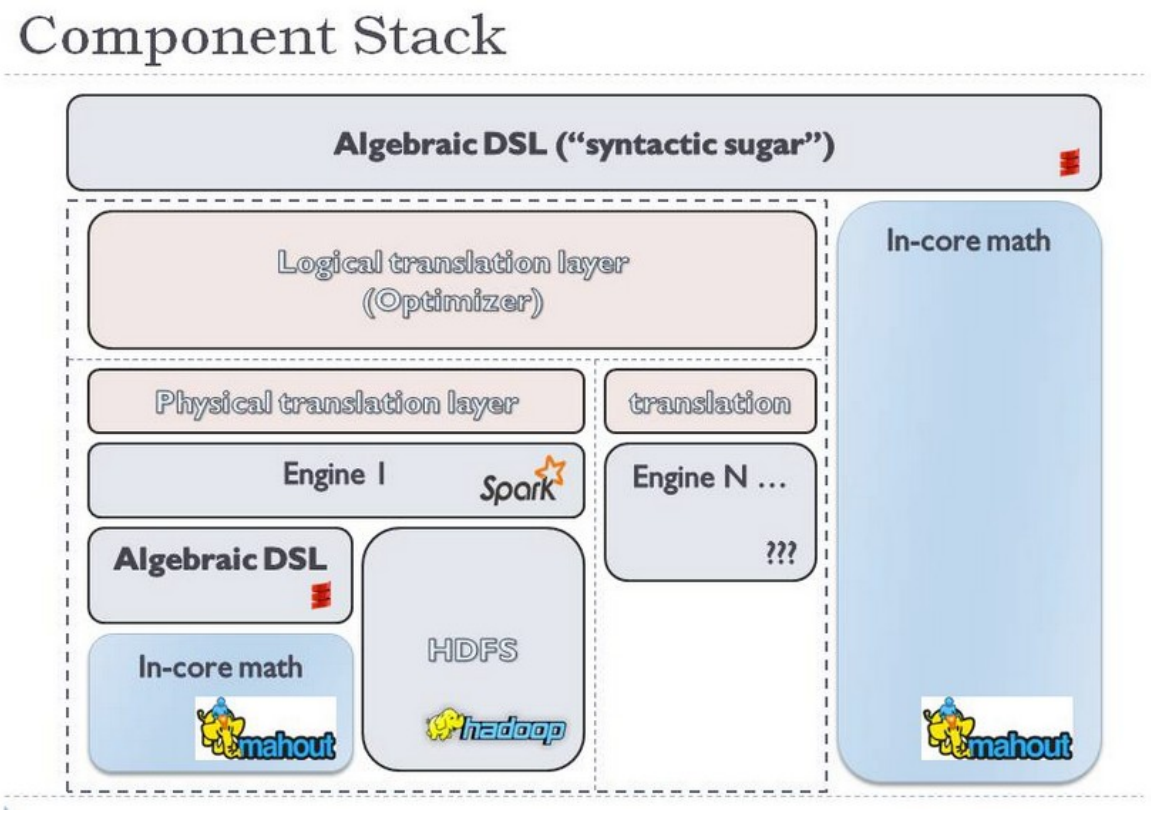
Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 80

```
1  5.1,3.5,1.4,0.2,Iris-setosa
2  4.9,3.0,1.4,0.2,Iris-setosa
3  4.7,3.2,1.3,0.2,Iris-setosa
4  4.6,3.1,1.5,0.2,Iris-setosa
5  5.0,3.6,1.4,0.2,Iris-setosa
6  5.4,3.9,1.7,0.4,Iris-setosa
7  4.6,3.4,1.4,0.3,Iris-setosa
8  5.0,3.4,1.5,0.2,Iris-setosa
9  4.4,2.9,1.4,0.2,Iris-setosa
10 4.9,3.1,1.5,0.1,Iris-setosa
11 5.4,3.7,1.5,0.2,Iris-setosa
12 4.8,3.4,1.6,0.2,Iris-setosa
13 4.8,3.0,1.4,0.1,Iris-setosa
14 4.3,3.0,1.1,0.1,Iris-setosa
15 5.8,4.0,1.2,0.2,Iris-setosa
16 5.7,4.4,1.5,0.4,Iris-setosa
17 5.4,3.9,1.3,0.4,Iris-setosa
18 5.1,3.5,1.4,0.3,Iris-setosa
19 5.7,3.8,1.7,0.3,Iris-setosa
20 5.1,3.8,1.5,0.3,Iris-setosa
21 5.4,3.4,1.7,0.2,Iris-setosa
22 5.1,3.7,1.5,0.4,Iris-setosa
23 4.6,3.6,1.0,0.2,Iris-setosa
24 5.1,3.3,1.7,0.5,Iris-setosa
```

```
inMultilayerPerceptron -i /tmp/irisdata.csv -labels Iris-setosa Iris-versicolor Iris-virginica -mo /tmp/model.model -ls 4 8 3
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/mahout/mahout-core-0.9.0.2.1.1.0-385-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
log4j:WARN No appenders could be found for logger (org.apache.mahout.classifier.mlp.TrainMultilayerPerceptron).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Input: /tmp/irisdata.csv, Model: /tmp/model.model, Update: false, Layer size: [4, 8, 3], Squashing function: Sigmoid, Learning rate: 0.500000, Momemtum weight: 0.100000
, Regularization Weight: 0.000000
```

```
ultilayerPerceptron -i /tmp/irisdata.csv -sh --columnRange -mo /tmp/model.model -o /tmp/labelResult.txt
LF4J: Class path contains multiple SLF4J bindings.
LF4J: Found binding in [jar:file:/usr/lib/hadoop/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
LF4J: Found binding in [jar:file:/usr/lib/mahout/mahout-core-0.9.0.2.1.1.0-385-job.jar!/org/slf4j/impl/StaticLoggerBinder.class]
LF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
LF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
og4j:WARN No appenders could be found for logger (org.apache.mahout.classifier.mlp.RunMultilayerPerceptron).
og4j:WARN Please initialize the log4j system properly.
og4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
```

Chapter 8



```
Launching sbt from sbt/sbt-launch-0.13.5.jar
Getting org.scala-sbt sbt 0.13.5 ...
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/sbt/0.13.5/jars/sbt.jar ...
        [SUCCESSFUL ] org.scala-sbt#sbt;0.13.5!sbt.jar (2287ms)
downloading http://repo1.maven.org/maven2/org/scala-lang/scala-library/2.10.4/scala-library-2.10.4.jar ...
        [SUCCESSFUL ] org.scala-lang#scala-library;2.10.4!scala-library.jar (13640ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/main/0.13.5/jars/main.jar ...
        [SUCCESSFUL ] org.scala-sbt#main;0.13.5!main.jar (8780ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/compiler-interface/0.13.5/jars/compiler-interface-src.jar ...
        [SUCCESSFUL ] org.scala-sbt#compiler-interface;0.13.5!compiler-interface-src.jar (1775ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/compiler-interface/0.13.5/jars/compiler-interface-bin.jar ...
        [SUCCESSFUL ] org.scala-sbt#compiler-interface;0.13.5!compiler-interface-bin.jar (2758ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/precompiled-2_8_2/0.13.5/jars/compiler-interface-bin.jar ...
        [SUCCESSFUL ] org.scala-sbt#precompiled-2_8_2;0.13.5!compiler-interface-bin.jar (3975ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/precompiled-2_9_2/0.13.5/jars/compiler-interface-bin.jar ...
        [SUCCESSFUL ] org.scala-sbt#precompiled-2_9_2;0.13.5!compiler-interface-bin.jar (2727ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/precompiled-2_9_3/0.13.5/jars/compiler-interface-bin.jar ...
        [SUCCESSFUL ] org.scala-sbt#precompiled-2_9_3;0.13.5!compiler-interface-bin.jar (4058ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/actions/0.13.5/jars/actions.jar ...
        [SUCCESSFUL ] org.scala-sbt#actions;0.13.5!actions.jar (3909ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/main-settings/0.13.5/jars/main-settings.jar ...
        [SUCCESSFUL ] org.scala-sbt#main-settings;0.13.5!main-settings.jar (5474ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/interface/0.13.5/jars/interface.jar ...
        [SUCCESSFUL ] org.scala-sbt#interface;0.13.5!interface.jar (2766ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/io/0.13.5/jars/io.jar ...
        [SUCCESSFUL ] org.scala-sbt#io;0.13.5!io.jar (7124ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/ivy/0.13.5/jars/ivy.jar ...
        [SUCCESSFUL ] org.scala-sbt#ivy;0.13.5!ivy.jar (5411ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/launcher-interface/0.13.5/jars/launcher-interface.jar ...
        [SUCCESSFUL ] org.scala-sbt#launcher-interface;0.13.5!launcher-interface.jar (1725ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/logging/0.13.5/jars/logging.jar ...
        [SUCCESSFUL ] org.scala-sbt#logging;0.13.5!logging.jar (2375ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/logic/0.13.5/jars/logic.jar ...
        [SUCCESSFUL ] org.scala-sbt#logic;0.13.5!logic.jar (3060ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/process/0.13.5/jars/process.jar ...
        [SUCCESSFUL ] org.scala-sbt#process;0.13.5!process.jar (2743ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/run/0.13.5/jars/run.jar ...
        [SUCCESSFUL ] org.scala-sbt#run;0.13.5!run.jar (2882ms)
downloading http://repo.typesafe.com/typesafe/ivy-releases/org.scala-sbt/command/0.13.5/jars/command.jar ...
        [SUCCESSFUL ] org.scala-sbt#command;0.13.5!command.jar (4085ms)
```

```
# git clone https://github.com/apache/mahout mahout
Initialized empty Git repository in /tmp/Mahout/mahout/.git/
remote: Counting objects: 79953, done.
Receiving objects: 100% (79953/79953), 39.36 MiB | 218 KiB/s, done.
remote: Total 79953 (delta 0), reused 0 (delta 0)
Resolving deltas: 100% (43010/43010), done.
```

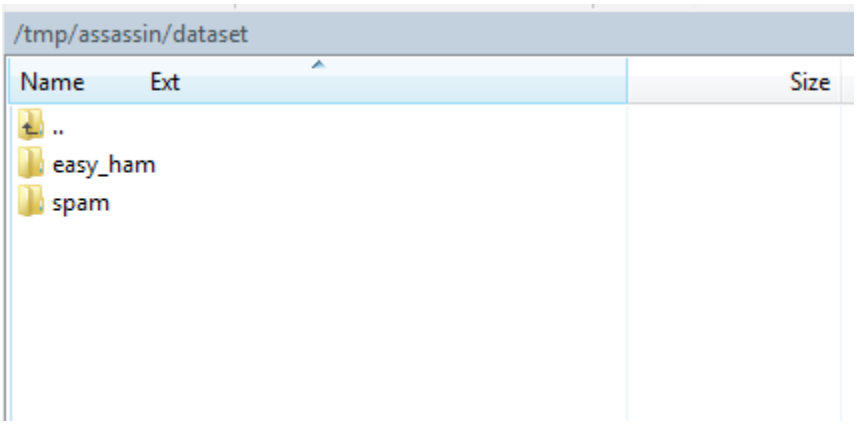
```
[# mvn -DskipTests clean install
/usr/local/maven/apache-maven-3.2.3/bin/mvn: line 53: uname: command not found
[INFO] Scanning for projects...
Downloading: https://repository.cloudera.com/artifactory/cloudera-repos/org/apache/apache/9/apache-9.pom
Downloading: http://repository.mapr.com/maven/org/apache/apache/9/apache-9.pom
Dec 03, 2014 9:19:33 AM org.apache.maven.wagon.providers.http.httpclient.client.protocol.ResponseProcessCookies processCookies
WARNING: Cookie rejected: "[version: 0][name: rememberMe][value: deleteMe][domain: repository.mapr.com][path: /nexus][expiry:
th attribute "/nexus". Path of origin: "/maven/org/apache/apache/9/apache-9.pom"
Downloading: https://repo.maven.apache.org/maven2/org/apache/apache/9/apache-9.pom
Downloaded: https://repo.maven.apache.org/maven2/org/apache/apache/9/apache-9.pom (15 KB at 8.1 KB/sec)
[INFO] ------------------------------------------------------------------------
[INFO] Reactor Build Order:
[INFO]
[INFO] Mahout Build Tools
[INFO] Apache Mahout
[INFO] Mahout Math
[INFO] Mahout MapReduce Legacy
[INFO] Mahout Integration
[INFO] Mahout Examples
[INFO] Mahout Release Package
[INFO] Mahout Math Scala bindings
[INFO] Mahout Spark bindings
[INFO] Mahout Spark bindings shell
[INFO] Mahout H2O backend
[INFO]
[INFO] ------------------------------------------------------------------------
[INFO] Building Mahout Build Tools 1.0-SNAPSHOT
[INFO] ------------------------------------------------------------------------
Downloading: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-remote-resources-plugin/1.1/maven-remote-reso
Downloaded: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-remote-resources-plugin/1.1/maven-remote-resou
Downloading: https://repo.maven.apache.org/maven2/org/apache/maven/plugins/maven-plugins/14/maven-plugins-14.pom
```

```
spark-1.1.1]# sbin/start-all.sh
starting org.apache.spark.deploy.master.Master, logging to /tmp/spark/spark-1.1.
1/sbin/../logs/spark-root-org.apache.spark.deploy.master.Master-1-sandbox.horton
works.com.out
localhost: starting org.apache.spark.deploy.worker.Worker, logging to /tmp/spark
/spark-1.1.1/sbin/../logs/spark-root-org.apache.spark.deploy.worker.Worker-1-san
dbox.hortonworks.com.out
```

Index of /publiccorpus

Name	Last modified	Size	Description
Parent Directory		-	
20021010_easy_ham.tar.bz2	2004-06-29 03:26	1.6M	
20021010_hard_ham.tar.bz2	2004-12-16 19:49	1.0M	
20021010_spam.tar.bz2	2004-06-29 03:26	1.1M	
20030228_easy_ham.tar.bz2	2004-06-29 03:26	1.5M	
20030228_easy_ham_2.tar.bz2	2004-06-29 03:26	1.0M	
20030228_hard_ham.tar.bz2	2004-12-16 19:49	1.0M	
20030228_spam.tar.bz2	2004-06-29 03:26	1.1M	
20030228_spam_2.tar.bz2	2004-06-29 03:26	2.0M	
20050311_spam_2.tar.bz2	2005-03-11 23:55	2.0M	
obsolete/	2014-02-04 16:26	-	
readme.html	2006-01-31 20:30	4.5K	

Apache/2.4.10 (Unix) OpenSSL/1.0.1i Server at spamassassin.apache.org Port 80



```
mahout seqdirectory -i /user/hue/assassin/dataset -o /user/hue/assassinseq-out
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 01:04:24 INFO common.AbstractJob: Command line arguments: {--charset=[UTF-8], --chunkSize=[64], --endPhase=[2147483647], --fileFilterClass=[org.
ext.PrefixAdditionFilter], --input=[/user/hue/assassin/dataset], --keyPrefix=[], --method=[mapreduce], --output=[/user/hue/assassinseq-out], --startPhase
=[temp]}
14/12/07 01:04:26 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
14/12/07 01:04:26 INFO Configuration.deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
14/12/07 01:04:26 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
14/12/07 01:04:30 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/12/07 01:04:37 INFO input.FileInputFormat: Total input paths to process : 3052
14/12/07 01:04:38 INFO input.CombineFileInputFormat: DEBUG: Terminated node allocation with : CompletedNodes: 1, size left: 12664944
14/12/07 01:04:39 INFO mapreduce.JobSubmitter: number of splits:1
14/12/07 01:04:40 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1417937528905_0001
14/12/07 01:04:41 INFO impl.YarnClientImpl: Submitted application application_1417937528905_0001
14/12/07 01:04:42 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1417937528905_0001/
14/12/07 01:04:42 INFO mapreduce.Job: Running job: job_1417937528905_0001
14/12/07 01:05:23 INFO mapreduce.Job: Job job_1417937528905_0001 running in uber mode : false
14/12/07 01:05:23 INFO mapreduce.Job:  map 0% reduce 0%
14/12/07 01:05:58 INFO mapreduce.Job:  map 6% reduce 0%
14/12/07 01:06:01 INFO mapreduce.Job:  map 13% reduce 0%
14/12/07 01:06:04 INFO mapreduce.Job:  map 22% reduce 0%
```

```
mahout seq2sparse -i /user/hue/assassinseq-out/part-m-00000 -o /user/hue/assassinvec -lnorm -nv -wt tfidf
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 01:32:37 INFO vectorizer.SparseVectorsFromSequenceFiles: Maximum n-gram size is: 1
14/12/07 01:32:37 INFO vectorizer.SparseVectorsFromSequenceFiles: Minimum LLR value: 1.0
14/12/07 01:32:37 INFO vectorizer.SparseVectorsFromSequenceFiles: Number of reduce tasks: 1
14/12/07 01:32:37 INFO vectorizer.SparseVectorsFromSequenceFiles: Tokenizing documents in /user/hue/assassinseq-out/part-m-00000
14/12/07 01:32:42 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/12/07 01:32:50 INFO input.FileInputFormat: Total input paths to process : 1
14/12/07 01:32:50 INFO mapreduce.JobSubmitter: number of splits:1
14/12/07 01:32:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1417937528905_0002
14/12/07 01:32:52 INFO impl.YarnClientImpl: Submitted application application_1417937528905_0002
14/12/07 01:32:52 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1417937528905_0002/
14/12/07 01:32:52 INFO mapreduce.Job: Running job: job_1417937528905_0002
14/12/07 01:33:25 INFO mapreduce.Job: Job job_1417937528905_0002 running in uber mode : false
14/12/07 01:33:25 INFO mapreduce.Job:  map 0% reduce 0%
14/12/07 01:33:52 INFO mapreduce.Job:  map 100% reduce 0%
14/12/07 01:33:55 INFO mapreduce.Job: Job job_1417937528905_0002 completed successfully
14/12/07 01:33:56 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=99468
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3309809
                HDFS: Number of bytes written=10717543
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=23700
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=23700
                Total vcore-seconds taken by all map tasks=23700
                Total megabyte-seconds taken by all map tasks=5925000
        Map-Reduce Framework
                Map input records=3052
```

```
mahout split -i /user/hue/assassinvec/tfidf-vectors --trainingOutput /user/hue/assassindatatrain --testOutput /user/hue/assassindatatest --randomSelectionPct 20 --overw
rite --sequenceFiles -xm sequential
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 02:04:13 WARN driver.MahoutDriver: No split.props found on classpath, will use command-line arguments only
14/12/07 02:04:14 INFO common.AbstractJob: Command line arguments: {--endPhase=[2147483647], --input=[/user/hue/assassinvec/tfidf-vectors], --method=[sequential], --ove
rwrite=null, --randomSelectionPct=[20], --sequenceFiles=null, --startPhase=[0], --tempDir=[temp], --testOutput=[/user/hue/assassindatatest], --trainingOutput=[/user/hue
/assassindatatrain]}
14/12/07 02:04:21 INFO utils.SplitInput: part-r-00000 has 50551 lines
14/12/07 02:04:21 INFO utils.SplitInput: part-r-00000 test split size is 10110 based on random selection percentage 20
14/12/07 02:04:21 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
14/12/07 02:04:21 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/12/07 02:04:21 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/12/07 02:04:23 INFO utils.SplitInput: file: part-r-00000, input: 50551 train: 2416, test: 636 starting at 0
14/12/07 02:04:23 INFO driver.MahoutDriver: Program took 9987 ms (Minutes: 0.16645)
```

```
mahout trainnb -i /user/hue/assassindatatrain -el -o /user/hue/prodmodel -li /user/hue/prodlabelindex -ow -c
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 02:12:57 WARN driver.MahoutDriver: No trainnb.props found on classpath, will use command-line arguments only
14/12/07 02:12:58 INFO common.AbstractJob: Command line arguments: {--alphaI=[1.0], --endPhase=[2147483647], --extractLabels=null, --input=[/user/hue/assassindatatrain]
, --labelIndex=[/user/hue/prodlabelindex], --output=[/user/hue/prodmodel], --overwrite=null, --startPhase=[0], --tempDir=[temp], --trainComplementary=null}
14/12/07 02:13:04 INFO common.HadoopUtil: Deleting temp
14/12/07 02:13:05 INFO zlib.ZlibFactory: Successfully loaded & initialized native-zlib library
14/12/07 02:13:05 INFO compress.CodecPool: Got brand-new decompressor [.deflate]
14/12/07 02:13:06 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
14/12/07 02:13:06 INFO Configuration.deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
14/12/07 02:13:06 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdir
14/12/07 02:13:07 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/12/07 02:13:10 INFO input.FileInputFormat: Total input paths to process : 1
14/12/07 02:13:11 INFO mapreduce.JobSubmitter: number of splits:1
14/12/07 02:13:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1417937528905_0011
14/12/07 02:13:12 INFO impl.YarnClientImpl: Submitted application application_1417937528905_0011
14/12/07 02:13:12 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1417937528905_0011/
14/12/07 02:13:12 INFO mapreduce.Job: Running job: job_1417937528905_0011
14/12/07 02:13:41 INFO mapreduce.Job: Job job_1417937528905_0011 running in uber mode : false
14/12/07 02:13:41 INFO mapreduce.Job:  map 0% reduce 0%
14/12/07 02:14:03 INFO mapreduce.Job:  map 100% reduce 0%
14/12/07 02:14:24 INFO mapreduce.Job:  map 100% reduce 100%
14/12/07 02:14:24 INFO mapreduce.Job: Job job_1417937528905_0011 completed successfully
14/12/07 02:14:25 INFO mapreduce.Job: Counters: 49
        File System Counters
```

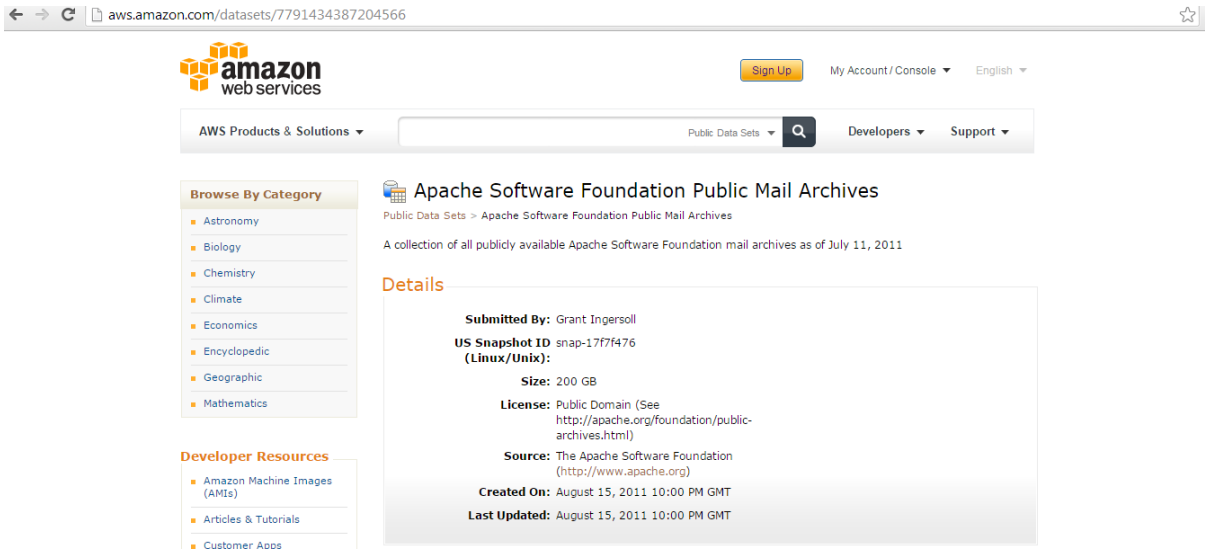
```
14/12/07 02:22:00 INFO test.TestNaiveBayesDriver: Standard NB Results:
=======================================================
Summary
-------------------------------------------------------
Correctly Classified Instances          :        633        99.5283%
Incorrectly Classified Instances        :          3         0.4717%
Total Classified Instances              :        636

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       <--Classified as
525     3        |  528          a     = easy_ham
0       108      |  108          b     = spam

=======================================================
Statistics
-------------------------------------------------------
Kappa                                       0.9678
Accuracy                                   99.5283%
Reliability                                66.4773%
Reliability (standard deviation)            0.5757

14/12/07 02:22:00 INFO driver.MahoutDriver: Program took 47941 ms (Minutes: 0.7990333333333334)
```

```
mv cmds 00000.7b1b73cf36cf9dbc3d64e3f2ee2b91f1
```

```
java -cp /tmp/assassinmodeltest/spamclassifier.jar:/usr/lib/mahout/* com.packt.spamfilter.TestClassifier /tmp/assassinmodeltest/ /tmp/assassinmodeltest/prodlabelindex
/tmp/assassinmodeltest/dictionary.file-0 /tmp/assassinmodeltest/df-count /tmp/assassinmodeltest/testemail
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.lib.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Number of labels: 2
Number of documents in training set: 3052
  easy_ham: -91.2525617072814  spam: -60.78830055665665 => spam
```

```
Delivered-To: exmh-workers@listman.example.com
```

```
java -cp /tmp/assassinmodeltest/spamclassifier.jar:/usr/lib/mahout/* com.packt.spamfilter.TestClassifier /tmp/assassinmodeltest/ /tmp/assassinmodeltest/prodlabelindex
/tmp/assassinmodeltest/dictionary.file-0 /tmp/assassinmodeltest/df-count /tmp/assassinmodeltest/testemail
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.lib.MutableMetricsFactory).
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.
Number of labels: 2
Number of documents in training set: 3052
  easy_ham: -88.32010546373571  spam: -159.80191214024765 => easy_ham
```



```
mahout org.apache.mahout.text.SequenceFilesFromMailArchives --charset "UTF-8" --body --subject --input /user/hue/asfmail/content --output
hue/asfmailout
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 09:34:05 WARN driver.MahoutDriver: No org.apache.mahout.text.SequenceFilesFromMailArchives.props found on classpath, will use com
14/12/07 09:34:08 INFO common.AbstractJob: Command line arguments: {--body=null, --bodySeparator=[
], --charset=[UTF-8], --chunkSize=[64], --endPhase=[2147483647], --input=[/user/hue/asfmail/content], --keyPrefix=[], --method=[mapreduce]
ut], --separator=[
], --startPhase=[0], --subject=null, --tempDir=[temp]}
14/12/07 09:34:10 INFO Configuration.deprecation: mapred.input.dir is deprecated. Instead, use mapreduce.input.fileinputformat.inputdir
14/12/07 09:34:10 INFO Configuration.deprecation: mapred.compress.map.output is deprecated. Instead, use mapreduce.map.output.compress
14/12/07 09:34:10 INFO Configuration.deprecation: mapred.output.dir is deprecated. Instead, use mapreduce.output.fileoutputformat.outputdi
14/12/07 09:34:18 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/12/07 09:34:25 INFO input.FileInputFormat: Total input paths to process : 573
14/12/07 09:34:25 INFO input.CombineFileInputFormat: DEBUG: Terminated node allocation with : CompletedNodes: 1, size left: 65487674
14/12/07 09:34:25 INFO mapreduce.JobSubmitter: number of splits:5
```

```
mahout seq2sparse --input /user/hue/asfmailout --output /user/hue/asfmailseqsp --norm 2 --weight TFIDF --namedVector --maxDFPercent 90 --minSupport 2 --analyzerName
org.apache.mahout.text.MailArchivesClusteringAnalyzer
MAHOUT_LOCAL is not set; adding HADOOP_CONF_DIR to classpath.
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/hadoop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.9.0.2.1.1.0-385-job.jar
14/12/07 09:59:16 INFO vectorizer.SparseVectorsFromSequenceFiles: Maximum n-gram size is: 1
14/12/07 09:59:16 INFO vectorizer.SparseVectorsFromSequenceFiles: Minimum LLR value: 1.0
14/12/07 09:59:16 INFO vectorizer.SparseVectorsFromSequenceFiles: Number of reduce tasks: 1
14/12/07 09:59:16 INFO vectorizer.SparseVectorsFromSequenceFiles: Tokenizing documents in /user/hue/asfmailout
14/12/07 09:59:21 INFO client.RMProxy: Connecting to ResourceManager at sandbox.hortonworks.com/10.0.2.15:8050
14/12/07 09:59:32 INFO input.FileInputFormat: Total input paths to process : 5
14/12/07 09:59:32 INFO mapreduce.JobSubmitter: number of splits:37
14/12/07 09:59:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1417968166805_0002
14/12/07 09:59:34 INFO impl.YarnClientImpl: Submitted application application_1417968166805_0002
14/12/07 09:59:34 INFO mapreduce.Job: The url to track the job: http://sandbox.hortonworks.com:8088/proxy/application_1417968166805_0002/
14/12/07 09:59:34 INFO mapreduce.Job: Running job: job_1417968166805_0002
14/12/07 10:00:05 INFO mapreduce.Job: Job job_1417968166805_0002 running in uber mode : false
14/12/07 10:00:05 INFO mapreduce.Job:  map 0% reduce 0%
14/12/07 10:06:47 INFO mapreduce.Job:  map 8% reduce 0%
```