<http://www.tuicool.com/articles/FbQ7Vb6>

训练

mahout trainlogistic --input donut.csv --output model2 --target color --categories 2 --predictors x y a b c --types numeric --features 20 --passes 100 --rate 50

评估

mahout runlogistic --input donut-test.csv --model model2 --scores --auc –confusion

版本：mahout0.9

Mahout里面使用逻辑回归（logistic regression）的主要两个类是org.apache.mahout.classifier.sgd.TrainLogistic、org.apache.mahout.classifier.sgd.RunLogistic，一个是建立模型，一个是进行模型评估。

首先是原始数据，格式如下：（可以在https://github.com/dirkweissenborn/mahout-rbmClassifier/blob/master/examples/src/main/resources/donut.csv#L1下载）

"x","y","shape","color","k","k0","xx","xy","yy","a","b","c","bias"

0.923307513352484,0.0135197141207755,21,2,4,8,0.852496764213146,0.0124828536260896,0.000182782669907495,0.923406490600458,0.0778750292332978,0.644866125183976,1

0.711011884035543,0.909141522599384,22,2,3,9,0.505537899239772,0.64641042683833,0.826538308114327,1.15415605849213,0.953966686673604,0.46035073663368,1

0.75118898646906,0.836567111080512,23,2,3,9,0.564284893392414,0.62842000028592,0.699844531341594,1.12433510339845,0.872783737128441,0.419968245447719,1

进入mahout的bin目录，运行：

./mahout trainlogistic --input /data/mahout-data/donut.csv --output /data/mahout-output/model2 --target color --categories 2 --predictors x y a b c --types numeric --features 20 --passes 100 --rate 50

这里各个参数说明如下：

input：输入数据；output：输出模型文件；--target 预测的变量（输入数据要求第一行为变量名称）；categories 预测变量的取值个数；predictors参与建模的变量；types 预测变量的类型（number、word、text其中一个，如果全部是一样的话，使用一个就可以）；pass训练的时候对输入数据测试的次数（这里也不是很清楚）；feature内部随机向量维度（用于建模，好像是这样理解，越大越好，但是时间会长 ）；rate学习速率（如果输入数据比较大，此值可以设置大点）。

得到下面的输出：

Running on hadoop, using /opt/hadoop2/bin/hadoop **and** HADOOP\_CONF\_DIR=

MAHOUT-JOB: /opt/mahout-distribution-0.9/examples/target/mahout-examples-0.9-job.jar

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding **in** [jar:file:/opt/hadoop2/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.**class**]

SLF4J: Found binding **in** [jar:file:/opt/hadoop2/share/hadoop/mapreduce/lib/mahout-core-0.9-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]

20

color ~

7.068\*Intercept Term + 0.581\*a + -1.369\*b + -25.059\*c + 0.581\*x + 2.319\*y

Intercept Term 7.06759

a 0.58123

b -1.36893

c -25.05945

x 0.58123

y 2.31879

0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 -1.368933989 0.000000000 0.000000000 0.000000000 0.000000000 0.581234210 0.000000000 0.000000000 7.067587159 0.000000000 0.000000000 0.000000000 2.318786209 0.000000000 -25.059452292

14/04/11 10:33:18 INFO driver.MahoutDriver: Program took 1758 ms (Minutes: 0.0293)

我这里有slf jar包的冲突，暂时不理这个。看后面的公式即可（公式变量前的值，每次训练不一定相同），应该是由这个公式算得最后的预测结果的，但是暂时不清楚Intercept是什么。

然后使用模型评估命令（测试数据：https://svn.apache.org/repos/asf/mahout/trunk/examples/src/main/resources/donut-test.csv）：

./mahout runlogistic --input /data/mahout-data/donut-test.csv --model /data/mahout-output/model2 --scores --auc --confusion

input就是测试数据；model是模型文件；scores打印预测值和原始值对比；auc打印auc值（评判主要标准，越大越好，最好接近1）；confusion打印模糊矩阵；

得到下面的结果：

Running on hadoop, using /opt/hadoop2/bin/hadoop **and** HADOOP\_CONF\_DIR=

MAHOUT-JOB: /opt/mahout-distribution-0.9/examples/target/mahout-examples-0.9-job.jar

SLF4J: Class path contains multiple SLF4J bindings.

SLF4J: Found binding **in** [jar:file:/opt/hadoop2/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.**class**]

SLF4J: Found binding **in** [jar:file:/opt/hadoop2/share/hadoop/mapreduce/lib/mahout-core-0.9-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]

"target","model-output","log-likelihood"

0,0.009,-0.009241

0,0.000,-0.000481

1,0.985,-0.015038

1,0.991,-0.009407

0,0.001,-0.000883

1,0.974,-0.026000

1,0.823,-0.194875

0,0.041,-0.042015

0,0.051,-0.052565

0,0.613,-0.950008

0,0.147,-0.158538

1,0.910,-0.094177

1,0.252,-1.377220

1,0.924,-0.078521

1,0.998,-0.001777

0,0.023,-0.023756

1,0.990,-0.009928

0,0.003,-0.003118

1,0.961,-0.039284

0,0.000,-0.000046

0,0.167,-0.183160

0,0.049,-0.049822

0,0.006,-0.005792

0,0.706,-1.222487

0,0.000,-0.000421

1,0.999,-0.001045

1,0.969,-0.031452

0,0.034,-0.034088

0,0.370,-0.461632

0,0.011,-0.011489

0,0.465,-0.624971

0,0.053,-0.054646

0,0.340,-0.414959

0,0.053,-0.054123

0,0.007,-0.006800

0,0.248,-0.285650

1,0.482,-0.728835

0,0.781,-1.516960

0,0.024,-0.023975

0,0.022,-0.022281

AUC = 0.97

confusion: [[24.0, 2.0], [3.0, 11.0]]

entropy: [[-0.2, -2.8], [-4.1, -0.1]]

14/04/11 10:43:39 INFO driver.MahoutDriver: Program took 414 ms (Minutes: 0.0069)

可以看到auc=0.97 说明模型还是比较好的；模糊矩阵中说明 有2个应该被分为1的被分为了0，有3个应该是0的结果被分为了1。

本来打算使用上面得到的公式带入测试数据，看能否得到第一行的输出，比如0.009，但是不知道哪个Interceptor值是什么，所以也是没有得到0.009的。大概浏览了下源码，好像要归一化的。具体下次在分析。

总结：

     目前遇到的问题有：1）如何使用上面的公式（Interceptor是什么？）；2）如何把这个在hadoop上面运行起来（从上面的结果来看，似乎mahout并没有运行在hadoop上面）。