Universidad de Granada

Cloud Computing: Servicios y Aplicaciones

Ciencia de Datos con Hadoop

Marvin Matías Agüero Torales

maguero@correo.ugr.es

Curso 2016-2017

Sumario

Objetivo

El objetivo de esta práctica es conocer las alternativas para realizar experimentaciones de Ciencia de Datos. Para ello, haremos uso del entorno que se ha convertido en un estándar de facto como es Hadoop, utilizando HDFS como sistema de archivos distribuido y Hadoop-MapReduce como mecanismo de ejecución. Por último, aplicaremos la biblioteca Mahout para lanzar algoritmos de clasificación sobre conjuntos tipo Big Data.

Enunciado

Para realizar nuestras pruebas, nos basaremos en el problema “Heart” del repositorio UCI, pero donde muchos ejemplos se han replicado de manera aleatoria con un modelo gaussiano. El conjunto de datos se encuentra en el directorio hdfs://user/ahilario/datasets/BNG_heart/

Adicionalmente, se ha puesto disponible un fichero JAR simplificado de mahout, que se puede encontrar en el directorio /tmp/mahout-distribution-sige.jar

Tareas

Las tareas a realizar serán las siguientes:

1. Ejecutar el algoritmo “Random Forest” sobre el conjunto de datos BNG_heart y comprueba el rendimiento alcanzado de acuerdo a los siguientes casos:

    a. Número de Maps:

        i. 64, 128, 256

    b. Número de árboles:

        i. 10, 100, 1000 árboles

2. Del punto anterior, obtener una tabla que indique los siguientes datos:

    a. Características del modelo: número de nodos (total y promedio), profundidad máxima del árbol.

    b. Tiempo de ejecución para entrenamiento.

    c. Medidas de calidad Accuracy estándar y media geométrica tanto para la partición de entrenamiento como para test.

Resultados

Se ha creado un script, README.P5.sh, al cuál se le pasan los parámetros (en este orden: n.º de maps, n.º de trees, ejecución inical [Y|N], usuario) por consola y nos muestra la salida al finalizar, la matriz de confusión tanto del conjunto de test, como el de train (ver Apartado *Salida*):

Por ejemplo

```
./README.P5 64 10 Y mcc4423998
```

Y luego

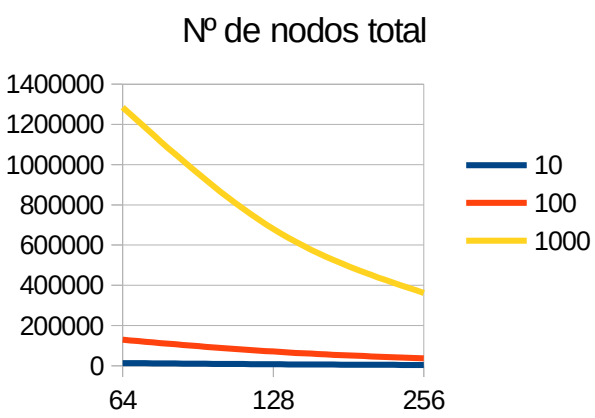
./README.P5 64 100 N mcc4423998

Al hacer la ejecución, si entrar en muchos detalles, cuando aumenta el número de maps, disminuye el tiempo. En un valor de map dado, el número de árboles, aumenta el tiempo. El accuracy mejora con el número de árboles, aunque, hay mejores resultados con 64 maps.

A continuación la tabla de resultados.

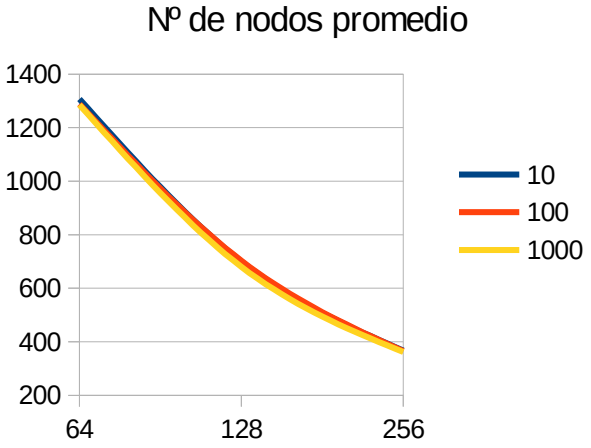
N.º de nodos total

	10	100	1000
64	13076	128888	1285123
128	7009	70338	680245
256	3674	36516	362097



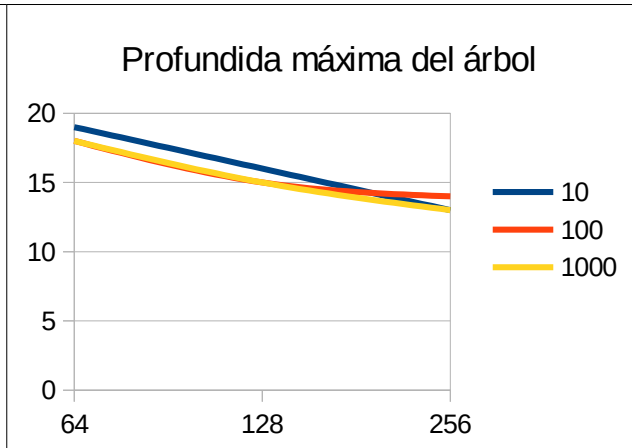
N.º de nodos promedio

	10	100	1000
64	1307	1288	1285
128	700	703	680
256	367	365	362



Profundidad máxima del árbol

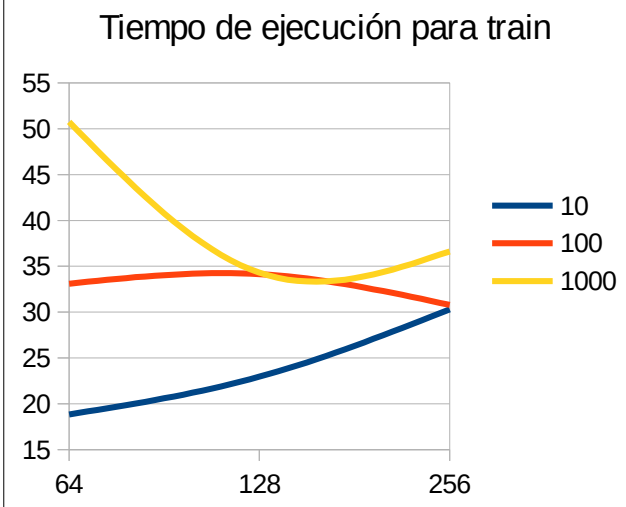
	10	100	1000
64	19	18	18
128	16	15	15
256	13	14	13



Train

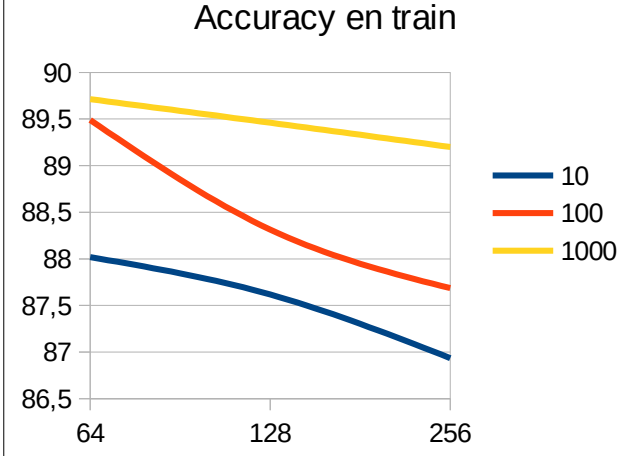
Tiempo de ejecución

	10	100	1000
64	18,845	33,1	50,761
128	22,95	34,141	34,298
256	30,293	30,783	36,627



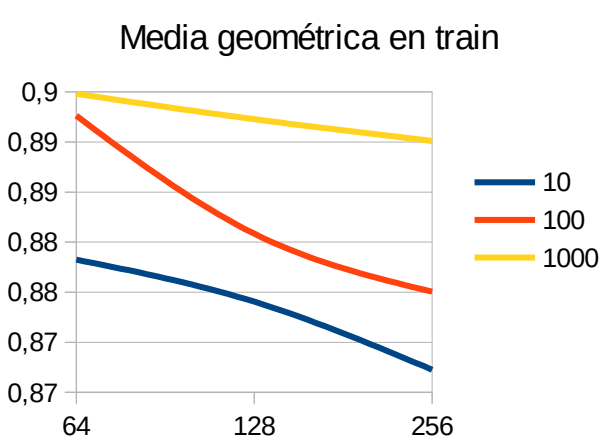
Accuracy estándar

	10	100	1000
64	88,0186	89,4899	89,713
128	87,6182	88,3144	89,4601
256	86,9358	87,6859	89,2004



Media geométrica

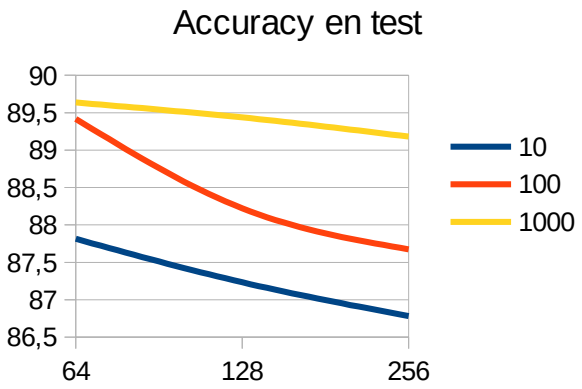
	10	100	1000
64	0,87824741	0,89262076	0,89480988
128	0,87405181	0,88084173	0,8922743
256	0,86725508	0,87504179	0,89010571



Test

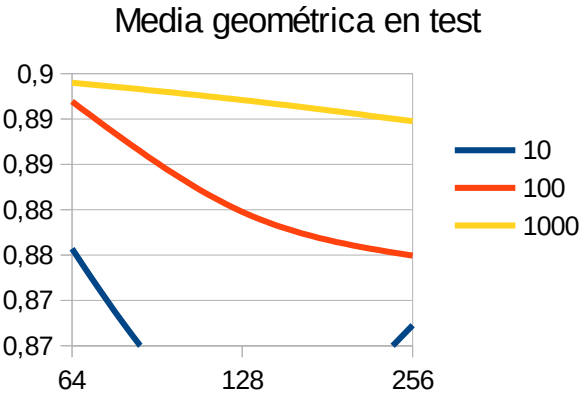
Accuracy estándar

	10	100	1000
64	87,8145	89,418	89,636
128	87,2335	88,222	89,437
256	86,782	87,674	89,181



Media geométrica

	10	100	1000
64	0,8757107	0,89194536	0,89400502
128	0,85567319	0,87977481	0,89209519
256	0,8672751	0,87494579	0,88975728



De la tabla de resultados se desprende lo siguiente:

El número de nodos total depende del número de árboles, a razón de mayor número de árboles, mayor número de nodos, como también del número de maps, a razón de mayor número de maps, menor número de nodos. Sin embargo, para el número de nodos promedio, así como la profundidad máxima del árbol, sólo hay dependencia con el número de maps.

Para el tiempo, actúa de forma distinta en dependencia del número de árboles, influyendo este en el resultado: cuanto mayor sea, arrojará un mejor resultado. Cabe destacar, que un mayor número de maps, influye de manera positiva.

Los resultados que se han obtenido sobre el conjunto de entrenamiento, son un poco mejores a los obtenidos con el conjunto de test, pero muy levemente, lo mismo ocurre con la media geométrica entre test y train. Los mejores resultados se obtienen con maps de 64.

Podemos concluir que Hadoop puede manejar grandes conjuntos de datos con facilidad, y Mahout, valiéndose de éste, puede procesar grandes conjuntos de datos y otorgar varios algoritmos, como Random Forests, incluso algoritmos de recomendación, como los utilizados en web como Amazon o Facebook, jugando un poco con Mahout y Hadoop, realicé unos experimentos (ver Anexos, Motor de recomendación Mahout en Hadoop).

Salida

Map	Tree	Output
64	10	sh ~/mahout/README.P5 64 10 Y mcc4423998 … 17/06/15 12:48:37 INFO mapreduce.BuildForest: Build Time: 0h 0m 18s 845 17/06/15 12:48:37 INFO mapreduce.BuildForest: Forest num Nodes: 13076 17/06/15 12:48:37 INFO mapreduce.BuildForest: Forest mean num Nodes: 1307 17/06/15 12:48:37 INFO mapreduce.BuildForest: Forest mean max Depth: 19 … Summary --- Correctly Classified Instances : 175629 87.8145% Incorrectly Classified Instances : 24371 12.1855% Total Classified Instances : 200000 === Confusion Matrix --- a b <--Classified as 76006 12762 | 88768 a = 1 11609 99623 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 76006 12762 11609 99623 --- Sensisivity or True Positive Rate (TPR) 0.85623198 --- Specificity or True Negative Rate (TNR) 0.89563255 --- AUC - Area Under the Curve ROC 0.87593226 --- GM - Geometric Mean 0.8757107 --- FM - F-Measure 0.86182909 --- Summary --- Correctly Classified Instances : 704149 88.0186% Incorrectly Classified Instances : 95851 11.9814% Total Classified Instances : 800000 === Confusion Matrix --- a b <--Classified as 306384 48902 | 355286 a = 1 46949 397765 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated.

		Instead use the hdfs command for it. #=== Confusion Matrix --- 306384 48902 46949 397765 --- Sensisivity or True Positive Rate (TPR) 0.86235878 --- Specificity or True Negative Rate (TNR) 0.89442878 --- AUC - Area Under the Curve ROC 0.87839378 --- GM - Geometric Mean 0.87824741 --- FM - F-Measure 0.86473549 ---
64	100	sh ~/mahout/README.P5 64 100 N … 17/06/15 13:26:02 INFO mapreduce.BuildForest: Build Time: 0h 0m 33s 100 17/06/15 13:26:02 INFO mapreduce.BuildForest: Forest num Nodes: 128888 17/06/15 13:26:02 INFO mapreduce.BuildForest: Forest mean num Nodes: 1288 17/06/15 13:26:03 INFO mapreduce.BuildForest: Forest mean max Depth: 18 … Summary --- Correctly Classified Instances : 178836 89.418% Incorrectly Classified Instances : 21164 10.582% Total Classified Instances : 200000 === Confusion Matrix --- a b <--Classified as 77574 11194 | 88768 a = 1 9970 101262 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 77574 11194 9970 101262 --- Sensisivity or True Positive Rate (TPR) 0.873896 --- Specificity or True Negative Rate (TNR) 0.91036752 --- AUC - Area Under the Curve ROC 0.89213176 --- GM - Geometric Mean 0.89194536 --- FM - F-Measure 0.87996279 --- Summary --- Correctly Classified Instances : 715919 89.4899% Incorrectly Classified Instances : 84081 10.5101% Total Classified Instances : 800000 === Confusion Matrix --- a b <--Classified as 310583 44703 | 355286 a = 1 39378 405336 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 310583 44703

		39378 405336 --- Sensisivity or True Positive Rate (TPR) 0.87417742 --- Specificity or True Negative Rate (TNR) 0.9114532 --- AUC - Area Under the Curve ROC 0.89281531 --- GM - Geometric Mean 0.89262076 --- FM - F-Measure 0.88077794 ---
64	1000	sh ~/mahout/README.P5 64 1000 N … 17/06/15 13:28:29 INFO mapreduce.BuildForest: Build Time: 0h 0m 50s 761 17/06/15 13:28:29 INFO mapreduce.BuildForest: Forest num Nodes: 1285123 17/06/15 13:28:29 INFO mapreduce.BuildForest: Forest mean num Nodes: 1285 17/06/15 13:28:29 INFO mapreduce.BuildForest: Forest mean max Depth: 18 … Summary --- Correctly Classified Instances : 179272 89.636% Incorrectly Classified Instances : 20728 10.364% Total Classified Instances : 200000 ======= Confusion Matrix --- a b <--Classified as 77678 11090 | 88768 a = 1 9638 101594 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 77678 11090 9638 101594 --- Sensisivity or True Positive Rate (TPR) 0.87506759 --- Specificity or True Negative Rate (TNR) 0.91335227 --- AUC - Area Under the Curve ROC 0.89420993 --- GM - Geometric Mean 0.89400502 --- FM - F-Measure 0.88228346 --- Summary --- Correctly Classified Instances : 717704 89.713% Incorrectly Classified Instances : 82296 10.287% Total Classified Instances : 800000 ======= Confusion Matrix --- a b <--Classified as 311249 44037 | 355286 a = 1 38259 406455 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 311249 44037 38259 406455 --- Sensisivity or True Positive Rate (TPR) 0.87605197 --- Specificity or True Negative Rate (TNR)

		0.91396943 --- AUC - Area Under the Curve ROC 0.8950107 --- GM - Geometric Mean 0.89480988 --- FM - F-Measure 0.88323397 ---
128	10	sh ~/mahout/README.P5 128 10 N … 17/06/15 13:32:26 INFO mapreduce.BuildForest: Build Time: 0h 0m 22s 95 17/06/15 13:32:26 INFO mapreduce.BuildForest: Forest num Nodes: 7009 17/06/15 13:32:26 INFO mapreduce.BuildForest: Forest mean num Nodes: 700 17/06/15 13:32:26 INFO mapreduce.BuildForest: Forest mean max Depth: 16 … Summary --- Correctly Classified Instances : 174467 87.2335% Incorrectly Classified Instances : 25533 12.7665% Total Classified Instances : 200000 ======= Confusion Matrix --- a b <--Classified as 75689 13079 | 88768 a = 1 12454 98778 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 75689 13079 12454 98778 --- Sensisivity or True Positive Rate (TPR) 0.85266087 --- Specificity or True Negative Rate (TNR) 0.88803582 --- AUC - Area Under the Curve ROC 0.87034834 --- GM - Geometric Mean 0.8701686 --- FM - F-Measure 0.85567319 --- Summary --- Correctly Classified Instances : 700946 87.6182% Incorrectly Classified Instances : 99054 12.3818% Total Classified Instances : 800000 ======= Confusion Matrix --- a b <--Classified as 304385 50901 | 355286 a = 1 48153 396561 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 304385 50901 48153 396561 --- Sensisivity or True Positive Rate (TPR) 0.85673232 --- Specificity or True Negative Rate (TNR) 0.89172142 --- AUC - Area Under the Curve ROC 0.87422687 --- GM - Geometric Mean

		0.87405181 --- FM - F-Measure 0.86005843 ---
128	100	sh ~/mahout/README.P5 128 100 N … 17/06/15 13:34:30 INFO mapreduce.BuildForest: Build Time: 0h 0m 34s 141 17/06/15 13:34:30 INFO mapreduce.BuildForest: Forest num Nodes: 70338 17/06/15 13:34:30 INFO mapreduce.BuildForest: Forest mean num Nodes: 703 17/06/15 13:34:30 INFO mapreduce.BuildForest: Forest mean max Depth: 15 … Summary --- Correctly Classified Instances : 176444 88.222% Incorrectly Classified Instances : 23556 11.778% Total Classified Instances : 200000 ======= Confusion Matrix --- a b <--Classified as 76359 12409 | 88768 a = 1 11147 100085 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 76359 12409 11147 100085 --- Sensisivity or True Positive Rate (TPR) 0.86020863 --- Specificity or True Negative Rate (TNR) 0.89978603 --- AUC - Area Under the Curve ROC 0.87999733 --- GM - Geometric Mean 0.87977481 --- FM - F-Measure 0.86636713 --- Summary --- Correctly Classified Instances : 706515 88.3144% Incorrectly Classified Instances : 93485 11.6856% Total Classified Instances : 800000 ======= Confusion Matrix --- a b <--Classified as 306343 48943 | 355286 a = 1 44542 400172 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 306343 48943 44542 400172 --- Sensisivity or True Positive Rate (TPR) 0.86224338 --- Specificity or True Negative Rate (TNR) 0.89984125 --- AUC - Area Under the Curve ROC 0.88104231 --- GM - Geometric Mean 0.88084173 --- FM - F-Measure 0.86761705 ---
128	1000	sh ~/mahout/README.P5 128 1000 N

		… 17/06/15 13:39:33 INFO mapreduce.BuildForest: Build Time: 0h 0m 34s 298 17/06/15 13:39:33 INFO mapreduce.BuildForest: Forest num Nodes: 680245 17/06/15 13:39:33 INFO mapreduce.BuildForest: Forest mean num Nodes: 680 17/06/15 13:39:33 INFO mapreduce.BuildForest: Forest mean max Depth: 15 … Summary --- Correctly Classified Instances : 178874 89.437% Incorrectly Classified Instances : 21126 10.563% Total Classified Instances : 200000 === Confusion Matrix --- a b <--Classified as 77561 11207 | 88768 a = 1 9919 101313 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 77561 11207 9919 101313 --- Sensisivity or True Positive Rate (TPR) 0.87374955 --- Specificity or True Negative Rate (TNR) 0.91082602 --- AUC - Area Under the Curve ROC 0.89228779 --- GM - Geometric Mean 0.89209519 --- FM - F-Measure 0.88013481 Summary --- Correctly Classified Instances : 715681 89.4601% Incorrectly Classified Instances : 84319 10.5399% Total Classified Instances : 800000 === Confusion Matrix --- a b <--Classified as 310332 44954 | 355286 a = 1 39365 405349 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 310332 44954 39365 405349 --- Sensisivity or True Positive Rate (TPR) 0.87347095 --- Specificity or True Negative Rate (TNR) 0.91148244 --- AUC - Area Under the Curve ROC 0.89247669 --- GM - Geometric Mean 0.8922743 --- FM - F-Measure 0.8803957 ---
256	10	sh ~/mahout/README.P5 256 10 N … 17/06/15 13:42:10 INFO mapreduce.BuildForest: Build Time: 0h 0m 30s 293 17/06/15 13:42:10 INFO mapreduce.BuildForest: Forest num Nodes: 3674 17/06/15 13:42:10 INFO mapreduce.BuildForest: Forest mean num Nodes: 367 17/06/15 13:42:10 INFO mapreduce.BuildForest: Forest mean max Depth: 13 …

		Summary --- Correctly Classified Instances : 173564 86.782% Incorrectly Classified Instances : 26436 13.218% Total Classified Instances : 200000 ======= Confusion Matrix --- a b <--Classified as 76567 12201 | 88768 a = 1 14235 96997 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 76567 12201 14235 96997 --- Sensisivity or True Positive Rate (TPR) 0.86255182 --- Specificity or True Negative Rate (TNR) 0.87202424 --- AUC - Area Under the Curve ROC 0.86728803 --- GM - Geometric Mean 0.8672751 --- FM - F-Measure 0.85278165 --- Summary --- Correctly Classified Instances : 695486 86.9358% Incorrectly Classified Instances : 104514 13.0642% Total Classified Instances : 800000 ======= Confusion Matrix --- a b <--Classified as 302049 53237 | 355286 a = 1 51277 393437 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======= Confusion Matrix --- 302049 53237 51277 393437 --- Sensisivity or True Positive Rate (TPR) 0.85015734 --- Specificity or True Negative Rate (TNR) 0.88469668 --- AUC - Area Under the Curve ROC 0.86742701 --- GM - Geometric Mean 0.86725508 --- FM - F-Measure 0.85250885 ---
256	100	sh ~/mahout/README.P5 256 100 N … 17/06/15 13:44:04 INFO mapreduce.BuildForest: Build Time: 0h 0m 30s 783 17/06/15 13:44:04 INFO mapreduce.BuildForest: Forest num Nodes: 36516 17/06/15 13:44:04 INFO mapreduce.BuildForest: Forest mean num Nodes: 365 17/06/15 13:44:04 INFO mapreduce.BuildForest: Forest mean max Depth: 14 … Summary --- Correctly Classified Instances : 175348 87.674% Incorrectly Classified Instances : 24652 12.326% Total Classified Instances : 200000

		=== Confusion Matrix --- a b <--Classified as 76360 12408 | 88768 a = 1 12244 98988 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 76360 12408 12244 98988 --- Sensisivity or True Positive Rate (TPR) 0.8602199 --- Specificity or True Negative Rate (TNR) 0.88992376 --- AUC - Area Under the Curve ROC 0.87507183 --- GM - Geometric Mean 0.87494579 --- FM - F-Measure 0.86101527 Confusion Matrix --- a b <--Classified as 305891 49395 | 355286 a = 1 49211 395503 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix --- 305891 49395 49211 395503 --- Sensisivity or True Positive Rate (TPR) 0.86097116 --- Specificity or True Negative Rate (TNR) 0.88934236 --- AUC - Area Under the Curve ROC 0.87515676 --- GM - Geometric Mean 0.87504179 --- FM - F-Measure 0.86119416 ---
256	1000	sh ~/mahout/README.P5 256 1000 N … 17/06/15 13:46:15 INFO mapreduce.BuildForest: Build Time: 0h 0m 36s 627 17/06/15 13:46:15 INFO mapreduce.BuildForest: Forest num Nodes: 362097 17/06/15 13:46:15 INFO mapreduce.BuildForest: Forest mean num Nodes: 362 17/06/15 13:46:15 INFO mapreduce.BuildForest: Forest mean max Depth: 13 … Summary --- Correctly Classified Instances : 178362 89.181% Incorrectly Classified Instances : 21638 10.819% Total Classified Instances : 200000 === Confusion Matrix --- a b <--Classified as 77500 11268 | 88768 a = 1 10370 100862 | 111232 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #=== Confusion Matrix

		---------- 77500 11268 10370 100862 ---------- Sensisivity or True Positive Rate (TPR) 0.87306236 ---------- Specificity or True Negative Rate (TNR) 0.90677143 ---------- AUC - Area Under the Curve ROC 0.8899169 ---------- GM - Geometric Mean 0.88975728 ---------- FM - F-Measure 0.87750088 Summary ---------- Correctly Classified Instances : 713603 89.2004% Incorrectly Classified Instances : 86397 10.7996% Total Classified Instances : 800000 ======== Confusion Matrix ---------- a b <--Classified as 310701 44585 | 355286 a = 1 41812 402902 | 444714 b = 0 DEPRECATED: Use of this script to execute hdfs command is deprecated. Instead use the hdfs command for it. #======== Confusion Matrix ---------- 310701 44585 41812 402902 ---------- Sensisivity or True Positive Rate (TPR) 0.87450955 ---------- Specificity or True Negative Rate (TNR) 0.90598002 ---------- AUC - Area Under the Curve ROC 0.89024479 ---------- GM - Geometric Mean 0.89010571 ---------- FM - F-Measure 0.87793569 ----------

Bibliografía consultada

Hilario, A., 2017. Guión de la Práctica 5: Ciencia de Datos con Hadoop, Universidad de Granada.

Anexos

Disponibles en https://github.com/mmaguero

Adjuntos

Se adjunta fichero README.P5 con el script que ejecuta las tareas de esta práctica, y Recommder.P5, con el motor de recomendación de Mahout sobre Hadoop.

Motor de recomendación Mahout en Hadoop

En este apartado se ejecuta el motor de recomendación de Mahout[1] en un conjunto de datos de calificaciones de películas (MovieLens[2]) y se muestran las recomendaciones de películas para cada usuario.

Como los requisitos son los mismos utilizados en esta práctica, no tendremos problemas con ello (aunque se recomienda descargar la distribucion completa de Mahout[3]):

Java → Hadoop → Mahout → Recommender

El motor recommender acepta cualquier archivo que contenga un conjunto de líneas con el userId, el itemId y un valor de preferencia (opcional) separados por una pestaña. El userId y itemId deben ser un entero y el valor de preferencia puede ser un entero o un doble. Aspectos cumplidos por MovieLens, este archivo contiene:

- **U.data**: contiene varias tuplas (user_id, movie_id, rating, timestamp)
- **U.user**: contiene varias tuplas (user_id, edad, sexo, ocupación, zip_code)
- **U.item**: contiene varias tuplas (movie_id, title, release_date, video_release_data, imdb_url, cat_unknown, cat_action, cat_adventure, cat_animation, cat_children, cat_comedy, cat_crime, cat_documentary, cat_drama, cat_fantasy, cat_film_noir, cat_horror, cat_musical, cat_mystery, cat_romance, cat_sci_fi, Cat_thriller, cat_war, cat_western)

En total, el conjunto de datos contiene 943 usuarios, 1,682 películas y 100,000 calificaciones.

Con el argumento "-s SIMILARITY_COOCURRENCE", le decimos al recomendador qué tipo de fórmula utilizar, aquí, dos elementos (películas) son muy similares si a menudo aparecen juntos en la clasificación de los usuarios: así que para encontrar las películas para recomendar a un usuario, tenemos que encontrar las 10 películas más similares a las películas que el usuario ha calificado. Es decir, si un usuario A da una buena calificación en la película X y otros usuarios dan una buena calificación en la película X y la película Y, entonces podemos recomendar la película Y al usuario A.

Mahout calcula las recomendaciones ejecutando varios trabajos de mapreduce de Hadoop, en un tiempo récord los trabajos están terminados y cada usuario tendrá las 10 películas que más le gusta sobre la base de la co-ocurrencia de cada película en las revisiones de los usuarios.

La salida es un archivo de texto, cada línea representa la recomendación para un usuario. El primer número es el identificador de usuario y los 10 pares de números representan un id de película y una puntuación.

El script utilizado es el siguiente:

```
#!/bin/bash

#Directorio de trabajo
mkdir recommender
cd recommender
```

1 https://www.slideshare.net/vangjee/a-quick-tutorial-on-mahouts-recommendation-engine-v-04
2 https://grouplens.org/datasets/movielens/
3 http://www.apache.org/dyn/closer.cgi/mahout/

```
#Descargar distribution completa de mahout
wget http://apache.rediris.es/mahout/0.11.0/apache-mahout-distribution-0.11.0.tar.gz
tar -xzvf apache-mahout-distribution-0.11.0.tar.gz

#Descargar dataset target
wget http://www.grouplens.org/system/files/ml-100k.zip
unzip ml-100k.zip

#Copiar dataset a mahout
cd ml-100k
hadoop fs -put u.data u.data

#Ejecutar tarea para el recommender de películas
hadoop jar apache-mahout-distribution-0.11.0/mahout-mr-0.11.0-job.jar org.apache.mahout.cf.taste.hadoop.item.RecommenderJob -s
SIMILARITY_COOCCURRENCE --input u.data --output output

#Copiar a local la salida de las recomendaciones
hadoop fs -getmerge output output.txt
cat output.txt
```

Y la salida resumida esta:

```
16    [121:5.0,62:5.0,514:5.0,347:5.0,117:5.0,231:5.0,88:5.0,515:5.0,275:5.0,690:5.0]
32    [42:4.777228,23:4.5793104,156:4.4673915,175:4.397321,240:4.370044,209:4.342984,137:4.3424125,150:4.340467,323:4.3301663,14:4.3236365]
48    [96:5.0,79:5.0,127:5.0,484:5.0,134:5.0,88:5.0,194:5.0,273:5.0,151:5.0,655:5.0]
64    [275:5.0,79:5.0,514:5.0,282:5.0,347:5.0,121:5.0,258:5.0,237:5.0,234:5.0,117:5.0]

…

911   [234:5.0,12:5.0,237:5.0,462:5.0,275:5.0,121:5.0,258:5.0,515:5.0,514:5.0,79:5.0]
927   [255:5.0,231:5.0,234:5.0,275:5.0,12:5.0,117:5.0,258:5.0,237:5.0,282:5.0,195:5.0]
943   [275:5.0,231:5.0,237:5.0,234:5.0,121:5.0,117:5.0,258:5.0,515:5.0,514:5.0,164:5.0]
```