Curso 03: Administração de Dados Complexos em Larga Escala

-- Demonstração Apache Mahout --

Prof. Jose Fernando Rodrigues Junior

Objetivo: demonstrar o uso do Apache Mahout

Instalação Linux do Apache Mahout

⇒ Passo a passo Linux

⇒ Setting JAVA_HOME

- Esta instalação permite a execução *standalone* do Mahout;
- Há várias funcionalidades de Aprendizado de Máquina via linha de comando;
- Se uma infraestrutura distribuída estiver configurada, o processamento distribuído em paralelo ocorrerá automaticamente.

Classificação Naïve Bayes de mensagens textuais de e-mails

- **Dados**: “20 newsgroups” dataset

  O conjunto de dados de notícias “20 newsgroups” compreende cerca de 18.000 postagens de grupos de notícias em 20 tópicos divididos em dois subconjuntos: um para treinamento e outro para teste (ou para avaliação de desempenho).

  ⇒ Sobre o dataset

Classificação Naïve Bayes de mensagens textuais de e-mails

- **Classificador**: Naïve Bayes
  - Baseia-se no teorema do estatístico Thomas Bayes
  - Em resumo: a estatística Bayesiana estipula que a probabilidade de um evento é condicionada à ocorrência de outros eventos relacionados

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

Classificação Naïve Bayes de mensagens textuais de e-mails

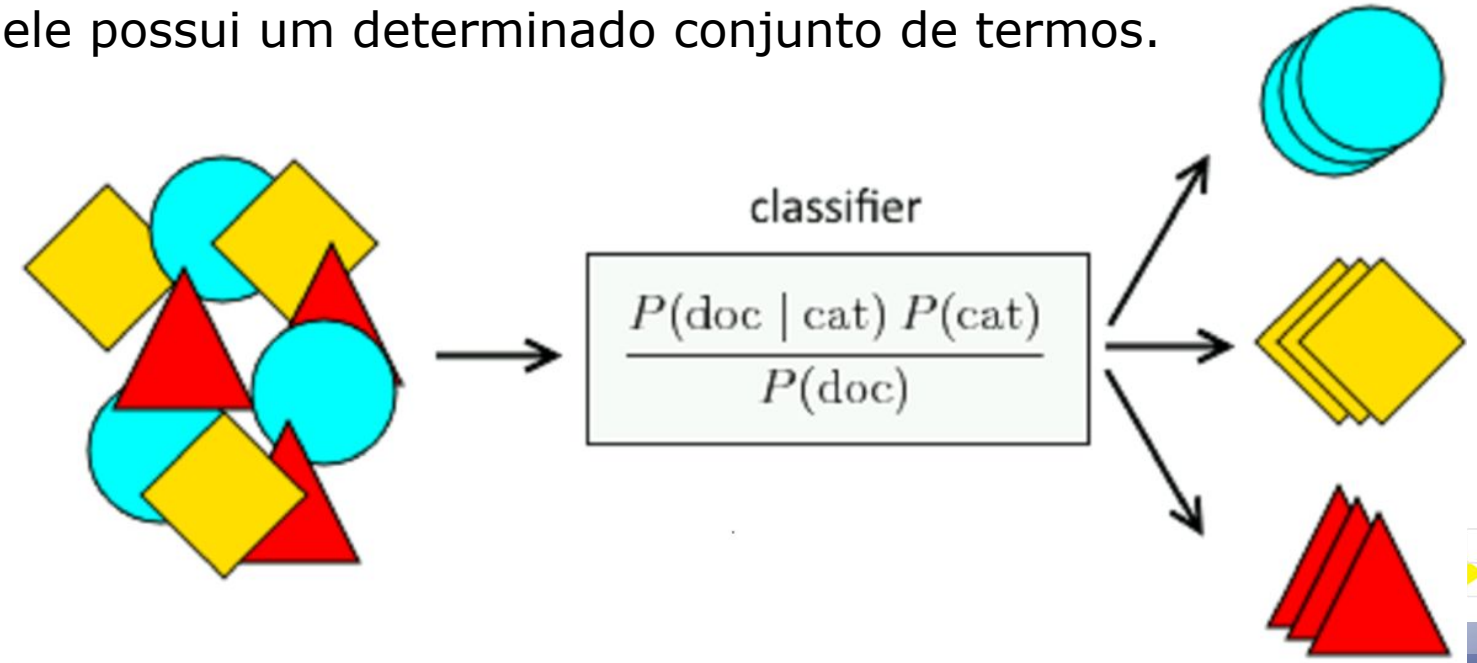
- **Classificador**: Naïve Bayes
  - Por exemplo: a probabilidade de uma pessoa usar cinto de segurança é condicionado à probabilidade desta pessoa ter uma criança no carro

$$P(\text{usar\_cinto}|\text{ter\_criança}) = \frac{P(\text{ter\_criança}|\text{usar\_cinto})P(\text{usar\_cinto})}{P(\text{ter\_criança})}$$

⇒ Naïve Bayes explained

Classificação Naïve Bayes de mensagens textuais de e-mails

- **Classificador**: Naïve Bayes
  - Na classificação de textos, deseja-se computar a probabilidade de um documento ser de uma determinada categoria dado que ele possui um determinado conjunto de termos.



Classificação Naïve Bayes - Passo a passo

1) Fazer o download de dados a partir de:

http://people.csail.mit.edu/jrennie/20Newsgroups/20news-bydate.tar.gz

O conjunto de dados de notícias "20 newsgroups" compreende cerca de 18.000 postagens de grupos de notícias em 20 tópicos divididos em dois subconjuntos: um para treinamento e outro para teste (ou para avaliação de desempenho).

⇒ Sobre o dataset

⇒ Descompactar o arquivo, por exemplo, dentro do diretório

`.../mahout/trunk/bin/`

Nota: como será usada a técnica tf-idf, é preciso fazer o merge dos dados de treino e teste, e um split posterior

Classificação Naïve Bayes - Passo a passo

2) Converter os dados em sequence files

- `mahout seqdirectory -i 20news-bydate -o 20news-seq --overwrite`

Em Hadoop, um *sequence file* é formato de entrada e de saída de dados.

Trata-se de uma estrutura de arquivo simples que consiste em pares de valores-chave serializados em formato binário. É também o formato em que os dados são armazenados internamente durante o processamento das tarefas MapReduce.

Classificação Naïve Bayes - Passo a passo

3) Extrair vetores a partir dos dados de notícias "20 newsgroups"

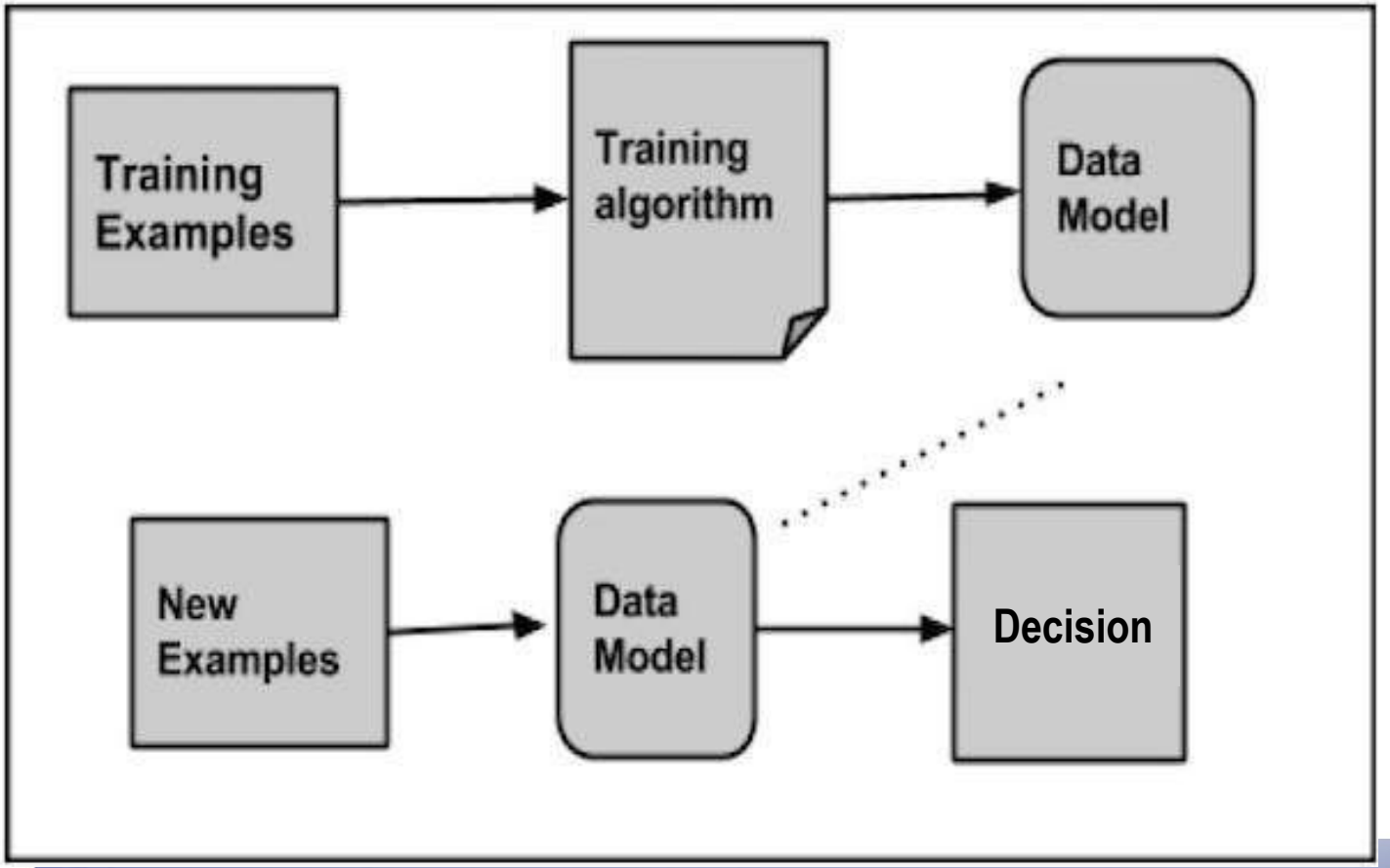
- `mahout seq2sparse -i 20news-seq -o 20news-vectors --logNormalize --namedVector --weight tfidf --overwrite`

Após o tf-idf, fazemos o Split dos dados

- `mahout split -i 20news-vectors/tfidf-vectors --trainingOutput 20news-vectors-train --testOutput 20news-vectors-test --randomSelectionPct 40 --overwrite --sequenceFiles -xm sequential`

Aqui, dados textuais (binarizados) são convertidos em vetores de características textuais. No caso, é usada a técnica TF-IDF: uma medida estatística que indica a importância de uma palavra em relação a uma coleção de documentos ⇒ mais detalhes no Curso 08.

Classificação Naïve Bayes - Passo a passo



Classificação Naïve Bayes - Passo a passo

4) Faz-se o treinamento supervisionado do classificador Naïve Bayes

- `mahout trainnb -i 20news-vectors-train --extractLabels -o ./model/20news-NN-model --labelIndex 20news-labelIndex --overwrite`

5) Testar o modelo sobre os dados de teste

- `mahout testnb -i 20news-vectors-test -m model/20news-NN-model -l 20news-labelIndex -o 20news-TEST-RESULTS --overwrite`

Classificação Naïve Bayes - Passo a passo

6) Avaliação dos resultados

```
Summary
-------------------------------------------------------
Correctly Classified Instances          :       6698       90,1723%
Incorrectly Classified Instances        :        730        9,8277%
Total Classified Instances              :       7428

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       h       i       j       k       l       m       n       o       p       q       r       s       t       <--Classified as
319     0       0       0       0       0       0       0       0       0       0       0       0       0       0       3       0       1       1       9       |  333        a     = alt.atheism
1       298     1       17      6       9       4       0       0       1       0       3       6       1       2       0       0       0       0       0       |  349        b     = comp.graphics
0       26      196     81      23      17      6       0       0       0       0       0       1       0       1       0       1       0       2       1       |  355        c     = comp.os.ms-windows.misc
0       7       2       325     23      4       9       2       0       0       0       0       5       0       1       0       0       0       0       0       |  378        d     = comp.sys.ibm.pc.hardware
1       0       0       14      351     1       3       1       0       0       1       1       4       1       0       0       0       0       1       0       |  379        e     = comp.sys.mac.hardware
0       31      2       5       6       375     1       1       1       0       0       1       0       0       2       0       0       0       0       0       |  425        f     = comp.windows.x
1       0       1       17      5       0       311     7       0       1       2       1       8       0       1       0       1       0       0       0       |  356        g     = misc.forsale
0       1       0       1       1       1       7       373     9       0       0       0       2       1       1       0       1       0       2       0       |  400        h     = rec.autos
0       0       0       0       0       0       5       5       380     0       0       0       0       1       1       0       0       0       0       0       |  392        i     = rec.motorcycles
1       1       0       3       1       0       1       2       1       366     8       0       2       1       0       0       0       0       0       0       |  387        j     = rec.sport.baseball
0       0       0       1       2       0       0       0       1       2       395     0       0       0       0       1       0       0       2       0       |  404        k     = rec.sport.hockey
0       4       0       1       2       2       1       1       0       1       0       374     1       2       0       0       3       0       2       1       |  395        l     = sci.crypt
0       9       0       15      14      3       7       3       0       0       1       2       337     1       1       0       1       0       1       0       |  395        m     = sci.electronics
0       3       0       2       4       2       0       0       1       0       0       0       3       354     3       0       0       1       5       1       |  379        n     = sci.med
2       4       0       1       2       1       2       2       0       1       0       0       0       2       391     0       0       0       0       1       |  409        o     = sci.space
4       1       0       0       1       1       2       0       1       0       0       0       0       5       0       386     1       1       3       0       |  406        p     = soc.religion.christian
0       0       0       0       0       1       0       1       1       0       0       2       0       0       0       0       353     0       13      0       |  371        q     = talk.politics.guns
0       0       0       0       0       0       0       0       0       0       0       0       3       0       0       0       2       370     4       0       |  379        r     = talk.politics.mideast
0       1       0       0       1       0       0       1       0       1       0       1       1       0       1       0       22      1       262     3       |  295        s     = talk.politics.misc
26      0       0       1       0       0       0       1       0       1       0       0       0       1       2       14      5       4       4       182     |  241        t     = talk.religion.misc

=======================================================
Statistics
-------------------------------------------------------
Kappa                                     0,8724
Accuracy                                 90,1723%
Reliability                              85,4427%
Reliability (standard deviation)          0,2181
```

Conclusões

- O Apache Mahout tem uma **ampla quantidade de algoritmos** de Aprendizado de Máquina

- Problemas como **classificação e recomendação** podem muito bem serem resolvidos via Mahout

- Apesar do mapreduce ser uma tecnologia defasada com relação ao Spark, ainda é a melhor **solução para problemas na escala de Big Data**, pois é mais escalável do que o Spark