Hadoop – Mahout
2-Classification

عادل قاضی خانی
هیئت علمی گروه کامپیوتر دانشگاه امام رضا (ع)

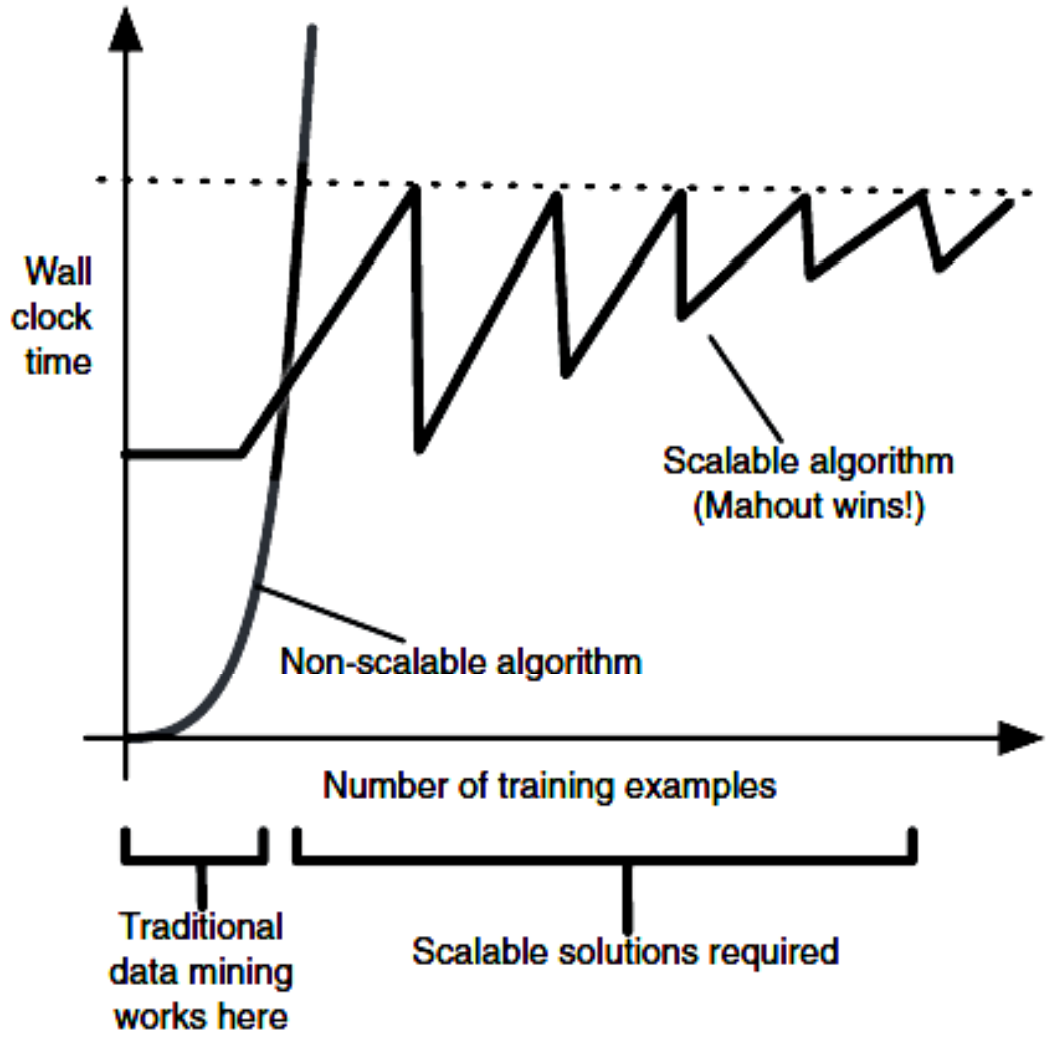
تعریف

- یکی از پروژه های Apache که از سال ۲۰۰۸ آغاز شد
- مجموعه ای از **تکنیکها و الگوریتمها** حول سه محور **سیستم های توصیه گر، خوشه بندی و طبقه بندی**
- قابلیت اعمال **الگوریتمهای یادگیری** روی **داده های حجیم**

طبقه بندی در Mahout

System size in number of examples	Choice of classification approach
< 100,000	Traditional, non-Mahout approaches should work very well. Mahout may even be slower for training.
100,000 to 1 million	Mahout begins to be a good choice. The flexible API may make Mahout a preferred choice, even though there is no performance advantage.
1 million to 10 million	Mahout is an excellent choice in this range.
> 10 million	Mahout excels where others fail.

طبقه بندی در Mahout

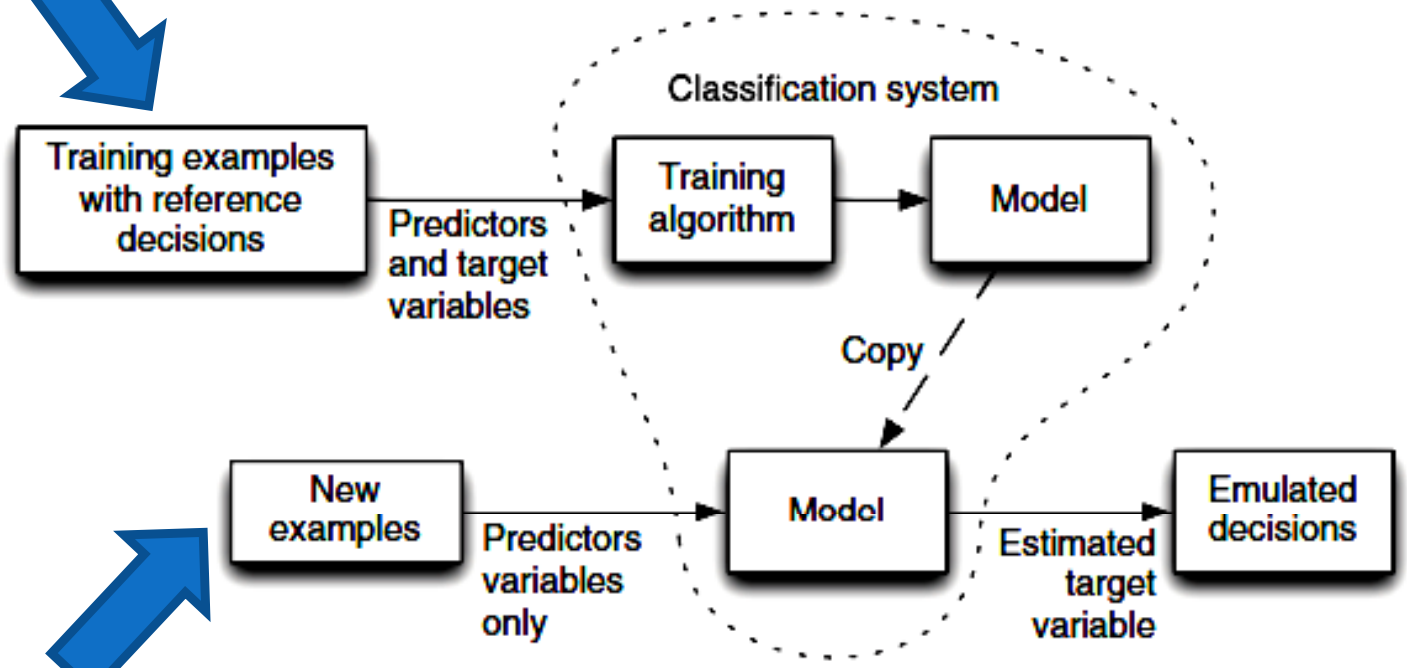


طبقه بندی در Mahout

- **فرایندی** که با استفاده **از یک یا چند ورودی یک یا چند خروجی** از میان **چند خروجی** موجود انتخاب می شود
- **یادگیری** که در آن داده ها **ورودی و خروجی** دارند
- **هدف** تخمین **وابستگی میان ورودی و خروجی**
- **مثلاً:** تشخیص چهره، شناسایی اسپم و ...

نحوه کار طبقه بند

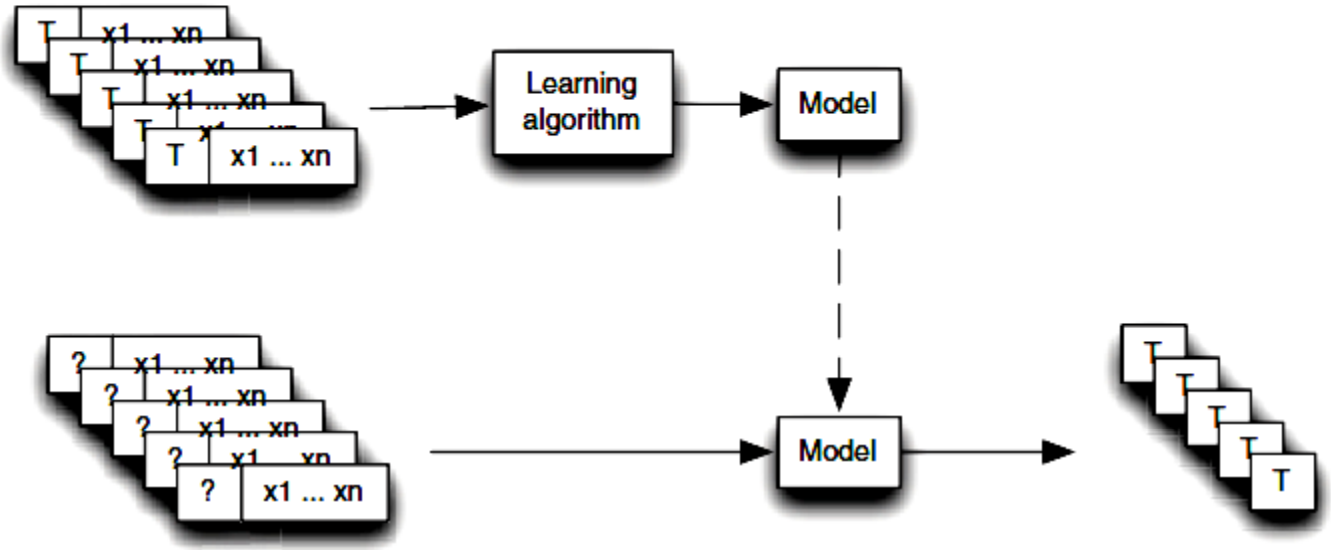
آموزش (یادگیری)



آزمایش (پیش بینی)

نحوه کار طبقه بند

آموزش (یادگیری)



آزمایش (پیش بینی)

انواع ویژگی

Type of value	Description
Continuous	This is a floating-point value. This type of value might be a price, a weight, a time, (anything else that has a numerical magnitude and where this magnitude is the key property of the value.
Categorical	A categorical value can have one of a set of prespecified values. Typically the set o categorical values is relatively small and may be as small as two, although the set can be quite large. Boolean values are generally treated as categorical values. Another example might be a vendor ID.
Word-like	A word-like value is like a categorical value, but it has an open-ended set of possible values.
Text-like	A text-like value is a sequence of word-like values, all of the same kind. Text is the classic example of a text-like value, but a list of email addresses or URLs is also text-like.

انواع ویژگی – مثال

Name	Type	Value
from-address	Word-like	George <george@fumble-tech.com>
in-address-book?	Categorical (TRUE, FALSE)	TRUE
non-spam-words	Text-like	Ted, Mahout, User, lunch
spam-words	Text-like	available
unknown-words	Continuous	0
message-length	Continuous	31

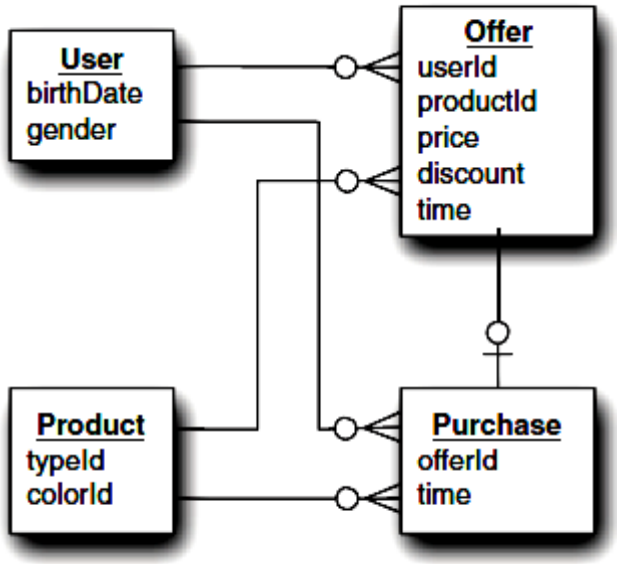
آماده سازی داده ها

- **در ابتدا** باید مشخص کنیم **ویژگی ها و خروجی** چیست؟
- **آماده سازی** برای یادگیری توسط یک طبقه بند

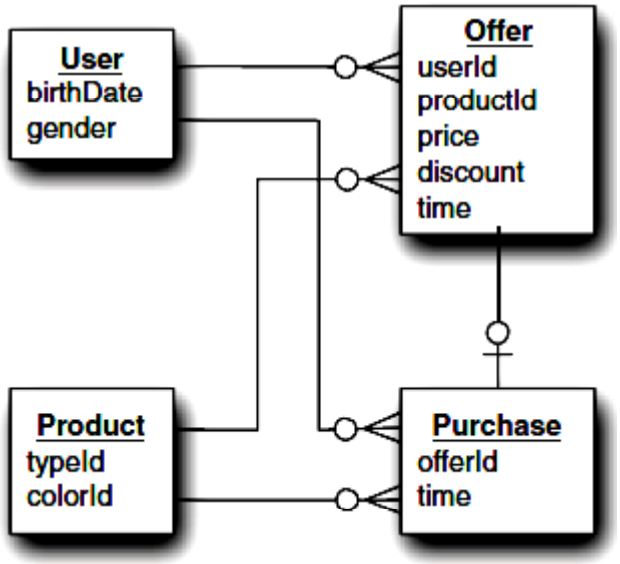
آماده سازی داده ها

- **مثال: طبقه بندی که مشخص کند یک مشتری پس از معرفی محصول آنرا خرید می کند یا خیر**

- **داده ها در پایگاه داده**



آماده سازی داده ها – مثال



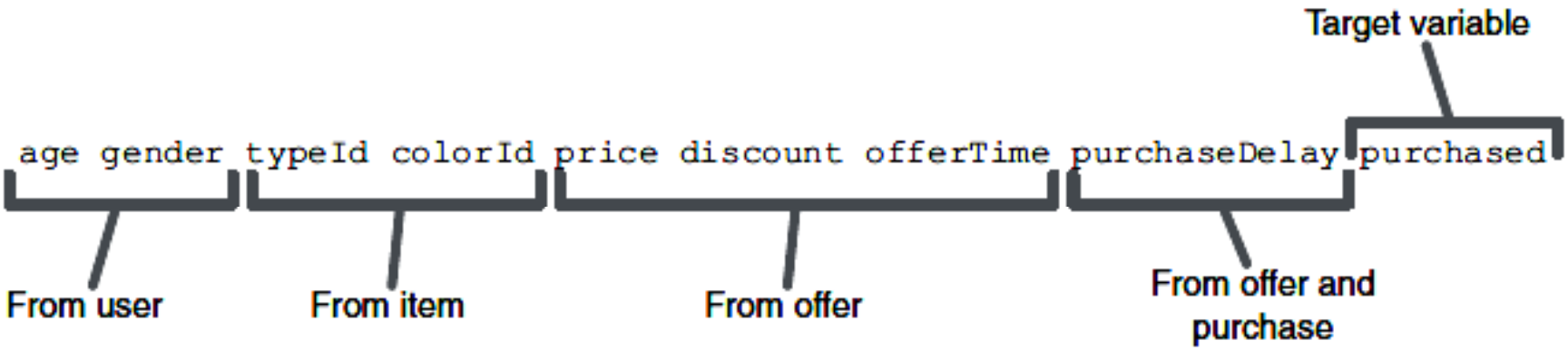
- داده ها در پایگاه داده

- دستور SQL برای خواندن داده های مورد نیاز

```
select
  now()-birthDate as age, gender,
  typeId, colorId, price, discount, offerTime,
  ifnull(purchase.time, 0, purchase.time - offer.time) as purchaseDelay,
  ifnull(purchase.time, 0, 1) as purchased
from
  offer
  join user using (userId)
  join product using (productId)
  left outer join purchase using (offerId);
```

آماده سازی داده ها – مثال

- داده ها در یک رکورد



طبقه بندی در Mahout – مثال

- **طبقه بندی داده های 20 newsgroups با الگوریتم یادگیری SGD**

- این داده ها مربوط به اطلاعات ارسالی به ۲۰ گروه خبری در Usennet در طی چند ماه است

طبقه بندی در Mahout – مثال

- طبقه بندی داده های 20 newsgroups با الگوریتم یادگیری SGD – بررسی داده ها (هر فایل)

```
From: rdippold@qualcomm.com (Ron "Asbestos" Dippold)
Subject: Re: text of White House announcement and Q&As
Originator: rdippold@qualcom.qualcomm.com
Nntp-Posting-Host: qualcom.qualcomm.com
Organization: Qualcomm, Inc., San Diego, CA
Lines: 12

ted@nmsu.edu (Ted Dunning) writes:
>nobody seems to have noticed that the clipper chip *must* have been
>under development for considerably longer than the 3 months that
>clinton has been president. this is not something that choosing
```

Header

Message

طبقه بندی در Mahout – مثال

- تعداد تکرار بخشهای Header در فایلهای مختلف داده ها

Header	Count	Comment	Action
Subject	11,314	Text	Keep
From	11,314	Sender of message	Keep
Lines	11,311	Number of lines in message	Keep
Organization	10,841	Related to sender	Try?
Distribution	2,533	Possible target leak	Try?
Nntp-Posting-Host	2,453	Related to sender	Try?
NNTP-Posting-Host	2,311	Same as previous except for capitalization	Try?
Reply-To	1,720	Possible cue, not common, experiment with this	Try?
Keywords	926	Describes content	Keep
Article-I.D.	673	Too specific	Drop
X-Newsreader	588	Software used by sender	Drop
Summary	391	Like “Keywords”	Keep
Originator	291	Similar to “From,” but much less common	Drop
In-Reply-To	219	Probably just an ID number, rare	Drop
News-Software	164	Software used by sender	Drop
Expires	113	Date-specific, rare	Drop
In-reply-to	101	Probably just an ID number, rare	Drop
To	80	Possible target leak, rare	Drop
X-Disclaimer	64	Noise, rare	Drop
Disclaimer	56	Possible cue due to newsgroup paranoia level, rare	Drop
...		...	
Weather	1	Bizarre header	Drop
Orginization	1	Spelling errors in headers are surprising	Drop
Oganization	1	Well, maybe not	Drop
Oanization	1		Drop
Moon-Phase	1	Yes, really, one posting had this; open source is wonderful	!?!

طبقه بندی در Mahout – مثال

- با توجه به بررسی داده ها، ویژگی های ***Subject, From*** ***Lines,Keywords,Summary*** در Header مناسب هستند
- در ابتدا باید متن tokenize (تکه تکه) شود

طبقه بندی در Mahout – تکه تکه شدن

```
for (File file : files) {
  BufferedReader reader = new BufferedReader(new FileReader(file));
  String ng = file.getParentFile().getName();
  int actual = newsGroups.intern(ng);
  Multiset<String> words = ConcurrentHashMultiset.create();

  String line = reader.readLine();
  while (line != null && line.length() > 0) {
    if (line.startsWith("Lines:")) {
      String count = Iterables.get(onColon.split(line), 1);
      try {
        lineCount = Integer.parseInt(count);
        averageLineCount += (lineCount - averageLineCount)
            / Math.min(k + 1, 1000);
      } catch (NumberFormatException e) {
        lineCount = averageLineCount;
      }
    }
    boolean countHeader = (
        line.startsWith("From:") || line.startsWith("Subject:")||
        line.startsWith("Keywords:")|| line.startsWith("Summary:"));
    do {
      StringReader in = new StringReader(line);
      if (countHeader) {
        countWords(analyzer, words, in);
      }
      line = reader.readLine();
    } while (line.startsWith(" "));
  }
  countWords(analyzer, words, reader);
  reader.close();
}
```

1. Identify newsgroup
2. Check for line count header
3. Count words in headers
4. Count words in body

طبقه بندی در Mahout – مثال

- سپس باید ویژگی های ***Subject, From Lines,Keywords,Summary*** از Header استخراج شوند و به بردار تبدیل شوند

طبقه بندی در Mahout – تبدیل به بردار

```
Vector v = new RandomAccessSparseVector(FEATURES);
bias.addToVector(null, 1, v);
lines.addToVector(null, lineCount / 30, v);
logLines.addToVector(null, Math.log(lineCount + 1), v);
for (String word : words.elementSet()) {
  encoder.addToVector(word, Math.log(1 + words.count(word)), v);
}
```

طبقه بندی در Mahout – ایجاد و تنظیم مدل

```
OnlineLogisticRegression learningAlgorithm =
    new OnlineLogisticRegression(
          20, FEATURES, new L1())
        .alpha(1).stepOffset(1000)
        .decayExponent(0.9)
        .lambda(3.0e-5)
        .learningRate(20);
```

طبقه بندی در Mahout – آموزش مدل

```
learningAlgorithm.train(actual, v);
k++;
int bump = bumps[(int) Math.floor(step) % bumps.length];
int scale = (int) Math.pow(10, Math.floor(step / bumps.length));
if (k % (bump * scale) == 0) {
  step += 0.25;
  System.out.printf("%10d %10.3f %10.3f %10.2f %s %s\n",
       k, ll, averageLL, averageCorrect * 100, ng,
       newsGroups.values().get(estimated));
}
learningAlgorithm.close();
```

طبقه بندی در Mahout

Size of data set	Mahout algorithm	Execution model	Characteristics
Small to medium (less than tens of millions of training examples)	Stochastic gradient descent (SGD) family: `OnlineLogisticRegression`, `CrossFoldLearner`, `AdaptiveLogisticRegression`	Sequential, online, incremental	Uses all types of predictor variables; sleek and efficient over the appropriate data range (up to millions of training examples)
	Support vector machine (SVM)	Sequential	Experimental still; sleek and efficient over the appropriate data range
Medium to large (millions to hundreds of millions of training examples)	Naive Bayes	Parallel	Strongly prefers text-like data; medium to high overhead for training; effective and useful for data sets too large for SGD or SVM
	Complementary naive Bayes	Parallel	Somewhat more expensive to train than naive Bayes; effective and useful for data sets too large for SGD, but has similar limitations to naive Bayes
Small to medium (less than tens of millions of training examples)	Random forests	Parallel	Uses all types of predictor variables; high overhead for training; not widely used (yet); costly but offers complex and interesting classifications, handles nonlinear and conditional relationships in data better than other techniques