Data mining with WEKA, Part 1: Introduction and regression

Skill Level: Intermediate

Michael Abernethy (mabernethy@yahoo.com)
Product Development Manager
Optimal Auctions

27 Apr 2010

Data mining is the talk of the tech industry, as companies are generating millions of data points about their users and looking for a way to turn that information into increased revenue. Data mining is a collective term for dozens of techniques to glean information from data and turn it into something meaningful. This article will introduce you to open source data-mining software and some of the most common techniques to interpret data.

Introduction

What is data mining? You may find yourself asking this question occasionally, as the topic seems to be getting increased attention in the world of technology. You read that companies like Google and Yahoo! are generating billions of data points about all their users, and you may wonder, "What do they plan to do with all that information?" You may also be surprised to know that Walmart is one of the most advanced companies that mines data and applies the results to their business. Virtually every company in the world is using data mining now, and those that don't will soon find themselves at an extreme disadvantage.

So, how do you get you and your company on board the data-mining bandwagon?

We hope to answer all of your initial questions about data mining. We also will introduce you to Waikato Environment for Knowledge Analysis (WEKA), free and open source software you can use to mine your own data and turn what you know about your users, your clients, and your business into useful information for increasing your revenue. You will see that it is not as difficult as you might think it is

to do a "pretty good" job of mining data.

Additionally, this article will discuss the first technique for data mining: regression, which transforms existing data into a numerical prediction for future data. It is likely the easiest method of mining data and even on a simple level something you may have done before in your favorite market-dominant spreadsheet software (though WEKA can do much more complex calculations). Future articles will touch upon other methods of mining data, including clustering, Nearest Neighbor, and classification trees. (If those terms mean nothing to you, don't worry. We'll cover it all in this series.)

What is data mining?

Data mining, at its core, is the transformation of large amounts of data into meaningful patterns and rules. Further, it could be broken down into two types: directed and undirected. In directed data mining, you are trying to predict a particular data point — the sales price of a house given information about other houses for sale in the neighborhood, for example.

In undirected data mining, you are trying to create groups of data, or find patterns in existing data — creating the "Soccer Mom" demographic group, for example. In effect, every U.S. census is data mining, as the government looks to gather data about everyone in the country and turn it into useful information.

For our purposes, modern data mining started in the mid-1990s, as the power of computing, and the cost of computing and storage finally reached a level where it was possible for companies to do it in-house, without having to look to outside computer powerhouses.

Additionally, the term data mining is all-encompassing, referring to dozens of techniques and procedures used to examine and transform data. Therefore, this series of articles will only scratch the surface of what is possible with data mining. Experts likely will have doctorates in statistics and have spent 10-30 years in the field. That may leave you with the impression that data mining is something only big companies can afford.

We hope to clear up many of these misconceptions about data mining, and we hope to make it clear that it is not as easy as simply running a function in a spreadsheet against a grid of data, yet it is not so difficult that everyone can't manage some of it themselves. This is the perfect example of the 80/20 paradigm — maybe even pushed further to the 90/10 paradigm. You can create a data-mining model with 90-percent effectiveness with only 10 percent of the expertise of one of these so-called data-mining experts. To bridge the remaining 10 percent of the model and create a perfect model would require 90-percent additional time and perhaps another 20 years. So unless you plan to make a career out of data mining, the "good

enough" is likely all that you need. Looking at it another way, good enough is probably better than what you're doing right now anyway.

The ultimate goal of data mining is to create a model, a model that can improve the way you read and interpret your existing data and your future data. Since there are so many techniques with data mining, the major step to creating a good model is to determine what type of technique to use. That will come with practice and experience, and some guidance. From there, the model needs to be refined to make it even more useful. After reading these articles, you should be able to look at your data set, determine the right technique to use, then take steps to refine it. You'll be able to create a good-enough model for your own data.

WEKA

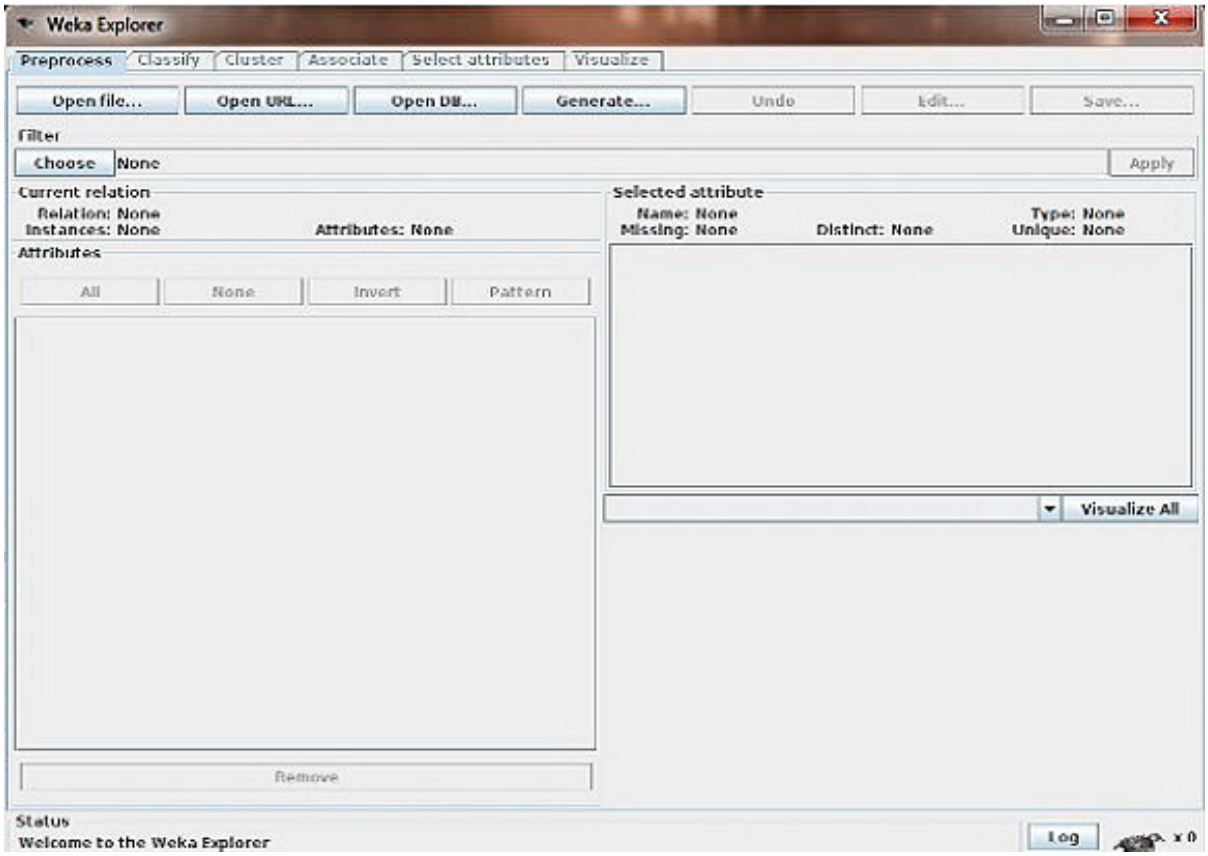
Data mining isn't solely the domain of big companies and expensive software. In fact, there's a piece of software that does almost all the same things as these expensive pieces of software — the software is called WEKA (see Resources). WEKA is the product of the University of Waikato (New Zealand) and was first implemented in its modern form in 1997. It uses the GNU General Public License (GPL). The software is written in the Java™ language and contains a GUI for interacting with data files and producing visual results (think tables and curves). It also has a general API, so you can embed WEKA, like any other library, in your own applications to such things as automated server-side data-mining tasks.

At this point, please go ahead and install WEKA. It's Java-based, so if you don't have a JRE installed on your computer, download the WEKA version that contains the JRE, as well.

Figure 1. WEKA startup screen

When you start WEKA, the GUI chooser pops up and lets you choose four ways to work with WEKA and your data. For all the examples in this article series, we will choose only the **Explorer** option. This option is more than sufficient for everything we need to do in these articles.

Figure 2. WEKA Explorer



Now that you're familiar with how to install and start up WEKA, let's get into our first data-mining technique: regression.

Regression

Regression is the easiest technique to use, but is also probably the least powerful (funny how that always goes hand in hand). This model can be as easy as one input variable and one output variable (called a Scatter diagram in Excel, or an XYDiagram in OpenOffice.org). Of course, it can get more complex than that, including dozens of input variables. In effect, regression models all fit the same general pattern. There are a number of independent variables, which, when taken together, produce a result — a dependent variable. The regression model is then used to predict the result of an unknown dependent variable, given the values of the independent variables.

Everyone has probably used or seen a regression model before, maybe even mentally creating a regression model. The example that immediately comes to mind is pricing a house. The price of the house (the dependent variable) is the result of many independent variables — the square footage of the house, the size of the lot, whether granite is in the kitchen, bathrooms are upgraded, etc. So, if you've ever bought a house or sold a house, you've likely created a regression model to price

the house. You created the model based on other comparable houses in the neighborhood and what they sold for (the model), then put the values of your own house into this model to produce an expected price.

Let's continue this example of a house price-based regression model, and create some real data to examine. These are actual numbers from houses for sale in my neighborhood, and I will be trying to find the value for my own house. (I'll also be taking the output from this model to protest my property-tax assessment).

Table 1. House values for regression model

House size (square feet)	Lot size	Bedrooms	Granite	Upgraded bathroom?	Selling price
3529	9191	6	0	0	$205,000
3247	10061	5	1	1	$224,900
4032	10150	5	0	1	$197,900
2397	14156	4	1	0	$189,900
2200	9600	4	0	1`	$195,000
3536	19994	6	1	1	$325,000
2983	9365	5	0	1	$230,000
3198	9669	5	1	1	????

The good news (or bad news, depending on your point of view) is that this little introduction to regression barely scratches the surface, and that scratch is really even barely noticeable. There are entire college semester courses on regression models, that will teach you more about regression models than you probably even want to know. But this scratch gets you acquainted with the concept and suffice for our WEKA tests in this article. If you have continued interest in regression models and all the statistical details that go into them, research the following terms with your favorite search engine: least squares, homoscedasticity, normal distribution, White tests, Lilliefors tests, R-squared, and p-values.

This data has been created so just read this part. don't do anything

Building the data det for WEKA

To load data into WEKA, we have to put it into a format that will be understood. WEKA's preferred method for loading data is in the Attribute-Relation File Format (ARFF), where you can define the type of data being loaded, then supply the data itself. In the file, you define each column and what each column contains. In the case of the regression model, you are limited to a `NUMERIC` or a `DATE` column. Finally, you supply each row of data in a comma-delimited format. The ARFF file we'll be using with WEKA appears below. Notice in the rows of data that we've left out my house. Since we are creating the model, we cannot input my house into it

since the selling price is unknown.

Listing 1. WEKA file format

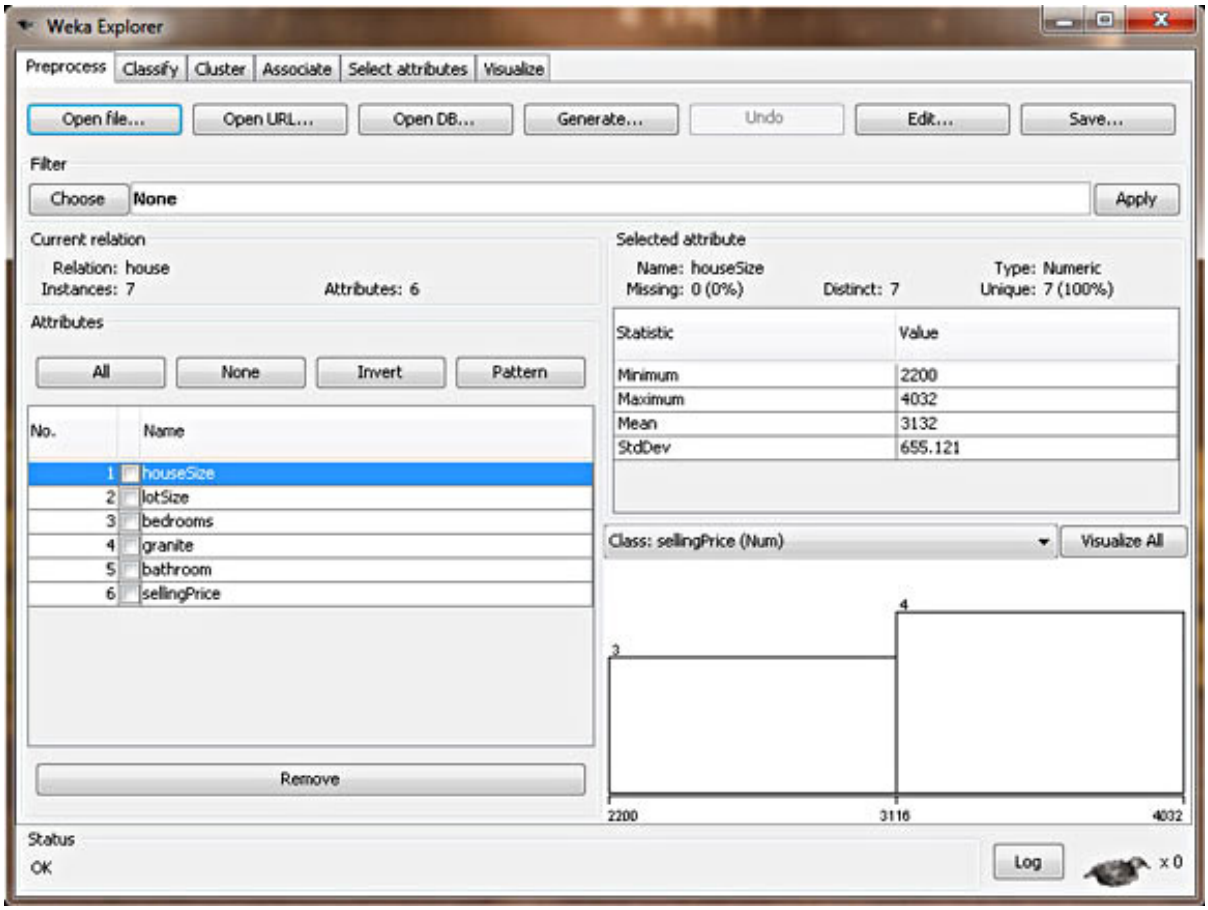
```
@RELATION house

@ATTRIBUTE houseSize NUMERIC
@ATTRIBUTE lotSize NUMERIC
@ATTRIBUTE bedrooms NUMERIC
@ATTRIBUTE granite NUMERIC
@ATTRIBUTE bathroom NUMERIC
@ATTRIBUTE sellingPrice NUMERIC

@DATA
3529,9191,6,0,0,205000
3247,10061,5,1,1,224900
4032,10150,5,0,1,197900
2397,14156,4,1,0,189900
2200,9600,4,0,1,195000
3536,19994,6,1,1,325000
2983,9365,5,0,1,230000
```

Loading the data into WEKA

Load houses.arff

Now that the data file has been created, it's time to create our regression model. Start WEKA, then choose the **Explorer.** You'll be taken to the Explorer screen, with the **Preprocess** tab selected. Select the **Open File** button and select the ARFF file you created in the section above. After selecting the file, your WEKA Explorer should look similar to the screenshot in Figure 3.

Figure 3. WEKA with house data loaded SCREENSHOT HERE



In this view, WEKA allows you to review the data you're working with. In the left section of the Explorer window, it outlines all of the columns in your data (Attributes) and the number of rows of data supplied (Instances). By selecting each column, the right section of the Explorer window will also give you information about the data in that column of your data set. For example, by selecting the **houseSize** column in the left section (which should be selected by default), the right-section should change to show you additional statistical information about the column. It shows the maximum value in the data set for this column is 4,032 square feet, and the minimum is 2,200 square feet. The average size is 3,131 square feet, with a standard deviation of 655 square feet. (Standard deviation is a statistical measure of variance.) Finally, there's a visual way of examining the data, which you can see by clicking the **Visualize All** button. Due to our limited number of rows in this data set, the visualization is not as powerful as it would be if there were more data points (in the hundreds, for example).

Enough looking at the data. Let's create a model and get a price for my house.

Creating the regression model with WEKA

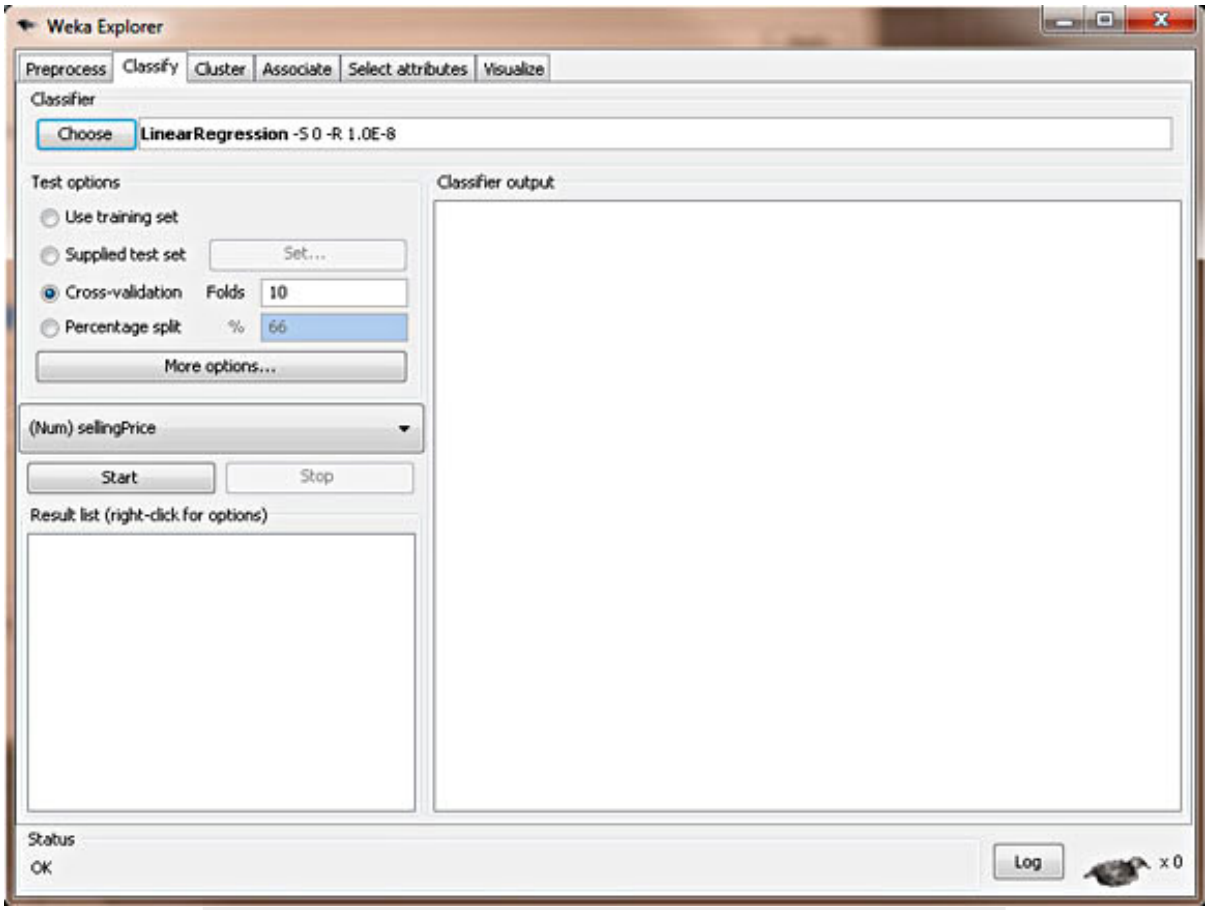
To create the model, click on the **Classify** tab. The first step is to select the model we want to build, so WEKA knows how to work with the data, and how to create the appropriate model:

1. Click the **Choose** button, then expand the **functions** branch.
2. Select the **LinearRegression** leaf.

This tells WEKA that we want to build a regression model. As you can see from the other choices, though, there are lots of possible models to build. Lots! This should give you a good indication of how we are only touching the surface of this subject. Also of note: There is another choice called **SimpleLinearRegression** in the same branch. Do not choose this because simple regression only looks at one variable, and we have six. When you've selected the right model, your WEKA Explorer should look like Figure 4.

SCREENSHOT HERE

Figure 4. Linear regression model in WEKA



Can I do this with a spreadsheet?

Short answer: No. Long answer: Yes. Most popular spreadsheet programs cannot easily do what we did with WEKA, which was

defining a linear regression model with multiple independent variables. However, you can do a Simple Linear Regression model (one independent variable) pretty easily. If you're feeling brave, it can do multi-variable regression, though it's quite confusing and difficult, definitely not as easy as WEKA. You can see an example video for Microsoft® Excel® in Resources.

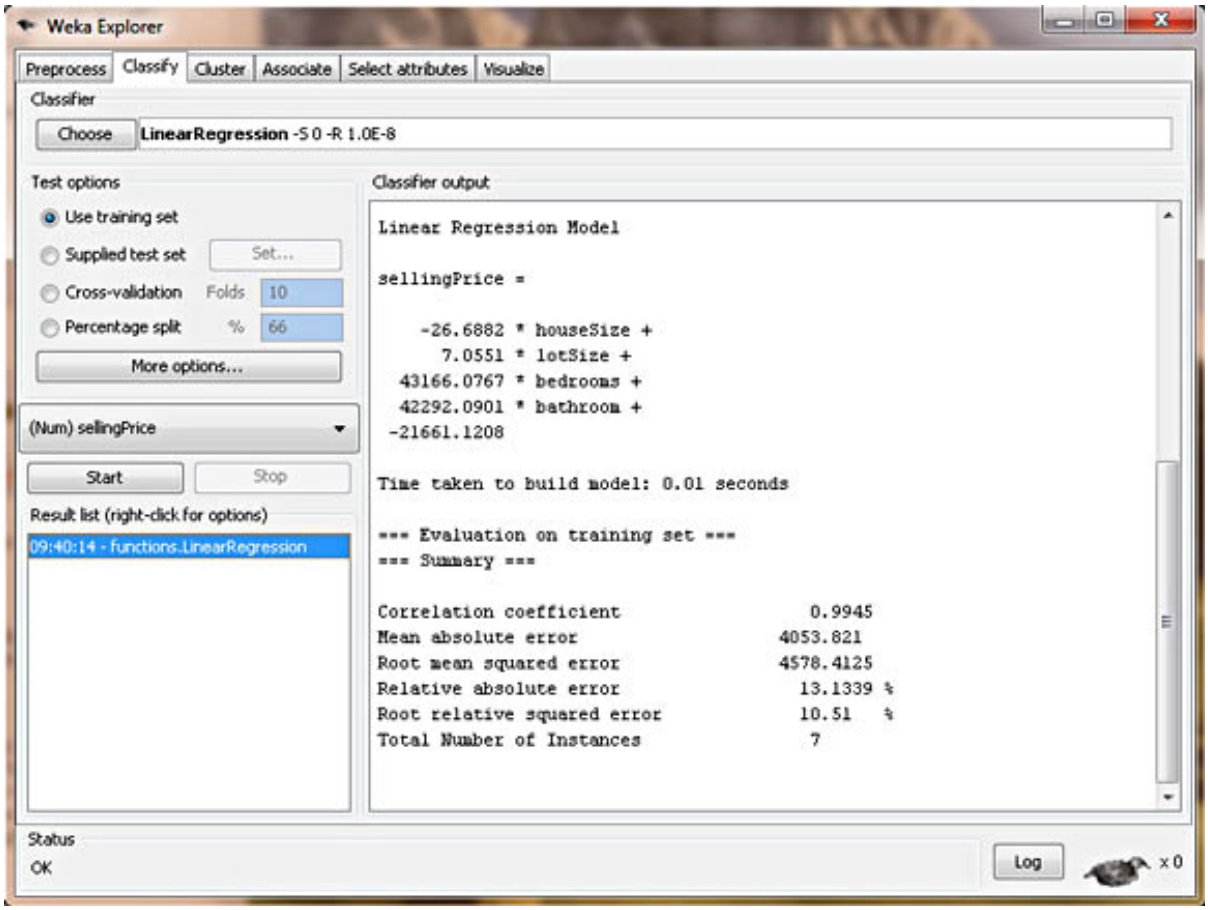
Now that the desired model has been chosen, we have to tell WEKA where the data is that it should use to build the model. Though it may be obvious to us that we want to use the data we supplied in the ARFF file, there are actually different options, some more advanced than what we'll be using. The other three choices are **Supplied test set,** where you can supply a different set of data to build the model; **Cross-validation,** which lets WEKA build a model based on subsets of the supplied data and then average them out to create a final model; and **Percentage split,** where WEKA takes a percentile subset of the supplied data to build a final model. These other choices are useful with different models, which we'll see in future articles. With regression, we can simply choose **Use training set.** This tells WEKA that to build our desired model, we can simply use the data set we supplied in our ARFF file.

Finally, the last step to creating our model is to choose the dependent variable (the column we are looking to predict). We know this should be the selling price, since that's what we're trying to determine for my house. Right below the test options, there's a combo box that lets you choose the dependent variable. The column **sellingPrice** should be selected by default. If it's not, please select it.

Now we are ready to create our model. Click **Start**. Figure 5 shows what the output should look like.

SCREENSHOT HERE !!!!

Figure 5. House price regression model in WEKA



Interpreting the regression model

WEKA doesn't mess around. It puts the regression model right there in the output, as shown in Listing 2.

Listing 2. Regression output

```
sellingPrice = (-26.6882   * houseSize) +
               (7.0551     * lotSize) +
               (43166.0767 * bedrooms) +
               (42292.0901 * bathroom)
             - 21661.1208
```

Listing 3 shows the results, plugging in the values for my house.

Listing 3. House value using regression model

```
sellingPrice = (-26.6882   * 3198) +
               (7.0551     * 9669) +
               (43166.0767 * 5) +
```

```
              (42292.0901 * 1)
            - 21661.1208

sellingPrice = 219,328
```

However, looking back to the top of the article, data mining isn't just about outputting a single number: It's about identifying patterns and rules. It's not strictly used to produce an absolute number but rather to create a model that lets you detect patterns, predict output, and come up with conclusions backed by the data. Let's take another step and interpret the patterns and conclusions that our model tells us, besides just a strict house value:

- **Granite doesn't matter** — WEKA will only use columns that statistically contribute to the accuracy of the model (measured in R-squared, but beyond the scope of this article). It will throw out and ignore columns that don't help in creating a good model. So this regression model is telling us that granite in your kitchen doesn't affect the house's value.
- **Bathrooms do matter** — Since we use a simple 0 or 1 value for an upgraded bathroom, we can use the coefficient from the regression model to determine the value of an upgraded bathroom on the house value. The model tells us it adds $42,292 to the house value.
- **Bigger houses reduce the value** — WEKA is telling us that the bigger our house is, the lower the selling price? This can be seen by the negative coefficient in front of the `houseSize` variable. The model is telling us that every additional square foot of the house reduces its price by $26? That doesn't make any sense at all. This is America! Bigger is better, especially where I live in Texas. How should we interpret this? This is a good example of garbage in, garbage out. The house size, unfortunately, isn't an independent variable because it's related to the bedrooms variable, which makes sense, since bigger houses tend to have more bedrooms. So our model isn't perfect. But we can fix this. Remember: On the **Preprocess** tab, you can remove columns from the data set. For your own practice, remove the **houseSize** column and create another model. How does it affect the price of my house? How does this new model make more sense? (My amended house value: $217,894).

> **Note to statisticians**
> This model breaks several requirements of a "proper" linear regression model, since every column isn't truly independent, and there aren't enough rows of data to produce a valid model. Since the primary purpose of this article is to introduce WEKA as a data-mining tool, we are oversimplifying the example data.

To take this simple example to the next level, let's take a look at a data file that the

WEKA Web site supplies to us as a regression example. Theoretically, this should be much more complex than our simple example with seven houses. This sample data file attempts to create a regression model to predict the miles per gallon (MPG) for a car based on several attributes of the car (this data is from 1970 to 1982, so keep that in mind). The model includes these possible attributes of the car: cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car make. Further, this data set has 398 rows of data and meets many of the statistical requirements that our earlier house price model didn't. In theory, this should be a much more complex regression model, and perhaps WEKA might have a hard time creating a model with this much data (though I'm sure you can predict at this point that WEKA will handle this just fine).

LOAD autoMPG.ARFF here

To produce the regression model with this data set, you should follow the exact same steps you followed for the house data, so I won't repeat it. So go ahead and create the regression model. It should produce the output shown in Listing 4.

Listing 4. MPG data regression model SCREENSHOT HERE !!!!

```
class (aka MPG) =

      -2.2744 * cylinders=6,3,5,4 +
      -4.4421 * cylinders=3,5,4 +
       6.74   * cylinders=5,4 +
       0.012  * displacement +
      -0.0359 * horsepower +
      -0.0056 * weight +
       1.6184 * model=75,71,76,74,77,78,79,81,82,80 +
       1.8307 * model=77,78,79,81,82,80 +
       1.8958 * model=79,81,82,80 +
       1.7754 * model=81,82,80 +
       1.167  * model=82,80 +
       1.2522 * model=80 +
       2.1363 * origin=2,3 +
       37.9165
```

When you do it yourself, you'll see that WEKA flies through the model in less than a second. So it's not a problem, computationally, to create a powerful regression model from a lot of data. This model may also appear much more complex than the house data, but it isn't. For example, the first line of the regression model, `-2.2744 * cylinders=6,3,5,4` means that if the car has six cylinders, you would place a 1 in this column, and if it has eight cylinders, you would place a 0. Let's take one example row from the data set (row 10) and plug those numbers into the regression model, and see if the output from the model approximates the output that was given to us in the data set.

END NO NEED TO EXECUTE FURTHER

Listing 5. Example MPG data

```
data = 8,390,190,3850,8.5,70,1,15

class (aka MPG) =
```

```
      -2.2744 * 0 +
      -4.4421 * 0 +
       6.74   * 0 +
       0.012  * 390 +
      -0.0359 * 190 +
      -0.0056 * 3850 +
       1.6184 * 0 +
       1.8307 * 0 +
       1.8958 * 0 +
       1.7754 * 0 +
       1.167  * 0 +
       1.2522 * 0 +
       2.1363 * 0 +
      37.9165

Expected Value = 15 mpg
Regression Model Output = 14.2 mpg
```

So our model did pretty well when we evaluate it with our randomly chosen test data, predicting a 14.2 MPG on a car whose actual value was 15 MPG.

Conclusion

This article strives to answer the question "what is data mining?" by giving you a background on the subject and introducing the goals of the field. Data mining strives to turn a lot of misinformation (in the form of scattered data) into useful information by creating models and rules. Your goal is to use the models and rules to predict future behavior, to improve your business, or to just explain things you might not otherwise be able to. These models may confirm what you've already thought, or even better, may find new things in your data you never knew existed. As a funny example, there is an urban data-mining legend (not sure how many of these exist) that, in the United States, Walmart moves beer to the end of the diaper aisles on weekends because its data mining showed that men typically buy diapers on weekends, and many men also like beer on weekends.

This article also introduced you to the free and open source software program WEKA. There are certainly complex commercial software products built for data mining, but, for the average person looking to start in data mining, there's a useful solution available that's open source. Remember, you are never going to be an expert in data mining unless you want to spend 20 years doing it. WEKA will let you get started and provide a pretty good solution to many of your initial problems. If you weren't doing any mining before, the pretty-good solution is all you need.

Finally, this article discussed the first data-mining model, the regression model (specifically, the linear regression multi-variable model), and showed how to use it in WEKA. This regression model is easy to use and can be used for myriad data sets. You may find it the most useful model I discuss in this series. However, data mining is much more than simply regression, and you'll find some other models are better solutions with different data sets and different output goals.

Finally, I want to reiterate that this article and the ones in the future parts of this series only are a brief introduction to the field of statistics and data mining. People spend an entire semester on statistics and an entire semester on data mining, and only then are they considered "beginners." Our goal is to explore the open source tooling available for a beginner and to foster an appreciation for the value that data mining might provide. Keep that in mind as we continue the series.

Downloads

Description	Name	Size	Download method
Sample code	os-weka1-Examples.zip	6KB	HTTP

Information about download methods

Resources

Learn

- WEKA requests that all publications about it cite the paper titled "The WEKA Data Mining Software: An Update," by Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, and Ian H. Witten.
- See how to use a spreadsheet for a simple regression model with this video on YouTube.
- Check out the WEKA Web site for all the documentation and an FAQ about the software.
- Read more about Regression analysis on Wikipedia, which probably has more technical details than you'd ever want to know.
- Read the details about ARFF, so you can get load your data into WEKA.
- IBM has its own data-mining software, and "Integrate InfoSphere Warehouse data mining with IBM Cognos reporting, Part 1" is a good starting point.
- To listen to interesting interviews and discussions for software developers, check out developerWorks podcasts.
- Stay current with developerWorks' Technical events and webcasts.
- Follow developerWorks on Twitter.
- Check out upcoming conferences, trade shows, webcasts, and other Events around the world that are of interest to IBM open source developers.
- Visit the developerWorks Open source zone for extensive how-to information, tools, and project updates to help you develop with open source technologies and use them with IBM's products, as well as our most popular articles and tutorials.
- The My developerWorks community is an example of a successful general community that covers a wide variety of topics.
- Watch and learn about IBM and open source technologies and product functions with the no-cost developerWorks On demand demos.

Get products and technologies

- Download WEKA to run it on your own system.
- You can also see specific details about the IBM DB2 Intelligent Miner software for comparison to WEKA.
- Innovate your next open source development project with IBM trial software, available for download or on DVD.

- Download IBM product evaluation versions or explore the online trials in the IBM SOA Sandbox and get your hands on application development tools and middleware products from DB2®, Lotus®, Rational®, Tivoli®, and WebSphere®.

Discuss

- Also, check out the new Data Mining group in My developerWorks.
- Participate in developerWorks blogs and get involved in the developerWorks community.

About the author

Michael Abernethy

In his 10 years in technology, Michael Abernethy has worked with a wide variety of technologies and a wide variety of clients. He currently works as the Product Development Manager for Optimal Auctions, an auction software company. His focus nowadays is on Rich Internet Applications and making them both more complex and simpler at the same time. When he's not working at his computer, he can be found on the beach in Mexico with a good book.