Regression with WEKA

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

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.

Building the data set 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.

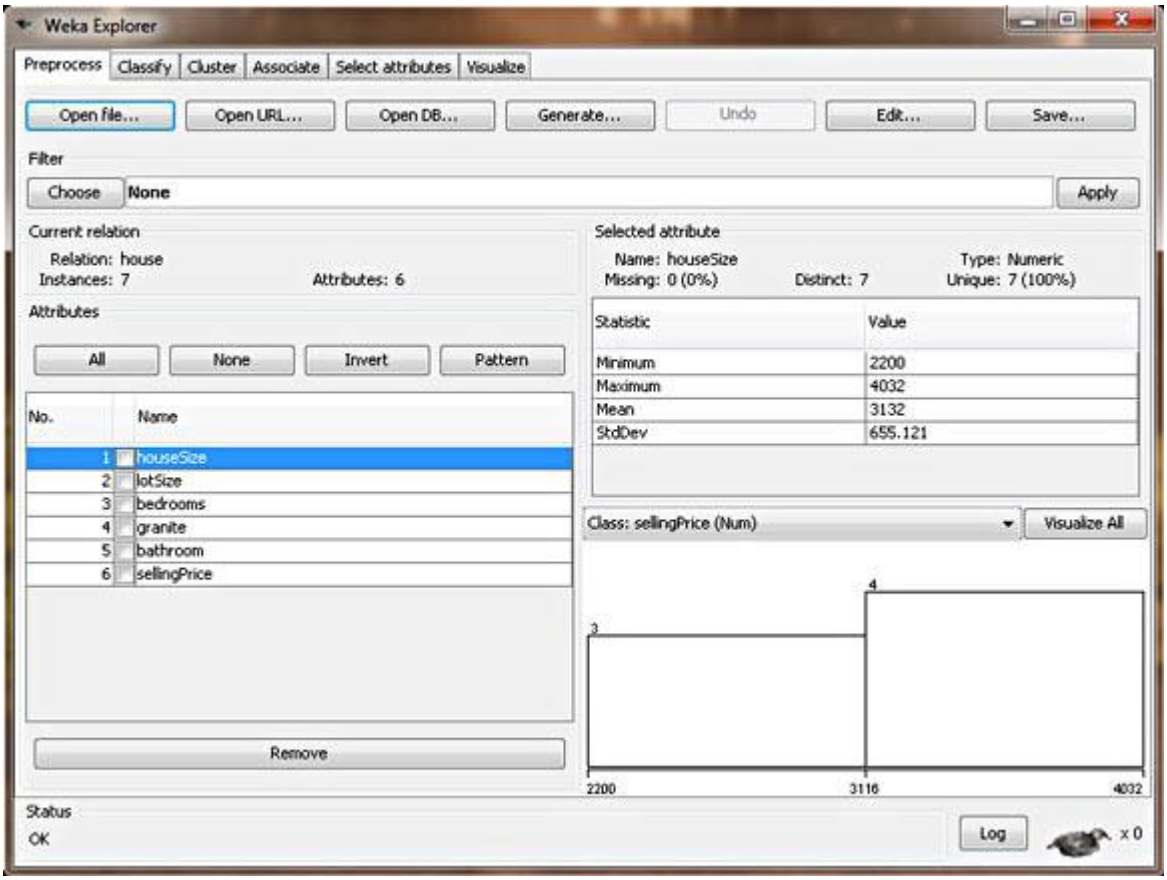
Listing 1. WEKA file format

```
1
2 @RELATION house
3
4 @ATTRIBUTE houseSize NUMERIC
5 @ATTRIBUTE lotSize NUMERIC
6 @ATTRIBUTE bedrooms NUMERIC
7 @ATTRIBUTE granite NUMERIC
8 @ATTRIBUTE bathroom NUMERIC
9 @ATTRIBUTE sellingPrice NUMERIC
10
11@DATA
123529,9191,6,0,0,205000
133247,10061,5,1,1,224900
144032,10150,5,0,1,197900
152397,14156,4,1,0,189900
162200,9600,4,0,1,195000
173536,19994,6,1,1,325000
182983,9365,5,0,1,230000
19
```

Loading the data into WEKA

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



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.

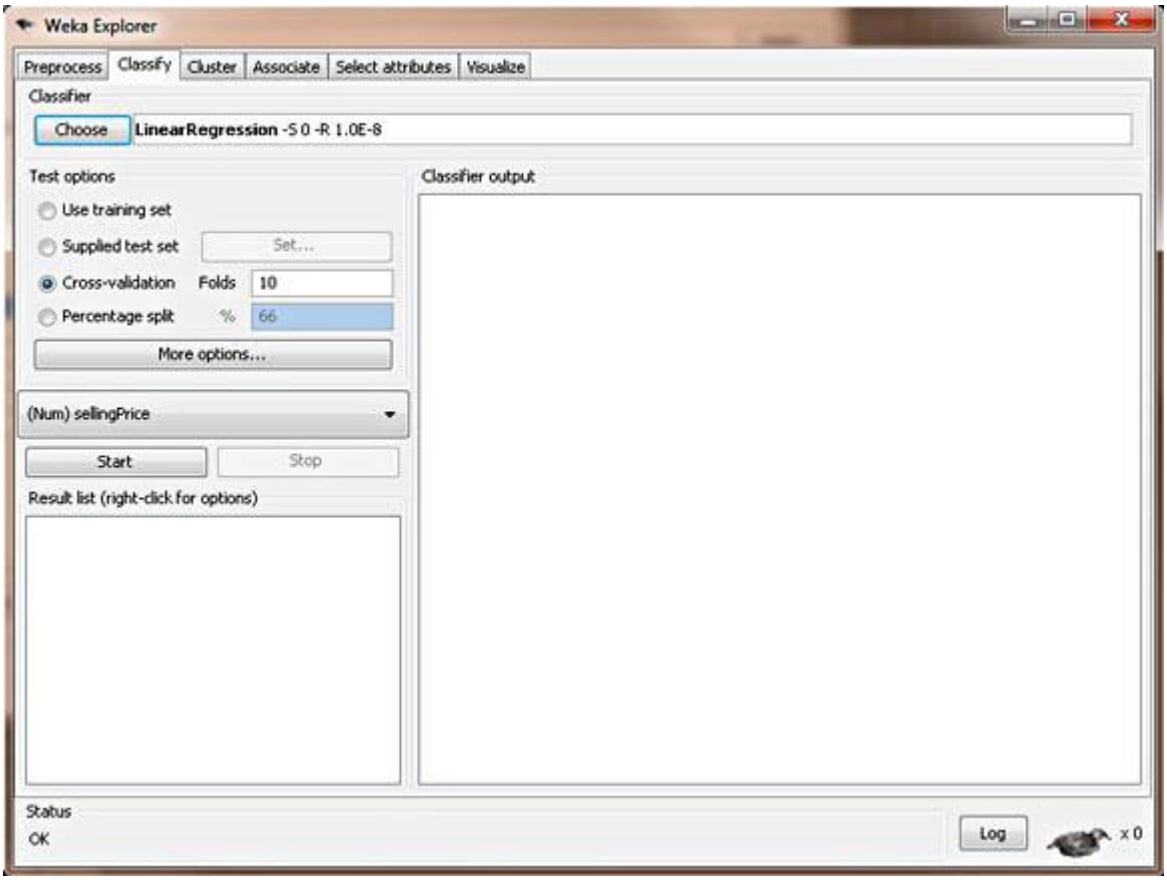
Creating the regression model with WEKA

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.

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.

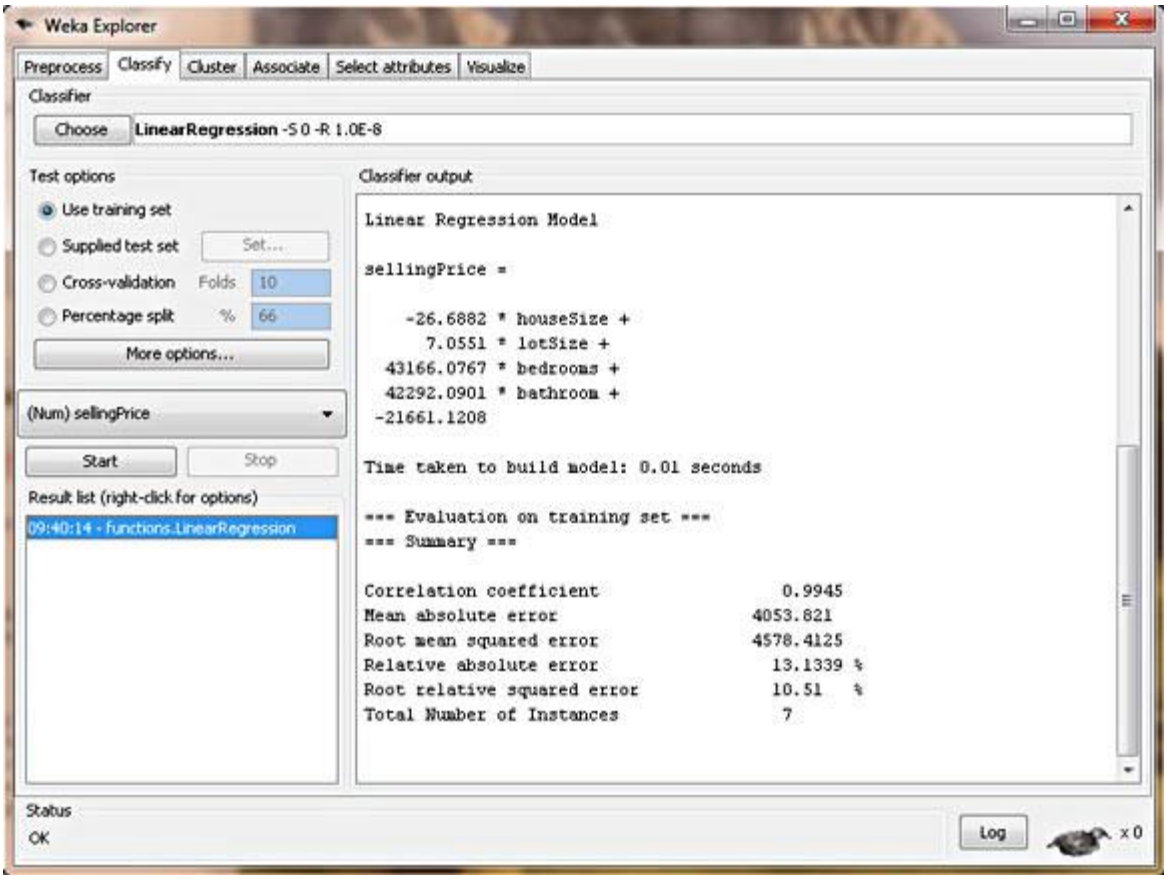
Now that the desired model has been chosen, we have to tell WEKA where the data is that it should use to build the model. Though it may be obvious to us that we want to use the

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.

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).

Conclusion

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.

Referencia: https://www.ibm.com/developerworks/library/os-weka1/index.html.