Machine Learning Mastery With Weka

Simple | Smart | Speed

CHOOLS

MACHINE
LEARNING
MASTERY

Machine Learning Mastery With Weka

Contents

Before We Get Started...

Machine learning is a fascinating study, but how do you actually use it on your own problems? You may be confused as to how best prepare your data for machine learning, which algorithms to use or how to choose one model over another. In this guide you will discover a 14-part crash course into applied machine learning using the Weka platform without a single mathematical equation or line of programming code.

After completing this mini course:

- You will know how to work through a dataset end-to-end and deliver a set of predictions or a high performance model.
- You will know your way around the Weka machine learning workbench including how to explore algorithms and design controlled experiments.
- You will know how to create multiple views of your problem, evaluate multiple algorithms and use statistics to choose the best performing model for your own predictive modeling problems.

Let's get started.

This is a long and useful guide. You might want to print it out.

Who Is This Mini-Course For?

Before we get started, let's make sure you are in the right place. The list below provides some general guidelines as to who this course was designed for. Don't panic if you don't match these points exactly, you might just need to brush up in one area or another to keep up.

You are a developer that knows a little machine learning. This means you know about some of the basics of machine learning like cross validation, some algorithms and the bias-variance trade-off. It does not mean that you are a machine learning PhD, just that you know the landmarks or know where to look them up.

This mini-course is not a textbook on machine learning. It will take you from a developer that knows a little machine learning to a developer who can use the Weka platform to work through a dataset from beginning to end and deliver a set of predictions or a high performance model.

Mini-Course Overview (what to expect)

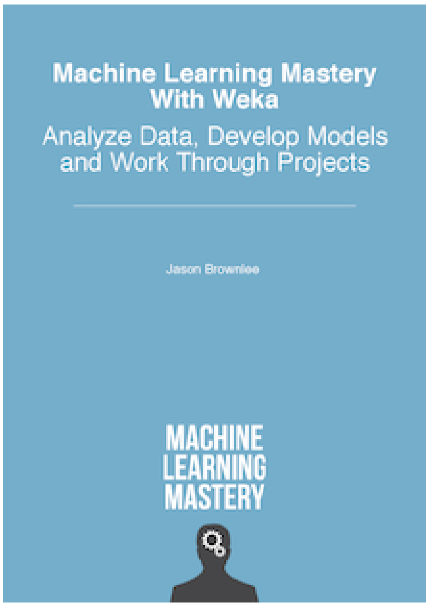
This mini-course is divided into 14 parts. Each lesson was designed to take you about 30 minutes. You might finish some much sooner and for others you may choose to go deeper and spend more time. You can complete each part as quickly or as slowly as you like. A comfortable schedule may be to complete one lesson per day over a two week period. Highly recommended. The topics you will cover over the next 14 lessons are as follows:

- **Lesson 01**: Download and Install Weka.
- **Lesson 02**: Load Standard Machine Learning Datasets.
- **Lesson 03**: Descriptive Stats and Visualization.
- **Lesson 04**: Rescale Your Data.
- **Lesson 05**: Perform Feature Selection on Your Data.
- **Lesson 06**: Machine Learning Algorithms in Weka.
- **Lesson 07**: Estimate Model Performance.
- **Lesson 08**: Baseline Performance On Your Data.
- **Lesson 09**: Classification Algorithms.
- **Lesson 10**: Regression Algorithms.
- **Lesson 11**: Ensemble Algorithms.
- **Lesson 12**: Compare the Performance of Algorithms.
- **Lesson 13**: Tune Algorithm Parameters.
- **Lesson 14**: Save Your Model.

You're going to have to do some work though, a little reading, a little tinkering in Weka. You want to get started in applied machine learning right? Here's a tip: All of the answers these lessons can be found on this blog
http://MachineLearningMastery.com. Use the search feature.

Hang in there, don't give up!

If you would like me to step you through each lesson in great detail (and much more), take a look at my book: **Machine Learning Mastery With Weka:**



Machine Learning Mastery
With Weka

Analyze Data, Develop Models
and Work Through Projects

Jason Brownlee

MACHINE
LEARNING
MASTERY

Learn more here:
https://machinelearningmastery.com/machine-learning-mastery-weka/

Lesson 01: Download and Install Weka

The first thing to do is install the Weka software on your workstation. Weka is free open source software. It is written in Java and can run on any platform that supports Java, including:

- Windows.
- Mac OS X.
- Linux.

You can download Weka as standalone software or as a version bundled with Java. If you do not already have Java installed on your system, I recommend downloading and installing a version bundled with Java.

1. Your task for this lesson is to visit the Weka download page[1], download and install Weka on your workstation.

Lesson 02: Load Standard Machine Learning Datasets

Now that you have Weka installed, you need to load data. Weka is designed to load data in a native format called ARFF. It is a modified CSV format that includes additional information about the types of each attribute (column). Your Weka installation includes a subdirectory with a number of standard machine learning datasets in ARFF format ready for you to load. Weka also supports loading data from raw CSV files as well as a database and converts the data to ARFF as needed. In this lesson you will load a standard dataset in the *Weka Explorer*.

1. Start Weka (click on the bird icon), this will start the *Weka GUI Chooser*.
2. Click the *Explorer* button, this will open the *Weka Explorer* interface.
3. Click the *Open file...* button and navigate to the `data/` directory in your Weka installation and load the `diabetes.arff` dataset.

Note, if you do not have a `data/` directory in your Weka installation, or you cannot find it, download the `.zip` version of Weka from the Weka download webpage[2], unzip it and access the `data/` directory.

You have just loaded your first dataset in Weka. Try loading some of the other datasets in the `data/` directory. Try downloading a raw CSV file from the UCI Machine Learning repository[3] and loading it in Weka.

Lesson 03: Descriptive Stats and Visualization

Once you can load data in Weka, it is important to take a look at it. Weka allows you to review descriptive statistics calculated from your data. It also provides visualization tools. In this lesson you will use Weka to learn more about your data.

1. Open the *Weka GUI Chooser*.
2. Open the *Weka Explorer*.
3. Load the `data/diabetes.arff` dataset.
4. Click on different attributes in the *Attributes* list and review the details in the *Selected attribute* pane.
5. Click the *Visualize All* button to review all attribute distributions.
6. Click the *Visualize* tab and review the scatter plot matrix for all attributes.

Get comfortable reviewing the details for different attributes in the *Preprocess* tab and tuning the scatter plot matrix in the *Visualize* tab.

Lesson 04: Rescale Your Data

Raw data is often not suitable for modeling. Often you can improve the performance of your machine learning models by rescaling attributes. In this lesson you will learn how to use data filters in Weka to rescale your data. You will normalize all of the attributes for a dataset, rescaling them to the consistent range of 0-to-1.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Choose* button in the *Filter* pane and select *unsupervised.attribute.Normalize*.
4. Click the *Apply* button.

Review the details for each attribute in the *Selected attribute* pane and note the change to the scale. Explore using other data filters such as the *Standardize* filter. Explore configuring filters by clicking on the name of the loaded filter and changing it's parameters. Test out saving modified datasets for later use by clicking the *Save...* button on the *Preprocess* tab.

Lesson 05: Perform Feature Selection on Your Data

Not all of the attributes in your dataset may be relevant to the attribute you want to predict. You can use feature selection to identify those attributes that are most relevant to your output variable. In this lesson you will get familiar with using different feature selection methods.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Select attributes* tab.
4. Click the *Choose* button in the *Attribute Evaluator* pane and select the *CorrelationAttributeEval*.
   (a) You will be presented with a dialog asking you to change to the *Ranker* search method, needed when using this feature selection method. Click the *Yes* button.
5. Click the *Start* button to run the feature selection method.

Review the output in the *Attribute selection output* pane and note the correlation scores for each attribute, the larger numbers indicating the more relevant features. Explore other feature selection methods such as the use of information gain (entropy). Explore selecting features to removal from your dataset in the *Preprocess* tab and the *Remove* button.

Lesson 06: Machine Learning Algorithms in Weka

A key benefit of the Weka workbench is the large number of machine learning algorithms it provides. You need to know your way around machine learning algorithms. In this lesson you will take a closer look at machine learning algorithms in Weka.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Classify* tab.
4. Click the *Choose* button and note the different groupings for algorithms.
5. Click the name of the selected algorithm to configure it.
6. Click the *More* button on the configuration window to learn more about the implementation.
7. Click the *Capabilities* button on the configuration window to learn more about how it can be used.
8. Note the *Open* and *Save* buttons on the window where different configurations can be saved and loaded.
9. Hover on a configuration parameter and note the tooltip help.
10. Click the *Start* button to run an algorithm.

Browse the algorithms available. Note that some algorithms are unavailable given whether your dataset is a classification (predict a category) or regression (predict a real value) type problem. Explore and learn more about the various algorithms available in Weka. Get confidence choosing and configuring algorithms.

Lesson 07: Estimate Model Performance

Now that you know how to choose and configure different algorithms, you need to know how to evaluate the performance of an algorithm. In this lesson you are going to learn about the different ways to evaluate the performance of an algorithm in Weka.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Classify* tab.

The *Test options* pane lists the various different techniques that you can use to evaluate the performance of an algorithm.

- The gold standard is 10-fold *Cross Validation*. This is selected by default. For a small dataset, the number of folds can be adjusted from 10 to 5 or even 3.
- If your dataset is very large and you want to evaluate algorithms quickly, you can use the *Percentage split* option. By default, this option will train on 66% of your dataset and use the remaining 34% to evaluate the performance of your model.
- Alternately, if you have a separate file containing a validation dataset, you can evaluate your model on that by selecting the *Supplied test set* option. Your model will be trained on the entire training dataset and evaluated on the separate dataset.
- Finally, you can evaluate the performance of your model on the whole training dataset. This is useful if you are more interested in a descriptive than a predictive model.

Click the *Start* button to run a given algorithm with your chosen test option. Experiment with different Test options. Further refine the test options in the configuration provided by clicking the *More options...* button.

Lesson 08: Baseline Performance On Your Data

When you start evaluating multiple machine learning algorithms on your dataset, you need a baseline for comparison. A baseline result gives you a point of reference to know whether the results for a given algorithm are good or bad, and by how much. In this lesson you will learn about the *ZeroR* algorithm that you can use as a baseline for classification and regression algorithms.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.

2. Load the `data/diabetes.arff` dataset.

3. Click the *Classify* tab. The *ZeroR* algorithm is chosen by default.

4. Click the *Start* button.

This will run the *ZeroR* algorithm using 10-fold cross validation on your dataset. The *ZeroR* algorithm also called the Zero Rule is an algorithm that you can use to calculate a baseline of performance for all algorithms on your dataset. It is the *worst* result and any algorithm that shows a better performance has some skill on your problem.

On a classification algorithm, the *ZeroR* algorithm will always predict the most abundant category. If the dataset has an equal number of classes, it will predict the first category value. On the diabetes dataset, this results in a classification accuracy of 65%. For regression problems, the *ZeroR* algorithm will always predict the mean output value.

Experiment with the *ZeroR* algorithm on a range of different datasets. It is the algorithm you should always run first before all others to develop a baseline.

Lesson 09: Tour of Classification Algorithms

Weka provides a large number of classification algorithms. In this lesson you will discover 5 top classification algorithms that you can use on your classification problems.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Classify* tab.
4. Click the *Choose* button.

5 Top algorithms that you can use for classification include:

- Logistic Regression (*functions.Logistic*).
- Naive Bayes (*bayes.NaiveBayes*).
- k-Nearest Neighbors (*lazy.IBk*).
- Classification and Regression Trees (*trees.REPTree*).
- Support Vector Machines (*functions.SMO*).

Experiment with each of these top algorithms. Try them out on different classification datasets, such as those with two classes and those with more.

Lesson 10: Tour of Regression Algorithms

Classification algorithms is Weka's specialty, but many of these algorithms can be used for regression. Regression is the prediction of a real valued outcome (like a dollar amount), different from classification that predicts a category (like *dog* or *cat*). In this lesson you will discover 5 top regression algorithms that you can use on your regression problems.

You can download a suite of standard regression machine learning datasets from the Weka dataset download webpage[4]. Download the `datasets-numeric.jar` archive of regression problems, titled:

- *A jar file containing 37 regression problems, obtained from various sources*

Use your favorite unzip program to unzip the `.jar` file and you will have a new directory called `numeric/` containing 37 regression problems that you can work with.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/housing.arff` dataset.
3. Click the *Classify* tab.
4. Click the *Choose* button.

5 Top algorithms that you can use for regression include:

- Linear Regression (*functions.LinearRegression*).
- Support Vector Regression (*functions.SMOReg*).
- k-Nearest Neighbors (*lazy.IBk*).
- Classification and Regression Trees (*trees.REPTree*).
- Artificial Neural Network (*functions.MultilayerPerceptron*).

Experiment with each of these top algorithms. Try them out on different regression datasets.

Lesson 11: Tour of Ensemble Algorithms

Weka is very easy to use and this may be its biggest advantage over other platforms. In addition to this, Weka provides a large suite of ensemble machine learning algorithms and this may be Weka's second big advantage over other platforms. It is worth spending your time to get good at using Weka's ensemble algorithms. In this lesson you will discover 5 top ensemble machine learning algorithms that you can use.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Classify* tab.
4. Click the *Choose* button.

5 Top ensemble algorithms that you can use include:

- Bagging (*meta.Bagging*).
- Random Forest (*trees.RandomForest*).
- AdaBoost (*meta.AdaBoost*).
- Voting (*meta.Voting*).
- Stacking (*meta.Stacking*).

Experiment with each of these top algorithms. Most of these ensemble methods let you choose the sub-models. Experiment using different combinations of sub-models. Combinations of techniques that work in very different ways and produce different predictions often result in better performance. Try them out on different classification and regression datasets.

Lesson 12: Compare the Performance of Algorithms

Weka provides a different tool specifically designed for comparing algorithms called the *Weka Experiment Environment*. The *Weka Experiment Environment* allows you to design and execute controlled experiments with machine learning algorithms and then analyze the results. In this lesson you will design your first experiment in Weka and discover how to use the *Weka Experiment Environment* to compare the performance of machine learning algorithms.

1. Open the *Weka Chooser GUI*.
2. Click the *Experimenter* button to open the *Weka Experiment Environment*.
3. Click the *New* button.
4. Click the *Add new...* button in the *Datasets* pane and select `data/diabetes.arff`.
5. Click the *Add new...* button in the *Algorithms* pane and add *ZeroR* and *IBk*.
6. Click the *Run* tab and click the *Start* button.
7. Click the *Analyse* tab and click the *Experiment* button and then the *Perform test* button.

You just designed, executed and analysed the results of your first controlled experiment in Weka. You compared the *ZeroR* algorithm to the *IBk* algorithm with default configuration on the diabetes dataset. The results show that *IBk* has a higher classification accuracy than *ZeroR* and that this difference is statistically significant (the little `v` character next to the result).

Expand the experiment and add more algorithms and rerun the experiment. Change the *Test base* on the *Analyse* tab to change which set of results is taken as the reference for comparison to the other results.

Lesson 13: Tune Algorithm Parameters

To get the most out of a machine learning algorithm you must tune the parameters of the method to your problem. You cannot know how to best do this beforehand, therefore you must try out lots of different parameters. The *Weka Experiment Environment* allows you to design controlled experiments to compare the results of different algorithm parameters and whether the differences are statistically significant. In this lesson you are going to design an experiment to compare the parameters of the k-Nearest Neighbors algorithm.

1. Open the *Weka Chooser GUI*.
2. Click the *Experimenter* button to open the *Weka Experiment Environment*
3. Click the *New* button.
4. Click the *Add new...* button in the *Datasets* pane and select `data/diabetes.arff`.
5. Click the *Add new...* button in the *Algorithms* pane and add 3 copes of the *IBk* algorithm.
6. Click each *IBk* algorithm in the list and click the *Edit selected...* button and change *KNN* to 1, 3, 5 for each of the 3 different algorithms.
7. Click the *Run* tab and click the *Start* button.
8. Click the *Analyse* tab and click the *Experiment* button and then the *Perform test* button.

You just designed, executed and analyzed the results of a controlled experiment to compare algorithm parameters. We can see that the results for large K values is better than the default of 1 and the difference is significant. Explore changing other configuration properties of *IBk* and build confidence in developing experiments to tune machine learning algorithms.

Lesson 14: Save Your Model

Once you have found a top performing model on your problem you need to finalize it for later use. In this final lesson you will discover how to train a final model and save it to a file for later use.

1. Open the *Weka GUI Chooser* and then the *Weka Explorer*.
2. Load the `data/diabetes.arff` dataset.
3. Click the *Classify* tab.
4. Change the *Test options* to *Use training set* and click the *Start* button.
5. Right click on the results in the *Result list* and click *Save model* and enter a filename like `diabetes-final`.

You have just trained a final model on the entire training dataset and saved the resulting model to a file. You can load this model back into Weka and use it to make predictions on new data.

1. Right-click on the *Result list* click *Load model* and select your model file (*diabetes-final.model*).
2. Change the *Test options* to *Supplied test set* and choose `data/diabetes.arff` (this could be a new file for which you do not have predictions)
3. Click *More options* in the *Test options* and change *Output predictions* to *Plain Text*
4. Right click on the loaded model and choose *Re-evaluate model on current test set.*

The new predictions will now be listed in the *Classifier output* pane. Experiment saving different models and making predictions for entirely new datasets.

Final Word Before You Go...

You made it. Well done! Take a moment and look back at how far you have come:

- You discovered how to start and use the *Weka Explorer* and *Weka Experiment Environment*, perhaps for the first time.
- You loaded data, analyzed it and used data filters and feature selection to prepare data for modeling.
- You discovered a suite of machine learning algorithms and how to design controlled experiments to evaluate their performance.

Don't make light of this, you have come a long way in a short amount of time. This is just the beginning of your journey in applied machine learning with Weka. Keep practicing and developing your skills.