A Tour of Machine Learning Algorithms

by **Jason Brownlee** on November 25, 2013 in **Machine Learning Algorithms**

In this post, we take a tour of the most popular machine learning algorithms.

It is useful to tour the main algorithms in the field to get a feeling of what methods are available.

There are so many algorithms available that it can feel overwhelming when algorithm names are thrown around and you are expected to just know what they are and where they fit.

I want to give you two ways to think about and categorize the algorithms you may come across in the field.

- The first is a grouping of algorithms by the **learning style**.
- The second is a grouping of algorithms by **similarity** in form or function (like grouping similar animals together).

Both approaches are useful, but we will focus in on the grouping of algorithms by similarity and go on a tour of a variety of different algorithm types.

After reading this post, you will have a much better understanding of the most popular machine learning algorithms for supervised learning and how they are related.

A cool example of an ensemble of lines of best fit. Weak members are grey, the combined prediction is red.
Plot from Wikipedia, licensed under public domain.

Algorithms Grouped by Learning Style

There are different ways an algorithm can model a problem based on its interaction with the experience or environment or whatever we want to call the input data.

It is popular in machine learning and artificial intelligence textbooks to first consider the learning styles that an algorithm can adopt.

There are only a few main learning styles or learning models that an algorithm can have and we'll go through them here with a few examples of algorithms and problem types that they suit.

This taxonomy or way of organizing machine learning algorithms is useful because it forces you to think about the roles of the input data and the model preparation process and select one that is the most appropriate for your problem in order to get the best result.

Let's take a look at four different learning styles in machine learning algorithms:

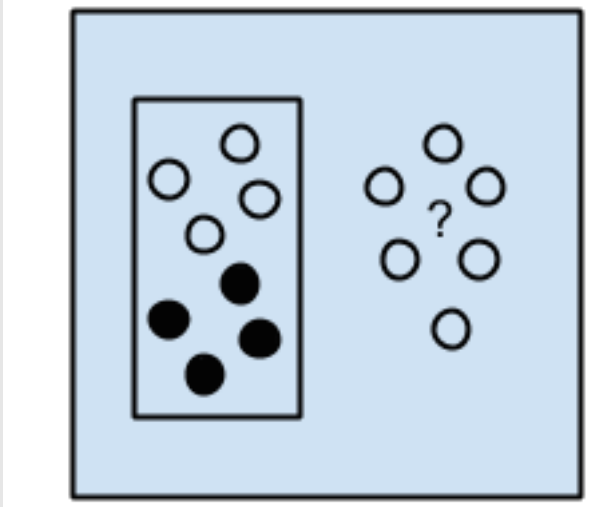
Supervised Learning

Input data is called training data and has a known label or result such as spam/not-spam or a stock price at a time.

A model is prepared through a training process in which it is required to make predictions and is corrected when those predictions are wrong. The training process continues until the model achieves a desired level of accuracy on the training data.

Example problems are classification and regression.

Example algorithms include Logistic Regression and the Back Propagation Neural Network.



Supervised Learning Algorithms

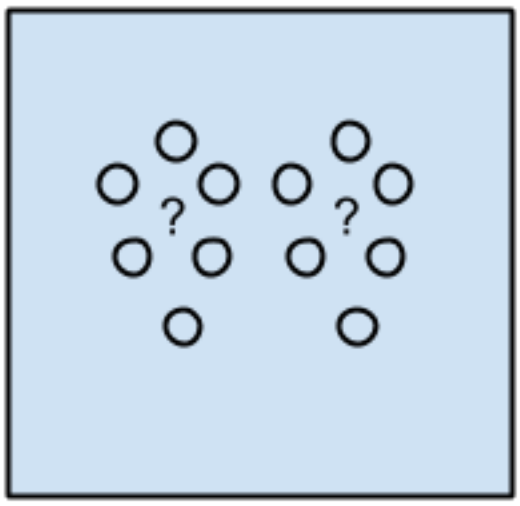
Unsupervised Learning

Input data is not labeled and does not have a known result.

A model is prepared by deducing structures present in the input data. This may be to extract general rules. It may be through a mathematical process to systematically reduce redundancy, or it may be to organize data by similarity.

Example problems are clustering, dimensionality reduction and association rule learning.

Example algorithms include: the Apriori algorithm and k-Means.



Unsupervised Learning Algorithms

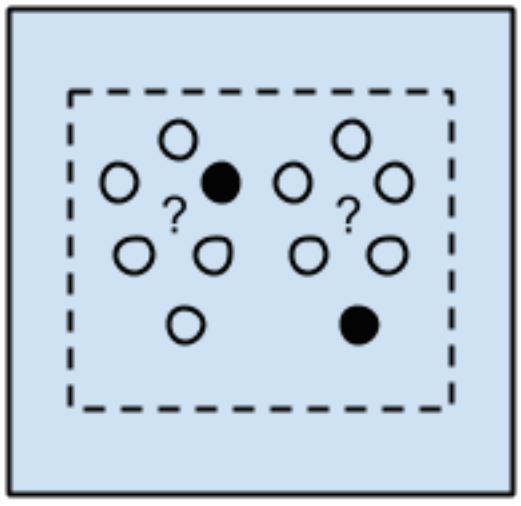
Semi-Supervised Learning

Input data is a mixture of labeled and unlabelled examples.

There is a desired prediction problem but the model must learn the structures to organize the data as well as make predictions.

Example problems are classification and regression.

Example algorithms are extensions to other flexible methods that make assumptions about how to model the unlabeled data.



Semi-supervised Learning Algorithms

Overview

When crunching data to model business decisions, you are most typically using supervised and unsupervised learning methods.

A hot topic at the moment is semi-supervised learning methods in areas such as image classification where there are large datasets with very few labeled examples.

Algorithms Grouped By Similarity

Algorithms are often grouped by similarity in terms of their function (how they work). For example, tree-based methods, and neural network inspired methods.

I think this is the most useful way to group algorithms and it is the approach we will use here.

This is a useful grouping method, but it is not perfect. There are still algorithms that could just as easily fit into multiple categories like Learning Vector Quantization that is both a neural network inspired method and an instance-based method. There are also categories that have the same name that describe the problem and the class of algorithm such as Regression and Clustering.

We could handle these cases by listing algorithms twice or by selecting the group that subjectively is the "best" fit. I like this latter approach of not duplicating algorithms to keep things simple.

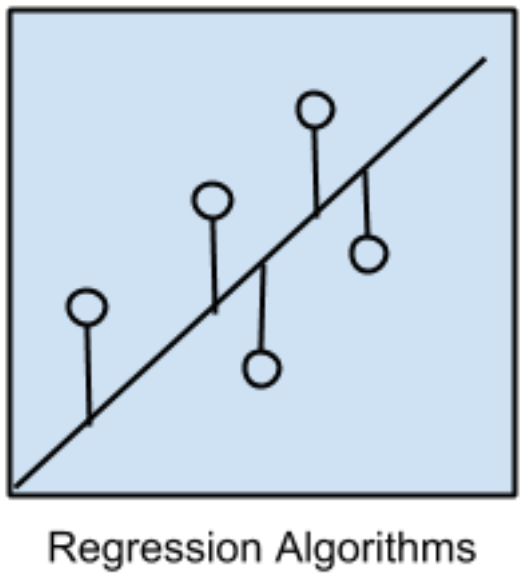
In this section, I list many of the popular machine learning algorithms grouped the way I think is the most intuitive. The list is not exhaustive in either the groups or the algorithms, but I think it is representative and will be useful to you to get an idea of the lay of the land.

Please Note: There is a strong bias towards algorithms used for classification and regression, the two most prevalent supervised machine learning problems you will encounter.

If you know of an algorithm or a group of algorithms not listed, put it in the comments and share it with us. Let's dive in.

Regression Algorithms

Regression is concerned with modeling the relationship between variables that is iteratively refined using a measure of error in the predictions made by the model.


Regression Algorithms

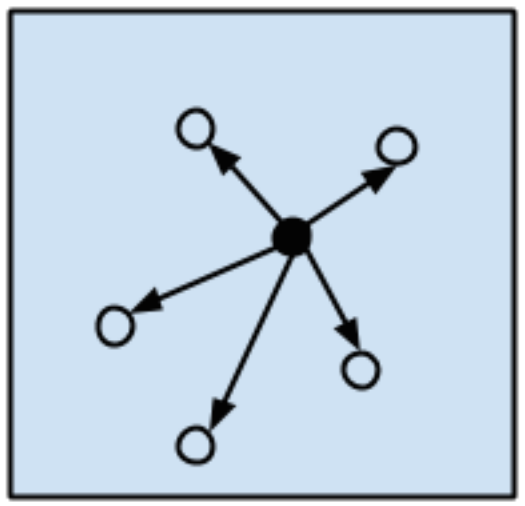
Regression methods are a workhorse of statistics and have been co-opted into statistical machine learning. This may be confusing because we can use regression to refer to the class of problem and the class of algorithm. Really, regression is a process.

The most popular regression algorithms are:

- Ordinary Least Squares Regression (OLSR)
- Linear Regression
- Logistic Regression
- Stepwise Regression
- Multivariate Adaptive Regression Splines (MARS)
- Locally Estimated Scatterplot Smoothing (LOESS)

Instance-based Algorithms

Instance-based learning model is a decision problem with instances or examples of training data that are deemed important or required to the model.


Instance-based Algorithms

Such methods typically build up a database of example data and compare new data to the database using a similarity measure in order to find the best match and make a prediction. For this reason, instance-based methods are also called winner-take-all methods and memory-based learning. Focus is put on the representation of the stored instances and similarity measures used between instances.

The most popular instance-based algorithms are:

- k-Nearest Neighbor (kNN)
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- Locally Weighted Learning (LWL)

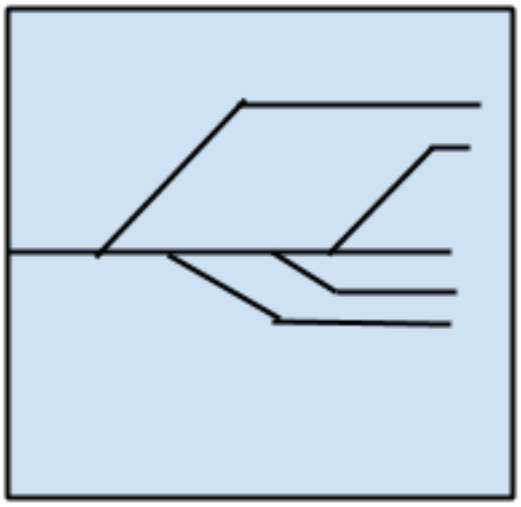
Regularization Algorithms

An extension made to another method (typically regression methods) that penalizes models based on their complexity, favoring simpler models that are also better at generalizing.

I have listed regularization algorithms separately here because they are popular, powerful and generally simple modifications made to other methods.

The most popular regularization algorithms are:

- Ridge Regression
- Least Absolute Shrinkage and Selection Operator (LASSO)
- Elastic Net
- Least-Angle Regression (LARS)



Regularization Algorithms

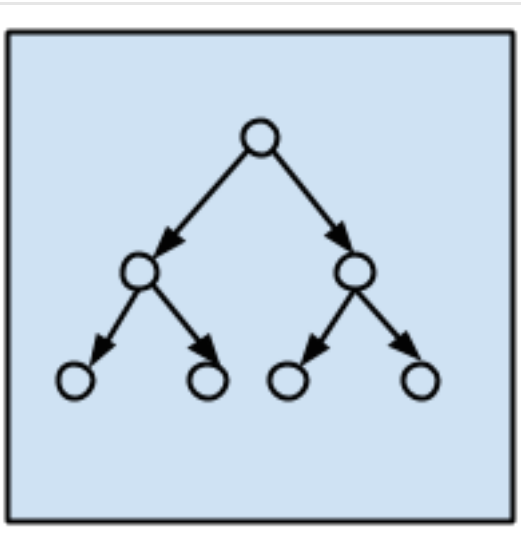
Decision Tree Algorithms

Decision tree methods construct a model of decisions made based on actual values of attributes in the data.

Decisions fork in tree structures until a prediction decision is made for a given record. Decision trees are trained on data for classification and regression problems. Decision trees are often fast and accurate and a big favorite in machine learning.

The most popular decision tree algorithms are:

- Classification and Regression Tree (CART)
- Iterative Dichotomiser 3 (ID3)
- C4.5 and C5.0 (different versions of a powerful approach)
- Chi-squared Automatic Interaction Detection (CHAID)
- Decision Stump
- M5
- Conditional Decision Trees



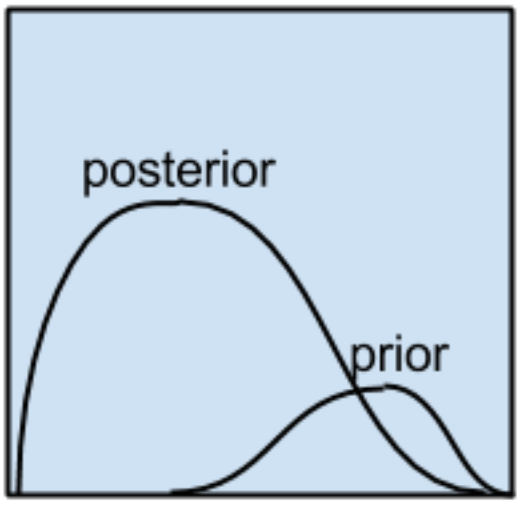
Decision Tree Algorithms

Bayesian Algorithms

Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification and regression.

The most popular Bayesian algorithms are:

- Naive Bayes
- Gaussian Naive Bayes
- Multinomial Naive Bayes
- Averaged One-Dependence Estimators (AODE)
- Bayesian Belief Network (BBN)
- Bayesian Network (BN)



Bayesian Algorithms

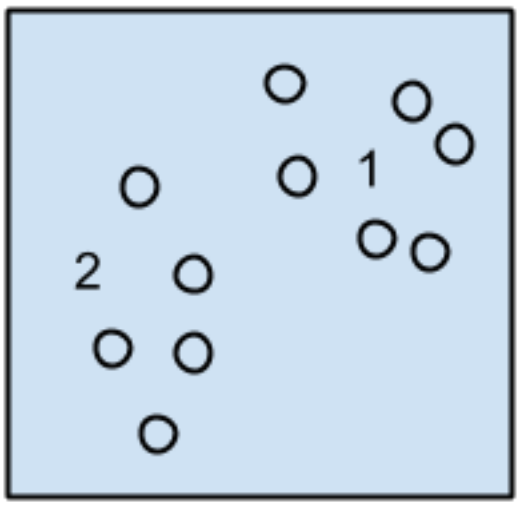
Clustering Algorithms

Clustering, like regression, describes the class of problem and the class of methods.

Clustering methods are typically organized by the modeling approaches such as centroid-based and hierarchal. All methods are concerned with using the inherent structures in the data to best organize the data into groups of maximum commonality.

The most popular clustering algorithms are:

- k-Means
- k-Medians
- Expectation Maximisation (EM)
- Hierarchical Clustering



Clustering Algorithms

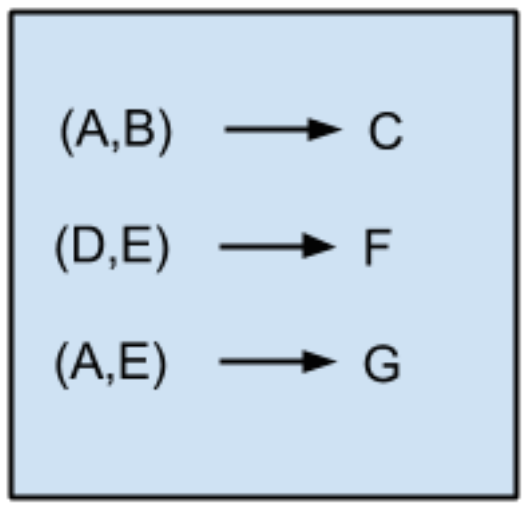
Association Rule Learning Algorithms

Association rule learning methods extract rules that best explain observed relationships between variables in data.

These rules can discover important and commercially useful associations in large multidimensional datasets that can be exploited by an organization.

The most popular association rule learning algorithms are:

- Apriori algorithm
- Eclat algorithm



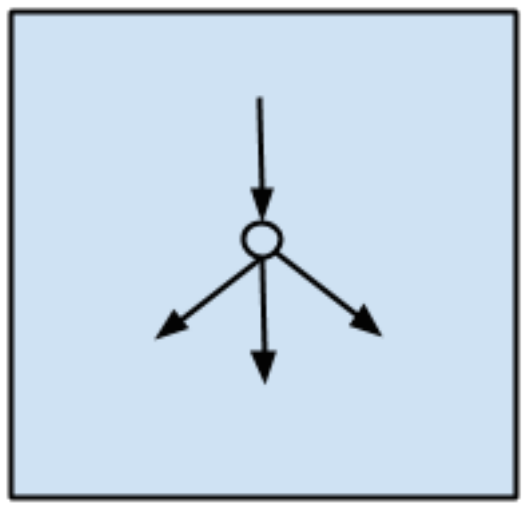
Association Rule Learning Algorithms

Artificial Neural Network Algorithms

Artificial Neural Networks are models that are inspired by the structure and/or function of biological neural networks.

They are a class of pattern matching that are commonly used for regression and classification problems but are really an enormous subfield comprised of hundreds of algorithms and variations for all manner of problem types.

Note that I have separated out Deep Learning from neural networks because of the massive growth and popularity in the field. Here we are concerned with the more classical methods.


Artificial Neural Network Algorithms

The most popular artificial neural network algorithms are:

- Perceptron
- Back-Propagation
- Hopfield Network
- Radial Basis Function Network (RBFN)

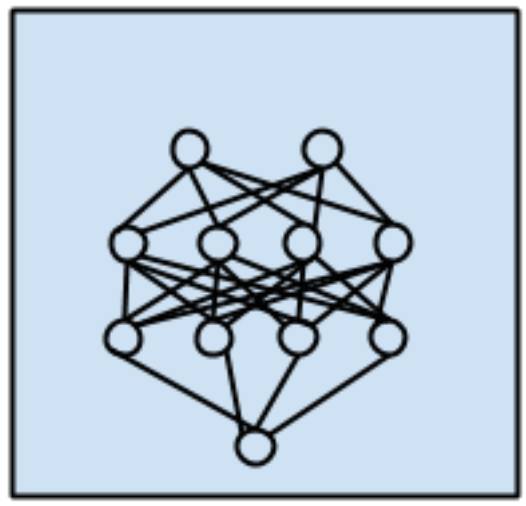
Deep Learning Algorithms

Deep Learning methods are a modern update to Artificial Neural Networks that exploit abundant cheap computation.

They are concerned with building much larger and more complex neural networks and, as commented on above, many methods are concerned with semi-supervised learning problems where large datasets contain very little labeled data.

The most popular deep learning algorithms are:

- Deep Boltzmann Machine (DBM)
- Deep Belief Networks (DBN)
- Convolutional Neural Network (CNN)
- Stacked Auto-Encoders



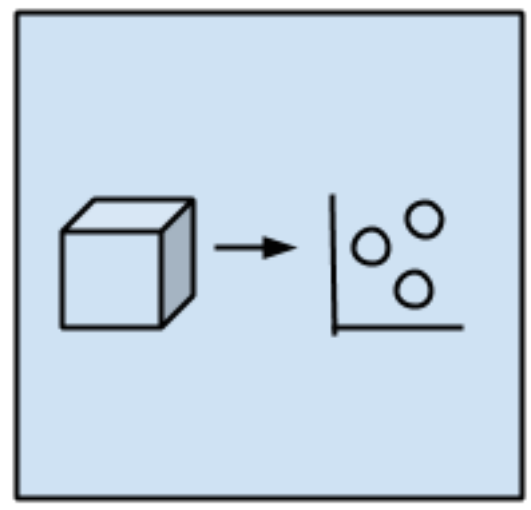
Deep Learning Algorithms

Dimensionality Reduction Algorithms

Like clustering methods, dimensionality reduction seek and exploit the inherent structure in the data, but in this case in an unsupervised manner or order to summarize or describe data using less information.

This can be useful to visualize dimensional data or to simplify data which can then be used in a supervised learning method. Many of these methods can be adapted for use in classification and regression.

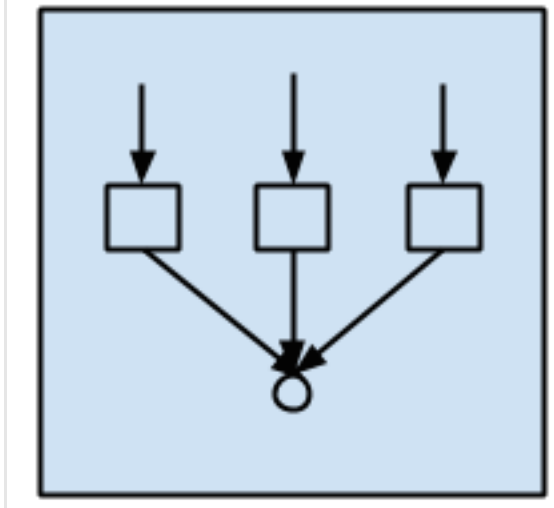
- Principal Component Analysis (PCA)
- Principal Component Regression (PCR)
- Partial Least Squares Regression (PLSR)
- Sammon Mapping
- Multidimensional Scaling (MDS)
- Projection Pursuit
- Linear Discriminant Analysis (LDA)
- Mixture Discriminant Analysis (MDA)
- Quadratic Discriminant Analysis (QDA)
- Flexible Discriminant Analysis (FDA)



Dimensional Reduction Algorithms

Ensemble Algorithms

Ensemble methods are models composed of multiple weaker models that are independently trained and whose predictions are combined in some way to make the overall prediction.



Ensemble Algorithms

Much effort is put into what types of weak learners to combine and the ways in which to combine them. This is a very powerful class of techniques and as such is very popular.

- Boosting
- Bootstrapped Aggregation (Bagging)
- AdaBoost
- Stacked Generalization (blending)
- Gradient Boosting Machines (GBM)
- Gradient Boosted Regression Trees (GBRT)
- Random Forest

Other Algorithms

Many algorithms were not covered.

For example, what group would Support Vector Machines go into? Its own?

I did not cover algorithms from specialty tasks in the process of machine learning, such as:

- Feature selection algorithms
- Algorithm accuracy evaluation
- Performance measures

I also did not cover algorithms from specialty subfields of machine learning, such as:

- Computational intelligence (evolutionary algorithms, etc.)
- Computer Vision (CV)
- Natural Language Processing (NLP)
- Recommender Systems
- Reinforcement Learning
- Graphical Models
- And more…

These may feature in future posts.

Further Reading

This tour of machine learning algorithms was intended to give you an overview of what is out there and

some ideas on how to relate algorithms to each other.

I've collected together some resources for you to continue your reading on algorithms. If you have a specific question, please leave a comment.

Other Lists of Algorithms

There are other great lists of algorithms out there if you're interested. Below are few hand selected examples.

- List of Machine Learning Algorithms: On Wikipedia. Although extensive, I do not find this list or the organization of the algorithms particularly useful.
- Machine Learning Algorithms Category: Also on Wikipedia, slightly more useful than Wikipedias great list above. It organizes algorithms alphabetically.
- CRAN Task View: Machine Learning & Statistical Learning: A list of all the packages and all the algorithms supported by each machine learning package in R. Gives you a grounded feeling of what's out there and what people are using for analysis day-to-day.
- Top 10 Algorithms in Data Mining: Published article and now a book (Affiliate Link) on the most popular algorithms for data mining. Another grounded and less overwhelming take on methods that you could go off and learn deeply.

How to Study Machine Learning Algorithms

Algorithms are a big part of machine learning. It's a topic I am passionate about and write about a lot on this blog. Below are few hand selected posts that might interest you for further reading.

- How to Learn Any Machine Learning Algorithm: A systematic approach that you can use to study and understand any machine learning algorithm using "algorithm description templates" (I used this approach to write my first book).
- How to Create Targeted Lists of Machine Learning Algorithms: How you can create your own systematic lists of machine learning algorithms to jump start work on your next machine learning problem.
- How to Research a Machine Learning Algorithm: A systematic approach that you can use to research machine learning algorithms (works great in collaboration with the template approach listed above).
- How to Investigate Machine Learning Algorithm Behavior: A methodology you can use to understand how machine learning algorithms work by creating and executing very small studies into their behavior. Research is not just for academics!
- How to Implement a Machine Learning Algorithm: A process and tips and tricks for implementing machine learning algorithms from scratch.

How to Run Machine Learning Algorithms

Sometimes you just want to dive into code. Below are some links you can use to run machine learning

algorithms, code them up using standard libraries or implement them from scratch.

- How To Get Started With Machine Learning Algorithms in R: Links to a large number of code examples on this site demonstrating machine learning algorithms in R.
- Machine Learning Algorithm Recipes in scikit-learn: A collection of Python code examples demonstrating how to create predictive models using scikit-learn.
- How to Run Your First Classifier in Weka: A tutorial for running your very first classifier in Weka (**no code required!**).

Final Word

I hope you have found this tour useful.

Please, leave a comment if you have any questions or ideas on how to improve the algorithm tour.

Update #1: Continue the discussion on HackerNews and reddit.

Update #2: I've added a bunch more resources and more algorithms. I've also added a handy mind map that you can download (see above).