Machine Learning Explained: Algorithms Are Your Friend

Data Science Basics | Machine Learning | Data Analysis | January 19, 2017 | Robert Kelley
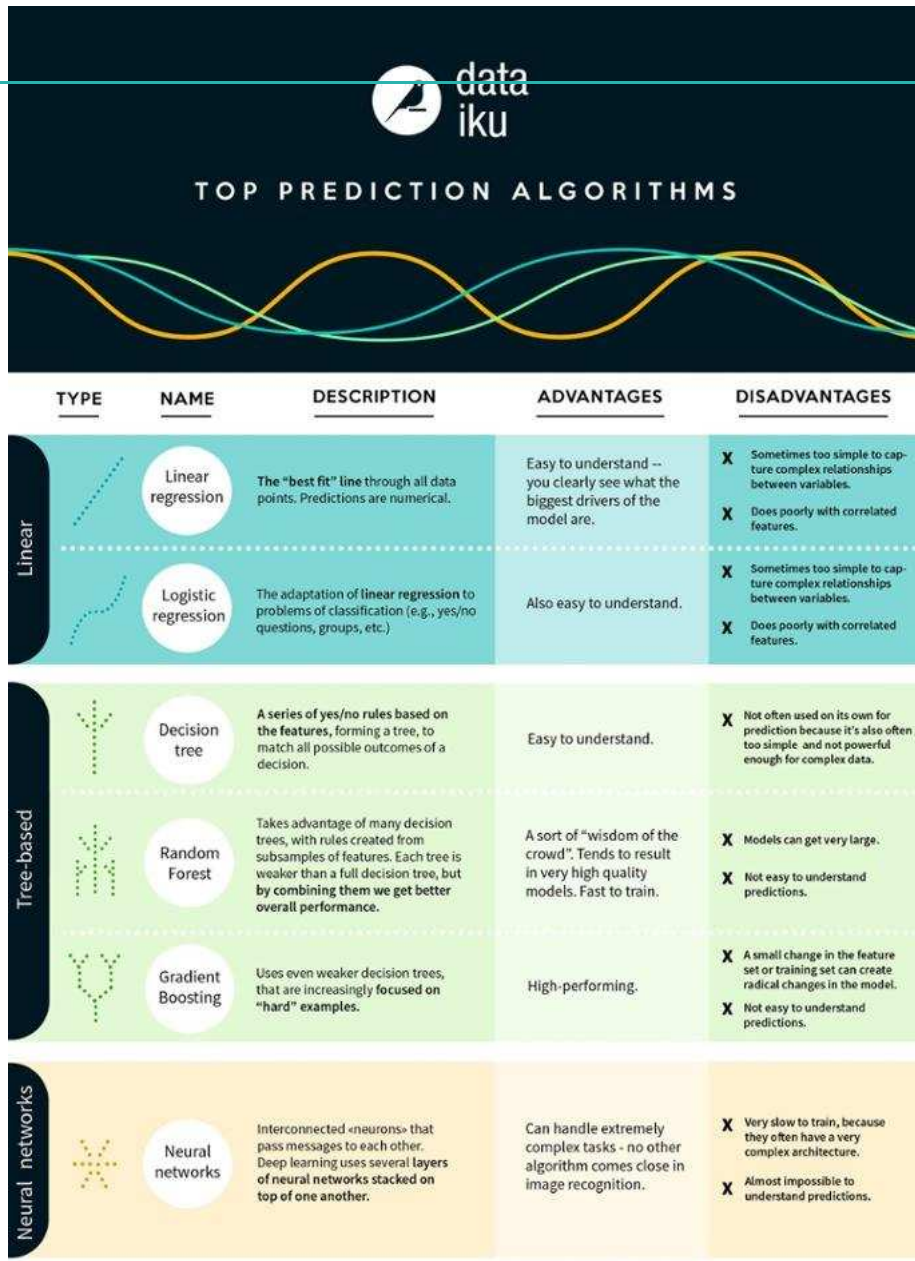(https://blog.dataiku.com/author/robert-kelley)

Tweet (https://twitter.com/share) | Share | | Like 115 | | Share |

We hear the term "machine learning" a lot these days, usually in the context of predictive analysis and artificial intelligence. Machine learning is, more or less, a way for computers to learn things without being specifically programmed. But how does that actually happen?

The answer is, in one word, algorithms. Algorithms are sets of rules that a computer is able to follow. Think about how you learned to do long division -- maybe you learned to take the denominator and divide it into the first digits of the numerator, then subtracting the subtotal and continuing with the next digits until you were left with a remainder. Well, that's an algorithm, and it's the sort of thing we can program into a computer, which can perform these sorts of calculations much, much faster than we can
Discover the fundamentals of machine learning with our free guidebook.
(https://cta-service-cms2.hubspot.com/ctas/v2/public/cs/c/?cta_guid=b2d0b2de-5a99-439f-bc93-3e3e95bac75f&placement_guid=e428bc67-4037-4e76-a03c-30876d9a52fb&portal_id=2123903&redirect_url=APefjpF9dmm1TZYOmTbKCoAo0ukBVEN-Ii39-TlLkEzqKDomIddnbK5P0PyyfjQxTLePB2L7uz9n6SroqDUNzX0NDY8Igrhm_1Z4O4pwYjzYsLfZswnO5WMKj_Tz262Uj6oZ13GwAxBj_dJ2qAKOznGQRF5NDGrwYqs8kp09RbayXjpV6FmVbiNRW0ROhJ-C4Ms3y2NEd0BFdIoi8AecFwc0QTxReFltWjWMdYbXS4QSxmcpWc5aNMueVX-ecbIrifKrU5N-ZFvVKf-MPxz57fpaQVyyg-Z9XNOsa49CyJghtKJejBbhvCA&hsutk=361bfb9efd13441e3e7dbbd647ca4603&canon=https%3A%2F%2Fblog.dataiku.com%2Fmachine-learning-explained-algorithms-are-your-friend&click=7df3cf3c-bd22-44da-b674-03231662ea3a&utm_referrer=https%3A%2F%2Fwww.datasciencecentral.com%2Fprofiles%2Fblogs%2Fprediction-algorithms-in-one-picture&pageId=4622395186&__hstc=186155446.361bfb9efd13441e3e7dbbd647ca4603.1527251482046.1527511619279.1528494139012.3&__hssc=186155446.1.1528494139012&__hsfp=3576934704)

(This blog post and infographic are part of our Machine Learning Basics illustrated guidebook -- click here (https://pages.dataiku.com/machine-learning-basics-illustrated-guidebook) for a free copy.)

We've put together a brief summary of the top algorithms used in predictive analysis, which you can see just below. Read on for more detail on these algorithms.

(//pages.dataiku.com/hubfs/Top%20Prediction%20Algorithms.jpg?t=1528490886919)

What Does Machine Learning Look Like?

In machine learning, our goal is either prediction or clustering. Today, we're going to focus on prediction (we'll cover clustering in a future article). Prediction is a process where, from a set of input variables, we estimate the value of an output variable. For example, using a set of characteristics of a house, we can predict its sale price.Prediction problems are divided into two main categories:

- **Regression** problems, where the variable to predict is numerical (e.g., the price of a house)
- **Classification** problems, where the variable to predict is part of one of some number of pre-defined categories, which can be as simple as "yes" or "no." (for example, predict whether a certain piece of equipment will experience a mechanical failure)

With that in mind, today we're not going to reveal a secret to do prediction better than the computers, or even how to be a data scientist! What we are going to do is introduce the most prominent and common algorithms used in machine learning historically and today.

These algorithms come in three groups: linear models, tree-based models, and neural networks.

Linear Model Approach

A **linear model** uses a simple formula to find the "best fit" line through a set of data points. This methodology dates back over 200 years, and it has been used widely throughout statistics and machine learning. It is useful for statistics because of its simplicity -- the variable you want to predict (the dependent variable) is represented as an equation of variables you know (independent variables), and so prediction is just a matter of inputting the independent variables and having the equation spit out the answer.

For example, you might want to know how long it will take to bake a cake, and your regression analysis might yield an equation $t = 0.5x + 0.25y$, where t is the baking time in hours, x is the weight of the cake batter in kg, and y is a variable which is 1 if it is chocolate and 0 if it is not. If you have 1 kg of chocolate cake batter (we love cake), then you plug your variables into our equation, and you get $t = (0.5 \times 1) + (0.25 \times 1) = 0.75$ hours, or 45 minutes.

Linear Regression

Linear regression, or more specifically "least squares regression," is the most standard form of linear model. For regression problems, linear regression is **the most simple linear model**. Its drawback is that there is a tendency for the model to "overfit" -- that is, for the model to adapt too exactly to the data on which it has been trained at the expense of the ability to generalize to previously unseen data. For this reason, linear regression (along with logistic regression, which we'll get to in a second) in machine learning is often "regularized," which means the model has certain penalties to prevent overfit.

Another drawback of linear models is that, since they're so simple, they tend to have trouble predicting more complex behaviors when the input variables are not independent.

Logistic Regression

Logistic regression is simply the adaptation of linear regression to classification problems (once again, discussed above). The drawbacks of logistic regression are the same as those of linear regression. Because it maps values between 0 and 1, it is suited for classification problems as it can represent the probabilities of being in each class.

data iku (http://www.dataiku.com/)

Tree-Based Model Approach

When you hear tree-based, think decision trees, i.e., a sequence of branching operations.

Decision Tree

A **decision tree** is a graph that uses a branching method to show each possible outcome of a decision. Like if you're ordering a salad, you first decide the type of lettuce, then the toppings, then the dressing. We can represent all possible outcomes in a decision tree. In machine learning, the branches used are binary yes/no answers.

To train a decision tree, we take the train data set (that is, the data set that we use to train the model) and **find which attribute best "splits" the train set** with regards to the target. For example, in a fraud detection case, we could find that the attribute which best predicts the risk of fraud is the country. After this first split, we have two subsets which are the best at predicting if we only know that first attribute. Then we can iterate on the second-best attribute for each subset and resplit each subset, continuing until we have used enough of the attributes to satisfy our needs.

Random Forest

A **random forest** is the average of many decision trees, each of which is trained with a random sample of the data. Each single tree in the forest is weaker than a full decision tree, but by putting them all together, we get better overall performance thanks to diversity.

Random forest is a very popular algorithm in machine learning today. **It is very easy to train, and it tends to perform quite well.** Its downside is that it can be slow to output predictions relative to other algorithms, so you might not use it when you need lightning-fast predictions.

Gradient Boosting

Gradient boosting, like random forest, is also made from "weak" decision trees. The big difference is that in gradient boosting, the trees are trained one after another. Each subsequent tree is trained primarily with data that had been incorrectly predicted by previous trees. This allows gradient boost to gradually focus less on the easy-to-predict cases and more on difficult cases.

Gradient boosting **performs very well.** However, small changes in the training data set can create radical changes in the model, so it may not produce the most explainable results.

Neural Networks

Neural networks refer to a biological phenomenon comprised of interconnected neurons that exchange messages with each other. This idea has now been adapted to the world of machine learning and is called ANN (Artificial Neural Networks). **Deep learning**, which you've heard a lot about, can be done with several layers of neural networks put one after the other.

ANNs are a family of models that are taught to adopt cognitive skills. No other algorithms can handle **extremely complex tasks**, such as image recognition, as well as neural networks. However, just like the human brain, it takes a very long time to train the model, and it requires a lot of power (just think about how much we eat to keep our brains working!).

If you're interested in learning more, check out our illustrated guidebook on the basics of machine learning (https://pages.dataiku.com/machine-learning-basics-illustrated-guidebook). Enjoy, and keep in touch!
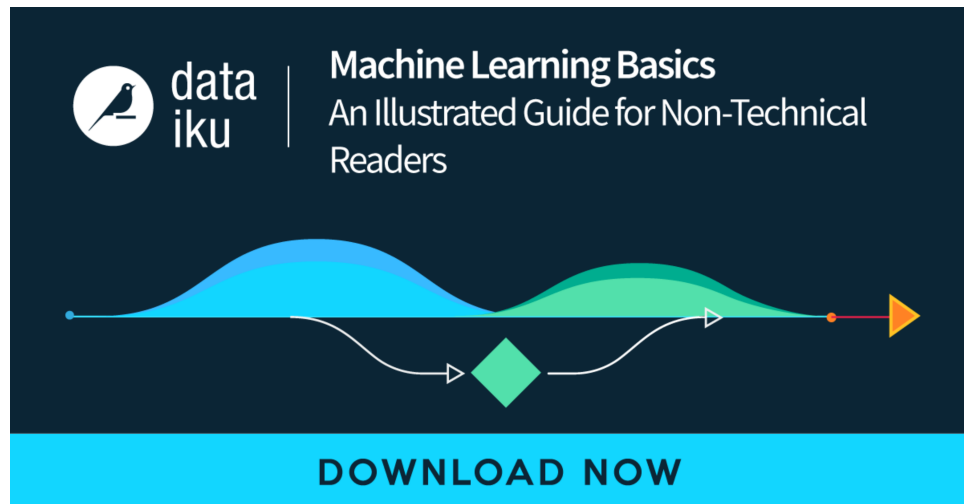
(https://cta-service-cms2.hubspot.com/ctas/v2/public/cs/c/?cta_guid=90b00821-88ed-4d77-a122-fd41c3f2d1c6&placement_guid=012ac079-edb4-481e-b008-eb9defee1011&portal_id=2123903&redirect_url=APefjpGI2IfXq-hP5pBwCBAC8uJ5y2kcrfA-QrnCUHpEEU0Qju-D2F2DwpKFY3TfcjaGh6gTNnpCDcueVFfuKw3y4Tr2-r6pQvnO-_KIkGGhKQgV7Pn1HSwGya9HGN5L_JjI7X3Kd9ubd_IWSy0ktrFMj08rpDJ3eqhMwTxqMXsxL_FaSxmp32XvZmo euZC_XFvq7k023Yz9QsgtgaGehqi8QZO9YNIqTWHsejObeVOce10YkmkLAqUKebQiLxXc_E5J_Q0XECpxiLze3m EZrlVHvJ0E-x1rFQ&hsutk=361bfb9efd13441e3e7dbbd647ca4603&canon=https%3A%2F%2Fblog.dataiku.com%2Fmachi ne-learning-explained-algorithms-are-your-friend&click=4c80ab8e-0f08-41d1-b1c0-5c0e6e8c6437&utm_referrer=https%3A%2F%2Fwww.datasciencecentral.com%2Fprofiles%2Fblogs%2Fpre diction-algorithms-in-one-picture&pageId=4622395186&__hstc=186155446.361bfb9efd13441e3e7dbbd647ca4603.1527251482046.152 7511619279.1528494139012.3&__hssc=186155446.1.1528494139012&__hsfp=3576934704)

Other Content You May Like

Disease Outbreak Assistance with Data: NATO ACT Innovation Challenge (https://blog.dataiku.com/data-science-and-disease-outbreak-assistance-nato-act-innovation-challenge)
Data Science, Machine Learning, Defense | June 05, 2018 | **Remi Meunier**