Top 5 Classification Algorithms in Machine Learning

[Classification](https://monkeylearn.com/text-classification/) is the process of recognizing, understanding, and grouping ideas and objects into preset categories or “sub-populations.” Using pre-categorized training datasets, machine learning programs use a variety of algorithms to classify future datasets into categories.

Classification algorithms in [machine learning](https://monkeylearn.com/machine-learning/) use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories.

Popular Classification Algorithms:

* [Logistic Regression](https://monkeylearn.com/blog/classification-algorithms/#logistic-regression)
* [Naive Bayes](https://monkeylearn.com/blog/classification-algorithms/#naive-bayes)
* [K-Nearest Neighbors](https://monkeylearn.com/blog/classification-algorithms/#knn)
* [Decision Tree](https://monkeylearn.com/blog/classification-algorithms/#decision-tree)
* [Support Vector Machines](https://monkeylearn.com/blog/classification-algorithms/#svm)

**Logistic Regression**

Logistic regression is a calculation used to predict a binary outcome: either something happens, or does not. This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc.

Independent variables are analyzed to determine the binary outcome with the results falling into one of two categories. The independent variables can be categorical or numeric, but the dependent variable is always categorical. Written like this:

***P(Y=1|X) or P(Y=0|X)***

It calculates the probability of dependent variable *Y*, given independent variable *X*.

This can be used to calculate the probability of a word having a positive or negative connotation (0, 1, or on a scale between). Or it can be used to determine the object contained in a photo (tree, flower, grass, etc.), with each object given a probability between 0 and 1.

**Naive Bayes**

[Naive Bayes](https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/) calculates the possibility of whether a data point belongs within a certain category or does not. In [text analysis](https://monkeylearn.com/text-analysis/), it can be used to categorize words or phrases as belonging to a preset “tag” (classification) or not. For example:



To decide whether or not a phrase should be tagged as “sports,” you need to calculate:

P(A/B) : *the probability of A, if B is true, is equal to the probability of B, if A is true, times the probability of A being true, divided by the probability of B being true*.

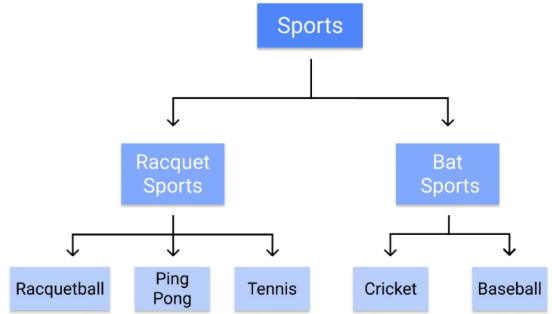
**K-nearest Neighbors**

K-nearest neighbors (k-NN) is a pattern recognition algorithm that uses training datasets **to find the *k* closest relatives** in future examples.

When k-NN is used in classification, you calculate to place data within the category of its nearest neighbor. If *k* = 1, then it would be placed in the class nearest 1. *K* is classified by a plurality poll of its neighbors.

**Decision Tree**

A decision tree is a supervised learning algorithm that is perfect **for classification problems**, as it’s able to order classes on a precise level. It works like a flow chart, separating data points into two similar categories at a time from the “tree trunk” to “branches,” to “leaves,” where the categories become more finitely similar. This creates categories within categories, allowing for organic classification with limited human supervision.



**Random Forest**

The random forest algorithm is an expansion of decision tree, in that you first construct a multitude of decision trees with training data, then fit your new data within one of the trees as a “random forest.”

It, essentially, averages your data to connect it to the nearest tree on the data scale. Random forest models are helpful as they remedy for the decision tree’s problem of “forcing” data points within a category unnecessarily.

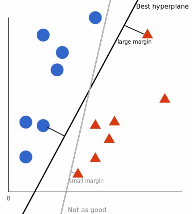
**Support Vector Machines**

A [support vector machine (SVM)](https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/) uses algorithms to train and classify data within degrees of polarity, taking it to a degree beyond *X/Y* prediction.

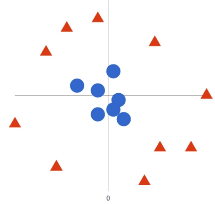
For a simple visual explanation, we’ll use two tags: *red* and *blue*, with two data features: *X* and *Y*, then train our classifier to output an*X/Y* coordinate as either *red* or *blue*.

The SVM then assigns a hyperplane that best separates the tags. In two dimensions this is simply a line. Anything on one side of the line is *red* and anything on the other side is *blue*. In sentiment analysis, for example, this would be *positive* and *negative*.

In order to maximize machine learning, the best hyperplane is the one with the largest distance between each tag:

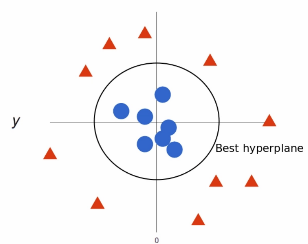


However, as data sets become more complex, it may not be possible to draw a single line to classify the data into two camps:



Using SVM, the more complex the data, the more accurate the predictor will become. Imagine the above in three dimensions, with a *Z-axis* added, so it becomes a circle.

Mapped back to two dimensions with the best hyperplane, it looks like this:



SVM allows for more accurate machine learning because it’s multidimensional.

4 Applications of Classification Algorithms

Okay, so now we understand a bit of the mathematics behind classification, but what can these machine learning algorithms do with real-world data?

* [Sentiment Analysis](https://monkeylearn.com/blog/classification-algorithms/#sentiment)
* [Email Spam Classification](https://monkeylearn.com/blog/classification-algorithms/#spam)
* [Document Classification](https://monkeylearn.com/blog/classification-algorithms/#document)
* [Image Classification](https://monkeylearn.com/blog/classification-algorithms/#image)

Sentiment Analysis

[Sentiment analysis](https://monkeylearn.com/sentiment-analysis/) is a machine learning text analysis technique that assigns sentiment (opinion, feeling, or emotion) to words within a text, or an entire text, on a polarity scale of *Positive*, *Negative*, or *Neutral*.

It can automatically read through thousands of pages in minutes or constantly monitor social media for posts about you. The tweet below, for example, about the messaging app, [Slack](https://slack.com/), would be analyzed to pull all of the individual statements as *Positive*. This allows companies to follow product releases and marketing campaigns in real-time, to see how customers are reacting.

Using advanced machine learning algorithms, sentiment analysis models can be trained to read for things like sarcasm and misused or misspelled words. Once properly trained, models produce consistently accurate results in a fraction of the time it would take humans.

 Dive right in to try MonkeyLearn’s [pre-trained sentiment classification tool](https://monkeylearn.com/sentiment-analysis-online/). Or learn how to build your own [sentiment classifier](https://monkeylearn.com/blog/sentiment-classification/) to the language and needs of your business.

Email Spam Classification

One of the most common uses of classification, working non-stop and with little need for human interaction, email spam classification saves us from tedious deletion tasks and sometimes even costly phishing scams.

Email applications use the above algorithms to calculate the likelihood that an email is either not intended for the recipient or unwanted spam. Using text analysis classification techniques, spam emails are weeded out from the regular inbox: perhaps a recipient’s name is spelled incorrectly, or certain scamming keywords are used.

Spam classifiers do still need to be trained to a degree, as we’ve all experienced when signing up for an email list of some sort that ends up in the spam folder.

Document Classification

Document classification is the ordering of documents into categories according to their content. This was previously done manually, as in the library sciences or hand-ordered legal files. Machine learning classification algorithms, however, allow this to be performed automatically.

Document classification differs from text classification, in that, entire documents, rather than just words or phrases, are classified. This is put into practice when using search engines online, cross-referencing topics in legal documents, and searching healthcare records by drug and diagnosis.

Image Classification

Image classification assigns previously trained categories to a given image. These could be the subject of the image, a numerical value, a theme, etc. Image classification can even use multi-label image classifiers, that work similarly to [multi-label text classifiers](https://monkeylearn.com/blog/multi-label-classification/), to tag an image of a stream, for example, into different labels, like “stream,” “water,” “outdoors,” etc.

Using [supervised learning algorithms](https://monkeylearn.com/machine-learning/#supervised-learning), you can tag images to train your model for appropriate categories. As with all machine learning models, the more you train it, the better it will work.

Wrap Up

Machine learning classification uses the mathematically provable guide of algorithms to perform analytical tasks that would take humans hundreds of more hours to perform. And with the proper algorithms in place and a properly trained model, classification programs perform at a level of accuracy that humans could never achieve.

[MonkeyLearn](https://monkeylearn.com/) is a text analysis platform with dozens of tools to move your business forward with data-driven insights. Try the pre-trained classification tools below to see how it works:

* [Sentiment Classifier](https://monkeylearn.com/sentiment-analysis-online/)
* [Intent and Email Classifier](https://monkeylearn.com/email-response-classifier-online/)
* [Survey Feedback Classifier](https://monkeylearn.com/nps-feedback-analysis-online/)

MonkeyLearn goes far beyond classification with text analysis tools that will give you the data results your business needs. [Request a demo](https://monkeylearn.com/contact/) to learn more about MonkeyLearn’s advanced text analysis tools.