# Theory: Introduction to classification



16 minutes 0 / 5 problems solved



Classification can be helpful, the primary technique here is to place a label for each sample. In terms of machine learning, the classification problem can be formulated to predict *finite and descriptive* sets. Basically, the classification is a supervised learning task of assigning a label to a sample.

## §1. Classification in real life

Classification is what we do every day in our heads. When we go to a supermarket to buy a watermelon, the main task is to determine whether a watermelon is good or not (*a binary classification*) based on tapping, weight, aroma, and sound.

Some tasks are more complex and ambiguous. In some cases, classes can overlap; when choosing a movie for the evening, one can choose a movie that is both comedy and fantasy. It is also worth mentioning that the data structure could be nested. This is a situation when classes are arranged in a hierarchy and are nested relative to each other. This is also a classification task, for example, nested word hierarchies — "class" in "classification".

A quick note on the difference between classification and clustering tasks. Classification implies that we already know the labels for some of the data. Clustering is a search for structure within the data without knowing the exact answer.

Mostly you will have to deal with **binary classifications**, **multiclass classifications**, and **multilabel classifications**. Binary classification implies a problem where a label is binary: yes or no, cat or dog, give a loan or request additional verification. Multiclass classification solves a problem with many classes, but each sample can only be in one class. For example, "apples, pears or plums". Multilabel classification assigns several labels at once, this task is very popular in text classification: drama, history, and comedy can be genres of one single book.

## §2. The formal explanation

In general, any task of supervised learning can be written like this:

Current topic:

Introduction to classification

Topic depends on:

× Typical ML pipeline

Topic is required for:

Classification performance metrics

Table of contents:

↑ Introduction to classification

§1. Classification in real life

§2. The formal explanation

§3. Basic ML algorithms for classification

§4. Non-linear algorithms

§5. Summary

Feedback & Comments

$f(x)$ is the pre-selected model, where $x$ is input data, $y$ output data. Our challenge as a machine learning engineer is to find *parameters* to the model $\theta$ that minimize the loss or error function. For example, the model should minimize the number of incorrectly classified pairs in the training set. Choosing the right model, the actual error, and optimizing the model parameters is the art of machine learning.

## §3. Basic ML algorithms for classification

We can divide all algorithms into linear *and* nonlinear. In the context of machine learning, linear models create a multidimensional dividing plane between classes. In contrast, nonlinear models create some complex surfaces.

How to choose the right training model so that it could produce results quickly and efficiently? First of all, it *depends on the data*. Features can be dependent or independent, linearly divided or divided by curves of a more complex order. So, it is so essential to understand the context of the problem before solving it. Knowing this will allow you to choose the best model.

First of all, let's give a quick overview of linear models.

*Naïve Bayes* is one of the earliest and oldest email spam search models that use the *Bayes rule*:

$$P(y|\vec{x}) = \frac{P(y)P(\vec{x}|y)}{P(\vec{x})}$$

This algorithm considers each feature as independent from other features but dependant on the output. Based on the formula, the algorithm predicts the probability of class.

*Logistic regression* is one of the main regression-based machine learning algorithms. Its main idea is that there is some *linear dividing surface*, on the opposite sides of which we have two different classes. The main task is to find the weight of each feature that would define this surface.

The *Support Vector Machines (*SVM) algorithm asks a more fundamental question: which of the lines is the most optimal in the figure*?*

The most optimal solution here would be the hyperplane that maximizes the margin between the two classes.

The advantages of linear models are that they are relatively fast and can be easily interpreted. But they do not work with the missing data and cannot restore nonlinear dependencies between features, which are often the case.

## §4. Non-linear algorithms

Nowadays, *artificial neural networks (multilayer perceptrons)* are one of the most popular tools for solving machine learning tasks. So, it is no surprise that they can be used for classification.

In the first approximation of a single neuron layer, this algorithm is closely related to logistic regression. The activation function adds nonlinearity to this system. This approach allows us to recover complex functions and find deeply hidden structures within the data. This is why they're so popular.

As for the disadvantages, it is worth noting that neural networks take a very long time to train in comparison with other models and require a larger amount of data.

Another popular algorithm is the *K-nearest neighbors* (k-NN). This algorithm's idea is simple, and as the saying goes — tell me who your friend is (or rather your neighbor), and I will tell you who you are. In other words, we look at the nearest neighbors of the sample we want to predict and say that it belongs to the same class.

The main disadvantage of this model is that it rarely considers the entire training set.

Another versatile algorithm is *a decision tree*. It divides the data into subsets according to features until an exact solution is given in each branch.

## §5. Summary

In this topic, we've discussed what a classification problem is and how it differs from a clustering problem. Classification is the task of determining the label for a sample, where labels come from a finite number of options. The classification task can be binary, multiclass, or multilabel. The basic algorithms can be divided into two types: linear and non-linear, which respectively create planes or complex surfaces.

If you feel a little overwhelmed with this amount of information, do not worry. In the following topics, we will discuss each of these methods separately.

Report a typo

Start practicing

Comments (1) Hints (0) Useful links (0) Show discussion