The classification problem in machine learning is concerned with how to divide data into categories. We have an input space X (all possible examples) and an output space Y (the categories of these examples). The goal is to create an algorithm that correctly matches examples to their categories, minimizing errors. There are 2 types of problems:

1. Supervised learning: The algorithm is trained on data where the correct answers are known (e.g. classifying objects).
2. Unsupervised learning: The algorithm looks for structure in data without provided answers (e.g. clustering customers by purchase profiles).

Classification is a supervised type.

In supervised learning, we work with two spaces: X (input data) and Y (output labels). In binary classification, objects are divided into two classes: -1 and +1. Our goal is to find a function $f:X\to Y$ that correctly maps objects and labels. To do this, we use a training set $(X1,Y1),...,(Xn,Yn)$, where each pair is selected independently from the general distribution. The classification algorithm, given this data, creates a function f that minimizes classification errors. This process helps the machine "learn" to classify new data. In general, the problem comes down to choosing the optimal function f that maps objects from space X to labels Y. The best classifier is known as Bayesian. It works like this: if the probability that an object belongs to class +1 is greater than or equal to 0.5, then assign the label +1, otherwise -1. This is the optimal choice that minimizes errors, but in reality it is difficult to calculate due to the unknown distribution of the data.

Instead, the binary classification problem boils down to the following: given a training sample $(X1,Y1),...,(Xn,Yn)$ drawn from an unknown distribution P, and a loss function, we aim to construct a function $f:X\to Y$ that minimizes the risk $R(f)$, approaching the risk of the Bayesian classifier.