1.1 Introduction

Machine learning is programming computers to optimize a performance criterion using example data or past experience.

Learning is used when:

- Human expertise does not exist (navigating on Mars),
- Humans are unable to explain their expertise (speech recognition)
- Solution changes in time (routing on a computer network)
- Solution needs to be adapted to particular cases (user biometrics)

To solve a problem on a computer, we need an algorithm. An algorithm is a sequence of instructions that should be carried out to transform the input to output. For example, one can devise an algorithm for sorting. The input is a set of numbers and the output is their ordered list. For the same task, there may be various algorithms and we may be interested in finding the most efficient one, requiring the least number of instructions or memory or both.

Application of machine learning methods to large databases is called *data mining*. In data mining, a large volume of data is processed to construct a simple model with valuable use, for example, having high predictive accuracy. Its application areas are abundant: In addition to retail, in finance banks analyze their past data to build models to use in credit applications, fraud detection, and the stock market. In manufacturing, learning models are used for optimization, control, and troubleshooting. In medicine, learning programs are used for medical diagnosis. In telecommunications, call patterns are analyzed for network optimization and maximizing the quality of service.

In science, large amounts of data in physics, astronomy, and biology can only be analyzed fast enough by computers. The World Wide Web is huge; it is constantly growing, and searching for relevant information cannot be done manually. But machine learning is not just a database problem; it is also a part of artificial intelligence. **To be intelligent**, a system that is in a changing environment should **have the ability to learn**. If the system can learn and adapt to such changes, the system designer provide solutions for all possible situations.

Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model define up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be ***predictive*** to make predictions in the future, or ***descriptive*** to gain knowledge from data, or both. Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample.

The role of computer science is twofold: **First, efficient algorithms** to solve the optimization problem, as well as to store and process the massive amount of data. **Second, efficient representation and algorithmic solution** for inference. In certain applications, the efficiency of the learning or inference algorithm, namely, its space and time complexity, may be as important as its predictive accuracy.

1.2 Machine Learning Applications

1.2.1 Learning Associations

In the case of retail-for example, a supermarket chain-one application of machine learning is **basket analysis**, which is finding associations between products bought by customers: If people who buy X typically also buy Y, and if there is a customer who buys X and does not buy Y, he or she is a potential Y customer. Once we find such customers, we can target them for cross-selling. In finding an ***association rule***, we are interested in learning a conditional probability of the form $P(Y|X)$ where Y is the product we would like to condition on X, which is the product or the set of products which we know that the customer has already purchased.

Let us say, going over our data, we calculate that $P(\text{chips}|\text{cola}) = 0.7$. Then, we can define the rule:

70 percent of customers who buy cola also buy chips.

We may want to make a distinction among customers and toward this, estimate $P(Y|X,D)$ where D is the set of customer attributes, for example, gender, age, marital status, and so on, assuming that we have access to this information. If this is a bookseller instead of a supermarket, products can be books or authors. In the case of a Web portal, items correspond to links to Web pages, and we can estimate the links a user is likely to click and use this information to download such pages in advance for faster access.

1.2.2 Classification

A credit is an amount of money loaned by a financial institution, for example, a bank, to be paid back with interest, generally in installments. It is important for the bank to be able to predict in advance the risk associated with a loan, which is the probability that the customer will default and not pay the whole amount back. This is both to make sure that the bank will make a profit and also to not inconvenience a customer with a loan over his or her financial capacity. In *credit scoring* (Hand 1998), the bank calculates the risk given the amount of credit and the information about the customer. The information about the customer includes data we have access to and is relevant in calculating his or her financial capacity—namely, income, savings, collaterals, profession, age, past financial history, and so forth. The bank has a record of past loans containing such customer data and whether the loan was paid back or not. From this data of particular applications, the aim is to infer a general rule coding the association between a customer's attributes and his risk. That is, the machine learning system fits a model to the past data to be able to calculate the risk for a new application and then decides to accept or refuse it accordingly.

This is an example of a *classification* problem where there are two classes: low-risk and high-risk customers. The information about a customer makes up the *input* to the classifier whose task is to assign the input to one of the two classes. After training with the past data, a classification rule learned may be of the form:

IF income> $\theta 1$ AND savings> $\theta 2$ THEN low-risk ELSE high-risk

for suitable values of $\theta 1$ and $\theta 2$ (see figure 1.1). This is an example of a *discriminant*; it is a function that separates the examples of different classes. Having a rule like this, the main application is *prediction*: Once we have a rule that fits the past data, if the future is similar to the past, then we can make correct predictions for novel instances. Given a new application with a certain income and savings, we can easily decide whether it is low risk or high-risk.

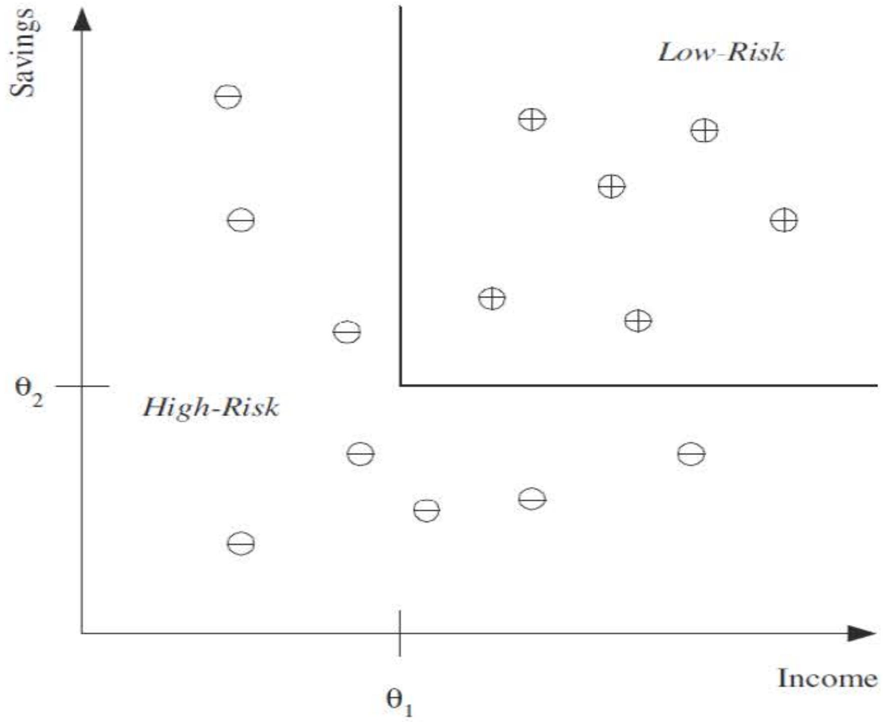

Figure 1.1 Example of a **training dataset** where each circle corresponds to one data instance with input values in the corresponding axes and its sign indicates the class. Two customer attributes used as an input, **income** and **savings**, and two classes are found: **low-risk ('+')** and **high-risk ('-').**

In some cases, instead of making a 0/1 (low-risk/high-risk) type decision, we may want to calculate a probability, namely, $P(Y|X)$, where X are the customer attributes and Y is 0 or 1 respectively for low-risk and high-risk. To see the classification as learning an association from X to Y. Then for a given $X = x$, if we have $P(Y = 1|X = x) = 0.8$, we say that the customer has an 80 percent probability of being high-risk, or equivalently a 20 percent probability of being low-risk. We then decide whether to accept or refuse the loan depending on the possible gain and loss.