Table of Content

What is Machine Learning ?

We are entering the era of big data. For example, there are about 1 trillion web pages1; one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day2; the genomes of 1000s of people, each of which has a length of 3.8×109 base pairs, have been sequenced by various labs; Walmart handles more than 1M transactions per hour and has databases containing more than 2.5 petabytes (2.5×1015) of information (Cukier 2010); and so on. This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, we define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!). This books adopts the view that the best way to solve such problems is to use the tools of probability theory. Probability theory can be applied to any problem involving uncertainty. In machine learning, uncertainty comes in many forms: what is the best prediction about the future given some past data? what is the best model to explain some data? what measurement should I perform next? etc. The probabilistic approach to machine learning is closely related to the field of statistics, but diers slightly in terms of its emphasis and terminology 3. We will describe a wide variety of probabilistic models, suitable for a wide variety of data and tasks. We will also describe a wide variety of algorithms for learning and using such models. The goal is not to develop a cook book of ad hoc techiques, but instead to present a unified view of the field through the lens of probabilistic modeling and inference. Although we will pay attention to computational eciency, details on how to scale these methods to truly massive datasets are better described in other books, such as (Rajaraman and Ullman 2011; Bekkerman et al. 2011).

Learning Algorithms

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning
4. Semi- Supervised Learning

Supervised Learning

Supervised learning is the machine learning task of inferring a function from supervised training data. The training data consist of a set of training examples. In supervised learning, each example is a pair consisting of an input object (typically a vector) and a desired output value (also called the supervisory signal). A supervised learning algorithm analyzes the training data and produces an inferred function, which is called a classifier (if the output is discrete, see classification) or a regression function (if the output is continuous, see regression). The inferred function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way (see inductive bias).

Unsupervised Learning

We now consider unsupervised learning, where we are just given output data, without any inputs. The goal is to discover “interesting structure” in the data; this is sometimes called knowledge discovery. Unlike supervised learning, we are not told what the desired output is for each input. Instead, we will formalize our task as one of density estimation, that is, we want to build models of the form $p(x_i|\theta)$. There are two di□erences from the supervised case. First, we have written $p(x_i|\theta)$ instead of $p(y_i|x_i,\theta)$; that is, supervised learning is conditional density estimation, whereas unsupervised learning is unconditional density estimation. Second, x_i is a vector of features, so we need to create multivariate probability models. By contrast, in supervised learning, y_i is usually just a single variable that we are trying to predict. This means that for most supervised learning problems, we can use univariate probability models (with input-dependent parameters), which significantly simplifies the problem. (We will discuss multi-output classification in Chapter 19, where we will see that it also involves multivariate probability models.) Unsupervised learning is arguably more typical of human and animal learning. It is also more widely applicable than supervised learning, since it does not require a human expert to manually label the data. Labeled data is not only expensive to acquire6, but it also contains relatively little information, certainly not enough to reliably estimate the parameters of complex models. Geo□ Hinton, who is a famous professor of ML at the University of Toronto, has said: When we’re learning to see, nobody’s telling us what the right answers are — we just look. Every so often, your mother says “that’s a dog”, but that’s very little information. You’d be lucky if you got a few bits of information — even one bit per second — that way. The brain’s visual system has 10^{14} neural connections. And you only live for 10^{9} seconds. So it’s no use learning one bit per second. You need more like 10^{5} bits per second. And there’s only one place you can get that much information: from the input itself. — Geo□rey Hinton, 1996 (quoted in (Gorder 2006)).

Semi Supervised Learning

Semi-supervised learning is a class of machine learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy over unsupervised learning (where no data is labeled), but without the time and costs needed for supervised learning (where all data is labeled).[1] The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning.

Classification

In this section, we discuss classification. Here the goal is to learn a mapping from inputs x to outputs y, where $y \in \{1,...,C\}$, with C being the number of classes. If C =2, this iscalled binary classification (in which case we often assume $y \in \{0,1\}$); if C>2, this is calledmulticlass classification . If the class labels are not mutually exclusive (e.g., somebody may be classified as tall and strong), we call it multi-label classification, but this is best viewed as predicting multiple related binary class labels (a so-called multiple output model). When we use the term "classification", we will mean multiclass classification with a single output, unless we state otherwise. One way to formalize the problem is as function approximation. We assume $y = f(x)$ for some unknown function f, and the goal of learning is to estimate the function f given a labeled training set, and then to make predictions using $\hat{y} = \hat{f}(x)$. (We use the hat symbol to denote an estimate.) Our main goal is to make predictions on novel inputs, meaning ones that we have not seen before (this is called generalization), since predicting the response on the training set is easy (we can just look up the answer).

Regression

Regression is just like classification except the response variable is continuous. Figure 1.7 shows a simple example: we have a single real-valued input $x_i \in R$, and a single real-valued response $y_i \in R$. We consider fitting two models to the data: a straight line and a quadratic function. (We explain how to fit such models below.) Various extensions of this basic problem can arise, such as having high-dimensional inputs, outliers, non-smooth responses, etc. We will discuss ways to handle such problems later in the book.

Dimension Reduction

In statistics, machine learning, and information theory, **dimensionality reduction** or **dimension reduction** is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction.