CHAPTER 1

Introduction: Classification, Learning, Features, and Applications

1.1 SCOPE

In this book we are concerned mainly with pattern classification—classifying an object into one of several categories on the basis of several observations or measurements of the object. The simplest case is classification of an object into one of two categories, but a more general case allows for any finite number of categories.

A second closely related task is estimation of a real number that is typically related to some property of the object. As in classification, several observations or measurements of the object are available, and our estimate is based on these observations.

Most of our discussion concerns issues arising about the first task, classification. But we occasionally say something about the second task, estimation. In either case, we are interested in *rules* for classifying objects or estimating values, given certain observations or measurements. More specifically, we are interested in methods for *learning* rules for classification or estimation.

We discuss some concrete examples further below. For now, think about learning to recognize handwritten characters or faces or other objects from visual data. Or, think about the problem of recognizing spoken words. While humans are extremely good at these types of classification problems in many natural settings, it is quite difficult to design automated algorithms for these tasks with performance and robustness anywhere near those of humans.

Even after more than a half century of effort in fields such as electrical engineering, mathematics, computer science, statistics, philosophy, and cognitive science, humans can still far outperform the best machine learning algorithms that have ever been developed. That said, enormous progress has been made in learning theory, algorithms, and applications. Results in this area are deep and practical and are

An Elementary Introduction to Statistical Learning Theory, First Edition.
Gilbert Harman and Sanjeev Kulkarni.

relevant to a range of disciplines such as those we have mentioned above. Many of the basic ideas are accessible to a broad audience. However, most treatments of this material are at an advanced level, requiring a rather technical background and expertise.

Our aim in this book is to provide an accessible introduction to this field, either as a first step for those wishing to pursue the subject in more depth, or for those desiring a broad understanding of the basic ideas. For most of the book, we focus on the problem of two-class pattern classification. This problem arises in many useful applications and is sufficiently rich to explain many of the key ideas in the field, yet removes some unnecessary complications. Although many important aspects of learning are not covered by this model, we provide many good references for more depth, generalizations, and other models. We hope this book will serve as a valuable entry point.

1.2 WHY MACHINE LEARNING?

Algorithms for recognizing patterns would be useful in a wide range of problems. This ability is one aspect of "artificial intelligence." But one might reasonably ask why we need to design automated methods for *learning* good rules for classification, as opposed to just figuring out what is a good rule for a given application and implementing it.

The main reason is that in many applications, the only way we can find a good rule is to use data to learn one. For example, it is very hard to describe exactly what constitutes a face in an image, and therefore it is hard to come up with a classification rule to decide whether or not a given image contains a face. But, given a good learning algorithm, we might be able to present the algorithm with many examples of images of a face and many examples of images without a face, and then let the algorithm come up with a good rule for recognizing whether or not a face is present. There are other benefits of having a learning algorithms as well, such as robustness to errors in assumptions or modelling, reduced need for explicit programming, and adaptation to changing conditions.

In general, for a classification problem, we want to decide to which of several categories the object belongs on the basis of some measurements of the object. To *learn* a good rule, we use data that consist of many examples of objects with their correct classification. The following questions immediately arise:

1. What do we mean by an "object" and "measurements" of the object?
2. In the classification problem, what are the categories to which we assign objects?
3. In the estimation problem, what are the values we attempt to estimate?
4. How do we measure the quality of a classification or estimation rule, and what is the best rule we could hope for?

5. What information is available to use for learning?
6. How do we go about learning a good classification or an estimation rule?

We describe the answers to the first three questions in this chapter. To answer the remaining questions, some background material on probability is provided in Chapters 2 and 3. With this background, the answer to the fourth question is discussed in Chapters 4 and 5. The answer to the fifth question is discussed in Chapter 6. The rest of the book is devoted to various aspects of and approaches to the last question.

1.3 SOME APPLICATIONS

Before discussing further details, it may be helpful to have some concrete examples in mind. There are a wide range of applications for learning, classification, and estimation. Here we mention just a few.

1.3.1 Image Recognition

There are many applications in which the object to be classified is a digital image. The "measurements" in this case might describe the outputs of each of the pixels in the image. In the case of a black and white image, the intensity of each pixel serves as one measurement. If the image has $N \times N$ pixels, then the total number of pixels (and hence measurements) is N^2. In the case of a color image, each pixel can be considered as providing three measurements, corresponding to the intensities of each of three color components, say RGB values. Hence, for an $N \times N$ color image, there are $3N^2$ measurements.

Depending on the application, there are many classification tasks based on using these measurements. Face detection or recognition is a common and useful application. In this case, the "categories" might be face versus no face present, or there might be a separate category for each person in a database of individuals.

A different application is character recognition. In this case, the writing can be segmented into smaller images that each contain a single character, and the *categories* might consist of the 26 letters of the alphabet (52 letters, if upper and lower case letters are to be distinguished), the 10 digits, and possibly some special characters (period, question mark, comma, colon, etc.).

In yet another application, the images might be of industrial parts and the categorization task is to decide whether the current part is defective or not.

1.3.2 Speech Recognition

In speech recognition, we are interested in recognizing the words uttered by a speaker. The *measurements* in this application might be a set of numbers that represent the speech signal. First, the signal is typically segmented into portions that contain distinct words or phonemes. In each segment, the speech signal can

be represented in a variety of ways. For example, the signal can be represented by the intensities or energy in different time-frequency bands. Although the details of the signal representation are outside the scope of this book, the signal can be represented ultimately by a set of real values.

In the simplest case, the *categories* might be as simple as deciding whether the utterance is "yes" versus "no." A slightly more complicated task might be to decide which of the 10 digits is uttered. Or there might be a category for each word from a large dictionary of acceptable words and the task might be to decide which, if any, of this large number of words has been uttered.

1.3.3 Medical Diagnosis

In medical diagnosis, we are interested in whether or not there is a disease present (and which disease). There is a separate *category* for each of the diseases under consideration and one category for the case where no disease is present.

The *measurements* in this application are typically the results of certain medical tests (e.g., blood pressure, temperature, and various blood tests) or medical diagnostics (such as medical images), presence/absence/intensity of various symptoms, and some basic physical information about the patient (age, sex, weight, etc.).

On the basis of the results of the measurements, we would like to decide which disease (if any) is present.

1.3.4 Statistical Arbitrage

In finance, statistical arbitrage refers to automated trading strategies that are typically of a very short term and involve a large number of securities. In such strategies, one tries to design a trading algorithm for the set of securities on the basis of quantities such as historical correlations among the large set of securities, price movements over recent time horizons, and general economic/financial variables. These can be thought of as the "measurements" and the prediction can be cast as a classification or estimation problem. In the case of classification, the categories might be "buy," "sell," or "do nothing" for each security. In the estimation case, one might try to predict the expected return of each security over some future time horizon. In this case, one typically needs to use the estimates of the expected return to make a trading decision (buy, sell, etc.).

1.4 MEASUREMENTS, FEATURES, AND FEATURE VECTORS

As we discussed in Sections 1.1 and 1.3, in classifying an object, we use observations about the object in order to make our decision. For example, when humans wish to classify an object, they might look at the object, pick it up, feel it, listen to it, etc. Or they might use some instruments to measure other properties of the object such as size, weight, and temperature.

Similarly, when designing a machine to automatically classify (or learn to classify) objects, we assume that the machine has access to measurements of various properties of the object. These measurements come from sensors that capture some physical variables of interest, or *features*, of the object.

For simplicity, in this book we model each measurement (or feature) as being captured by a single real number. Although in some applications, certain features may not be very naturally represented by a number, this assumption allows discussion of the most common learning techniques that are useful in the most common applications.

We assume that all the relevant and available aspects of the objects can be captured in a finite number of measurements/features. These finite number of features can be put together to form a *feature vector*. Suppose there are d features with the value of the features given by $x_1, x_2, \ldots, x_d$. The feature vector is $\overline{x} = (x_1, x_2, \ldots, x_d)$. This feature vector can be thought of as a point or a vector in d-dimensional space $\mathbf{R}^d$, which we call the *feature space*. Each component of the feature vector, indicating the value of the corresponding feature, is the value along a particular dimension of the feature space.

In the case of image recognition with an $N \times N$ image, the number of features is N^2 for a black and white image and $3N^2$ for a color image.

In speech recognition, the number of features is equal to the number of real values used to represent the speech segment to be classified.

1.5 THE NEED FOR PROBABILITY

In most applications, the category of the object is not uniquely and definitively determined by the value of the feature vector. There are some fundamental reasons for this. First, although it would be nice if the measured features capture all the properties of the object important for classification, this is usually not the case. The measured features might fail to capture some important details. This should be clear in the examples given above.

Second, depending on the application and the specific measurements, the feature values may be noisy. That is, there may be some inherent uncertainty or randomness in the observed values of the features so that even the same object might give rise to different values on different occasions.

For these reasons, it is helpful to use tools from probability to formulate the problem precisely and guide the solution. In Chapters 2 and 3, we review some of the basic tools from probability that we need for the rest of the book.

1.6 SUPERVISED LEARNING

After providing the necessary background from probability, in Chapter 4, we formulate the pattern recognition problem. In the ideal (and unusual) case, where the underlying probabilistic structure is known, the solution to the classification

problem is well known and is a basic result from statistics. This is discussed in Chapter 5.

However, in the much more typical case in applications, the underlying probability distributions are not known. In this case, we try to overcome this lack of knowledge by resorting to labeled examples as we discuss in Chapter 6. The learning problem, as formulated in Chapter 6, is just one type of machine learning problem known by various terms such as *learning from examples*, *supervised learning*, *statistical pattern classification*, *statistical pattern recognition*, and *statistical learning*.

The term "supervised" learning arises from the fact that examples we assume that we have access to are properly labeled by a "supervisor" or "teacher." This contrasts with "unsupervised learning," in which many examples of objects are available, but the class to which the objects belong are unknown. There are also other formulations of machine learning problems such as semi-supervised learning and reinforcement learning, as well as many other related problems in statistics, computer science, and other fields. But in this book, we focus exclusively on the case of supervised learning.

1.7 SUMMARY

In this chapter, we described the general problems of classification and estimation and discussed several concrete and important applications. We then introduced the terminology of features, feature vectors, and feature space. The need for introducing probability and learning was described.

We have mentioned both classification and estimation. We focus mainly on classification in this book, with some discussion of extensions to estimation.

In the next two chapters, we review some principles of probability that are important for aspects discussed in the rest of the book. After this, we formalize the classification (or pattern recognition) problem and discuss general issues in learning from data, before moving on to a discussion of specific learning methods and results.

1.8 APPENDIX: INDUCTION

The appendices at the end of each chapter briefly discuss certain side issues, perhaps of a philosophical nature.

In this book, we are concerned primarily with inductive learning rather than deductive learning. Deductive learning consists in deriving a new conclusion from premises whose truth guarantees the truth of the conclusion. For example, you might learn that the area of a parallelogram is equal to its base times its height by deducing this from what you already know about rectangles and about how the area of a parallelogram is related to the area of a rectangle with the same base and

height. You might then learn that the area of a triangle is equal to its base times *half* its height, by deducing this from the fact that any triangle is exactly half of a certain parallelogram.

Inductive learning consists in reaching a conclusion from evidence that does not guarantee the truth of the conclusion. For example, you might infer from the fact that mail has almost always been delivered before noon on Saturdays up until now to the conclusion that mail will be delivered before noon next Saturday. This is an inductive inference, because the data do not guarantee the truth of the conclusion. Sometimes, the conclusion of an inductive inference is false even though the "premises" of the inference are all true.

The philosophical "problem of induction" asks how one can be *justified* in believing inductive conclusions from true premises. Certainly, it is not possible to prove deductively that any such inductive conclusion is true if its premises are, since typical inductive inferences do not provide such a guarantee. Even if you are justified inductively in thinking that your mail will be delivered before noon next Saturday, it is compatible with your evidence that your mail is not delivered before noon next Saturday. Induction is not a special case of deduction.

It might be suggested that induction has almost always led to true conclusions in the past, so it is reasonable to conclude that it will almost always lead to true conclusions in the future. The objection to this suggestion is that this is circular reasoning: we are assuming that induction is justified in order to argue that induction is justified!

On the other hand, is it possible to offer a noncircular justification of deduction? Wouldn't any such justification take the form of a deductive argument and so also be circular?

It will emerge that statistical learning theory provides partial deductive mathematical justifications for certain inductive methods, given certain assumptions.

1.9 QUESTIONS

1. What is a feature space? What do the dimensions of such a space represent? What is a vector? What is a feature vector?

2. If we want to use the values of F different features, in order to classify objects, where each feature can have any of G different values, what is the dimension of the feature space?

3. For a 12×12 grayscale image (256 grayscale levels), how many *dimensions* are there for the feature vector? How many different possible *feature vectors* are there?

4. Is classification a special case of estimation? What differences are there between typical cases of classification and typical cases of estimation?

5. About the problem of induction

(a) What is the problem of induction?
(b) How does the reliability of induction compare with the reliability of deduction?
(c) How might statistical learning theory say something about the reliability of induction?

1.10 REFERENCES

Statistical pattern recognition as a distinct field has been an active area of research for about half a century, though its foundations are based on probability and statistics which go back much further than that. Statistical pattern recognition (or statistical learning) is part of the broad area of machine learning and spans many disciplines such as mathematics, probability, statistics, electrical engineering, computer science, cognitive science, econometrics, and philosophy. There are a number of conference, journals, and books devoted to machine learning, and among these much of the material is devoted to statistical learning.

Mitchell (1997) is an introduction to issues about machine learning generally. Vickers (2010) is an up-to-date discussion of the problem of induction. The other references below are just some of the many classic and recent references that discuss statistical pattern recognition and related areas at various levels.

Bishop C. Pattern recognition and machine learning. New York: Springer; 2006.

Bongard M. Pattern recognition. Washington (DC): Spartan Books; 1970.

Devijver PR, Kittler J. Pattern recognition: a statistical approach. Englewood Cliffs (NJ): Prentice-Hall; 1982.

Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. New York: Springer Verlag; 1996.

Duda RO, Hart PE. Pattern classification and scene analysis. New York: Wiley; 1973.

Duda RO, Hart PE, Stork DG. Pattern classification. 2nd ed. New York: Wiley; 2001.

Fukunaga K. Introduction to statistical pattern recognition. 2nd ed. San Diego (CA): Academic Press; 1990.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.

Ho YC, Agrawala A. On pattern classification algorithms: introduction and survey. Proc IEEE 1968;56:2101–2114.

Kulkarni SR, Lugosi G, Venkatesh S. Learning pattern classification - A survey. IEEE Trans Inf Theory 1998; 44(6): 2178–2206.

Mitchell T. Machine learning. Boston (MA): McGraw-Hill; 1997.

Nilsson NJ. Learning machines. New York: McGraw-Hill; 1965.

Schalkoff RJ. Pattern recognition: statistical, structural, and neural approaches. New York: Wiley; 1992.

Theodoridis S, Koutroumbas K. Pattern recognition. 4th ed. Amsterdam: Academic Press; 2008.

Theodoridis S, Pikrakis A, Koutroumbas K, Cavouras D. Introduction to pattern recognition: a matlab approach. Amsterdam: Academic Press; 2010.

Vapnik VN. The nature of statistical learning theory. New York: Springer; 1999.

Vickers J. The Problem of Induction, in The Stanford Encylopedia of Philosophy; 2010, http://plato.stanford.edu/entries/induction-problem/.

Watanabe MS. Knowing and guessing. New York: Wiley; 1969.