1

Introduction

Cluster analysis and classification are two important tasks which occur daily in everyday life. As humans, our brain naturally clusters and classifies animals, objects or even ideas thousands of times a day, without fatigue. The emergence of science has led to many data sets with clustering structure that cannot be easily detected by the human brain, and so require automated algorithms. Also, with the advent of the "Data Age," clustering and classification tasks are often repeated large numbers of times, and so need to be automated even if the human brain could carry them out.

This has led to a range of clustering and classification algorithms over the past century. Initially these were mostly heuristic, and developed without much reference to the statistical theory that was emerging in parallel. In the 1960s, it was realized that cluster analysis could be put on a principled statistical basis by framing the clustering task as one of inference for a finite mixture model. This has allowed cluster analysis to benefit from the inferential framework of statistics, and provide principled and reproducible answers to questions such as: how many clusters are there? what is the best clustering algorithm? how should we deal with outliers?

In this book, we describe and review the model-based approach to cluster analysis which has emerged in the past half-century, and is now an active research field. We describe the basic ideas, and aim to show the advantages of thinking in this way, as well as to review recent developments, particularly for newer types of data such as high-dimensional data, network data, textual data and image data.

1.1 Cluster Analysis

The goal of cluster analysis is to find meaningful groups in data. Typically, in the data these groups will be internally cohesive and separated from one another. The purpose is to find groups whose members have something in common that they do not share with members of other groups.

1.1.1 From Grouping to Clustering

The grouping of objects according to things they have in common goes back at least to the beginning of language. A noun (such as "hammer") refers to any one of a set of different individual objects that have characteristics in common. As

Greene (1909) remarked, "naming is classifying." Plato was among the first to formalize this with his Theory of Forms, defining a Form as an abstract unchanging object or idea, of which there may be many instances in practice. For example, in Plato's *Cratylus* dialogue, he has Socrates giving the example of a blacksmith's tool, such as a hammer. There are many hammers in the world, but just one Platonic Form of "hammerness" which is the essence of all of them.

Aristotle, in his *History of Animals*, classified animals into groups based on their characteristics. Unlike Plato, he drew heavily on empirical observations. His student Theophrastus did something similar for plants in his *Enquiry Into Plants*.

An early and highly influential example of the systematic grouping of objects based on measured empirical characteristics is the system of biological classification or taxonomy of Linnaeus (1735), applied to plants by Linnaeus (1753) and to animals by Linnaeus (1758). For example, he divided plants into 24 classes, including flowers with one stamen (Monandria), flowers with two stamens (Diandria) and flowerless plants (Cryptogamia). Linnaeus' methods were based on data but were subjective. Adanson (1757, 1763) developed less subjective methods using multiple characteristics of organisms.

Cluster analysis is something more: the search for groups in quantitative data using systematic numerical methods. Perhaps the earliest methods that satisfy this description were developed in anthropology, and mainly consisted of defining quantitative measures of difference and similarity between objects (Czekanowski, 1909, 1911; Driver and Kroeber, 1932). Most of the early clustering methods were based on measures of similarity between objects, and Czekanowski (1909) seems to have been the first to define such a measure for clustering purposes.

Then development shifted to psychology, where Zubin (1938) proposed a method for rearranging a correlation matrix to yield clusters. Stephenson (1936) proposed the use of factor analysis to identify clusters of people, while, in what seems to have been the first book on cluster analysis, Tryon (1939) proposed a method for clustering variables similar to what is now called multiple group factor analysis. Cattell (1944) also introduced several algorithmic and graphical clustering methods.

In the 1950s, development shifted again to biological taxonomy, the original problem addressed by the ancient Greeks and the eighteenth-century scientists interested in classification. It was in this context that the single link hierarchical agglomerative clustering method (Sneath, 1957), the average link method and the complete link method (Sokal and Michener, 1958) were proposed. These are sometimes thought of as marking the beginning of cluster analysis, but in fact they came after a half-century of previous, though relatively slow development. They did mark the takeoff of the area, though. The development of computational power and the publication of the important book of Sokal and Sneath (1963) led to a rapid expansion of the use and methodology of cluster analysis, which has not stopped in the past 60 years.

Many of the ensuing developments in the 1970s and 1980s were driven by applications in market research as well as biological taxonomy. From the 1990s

there was a further explosion of interest fueled by new types of data and questions, often involving much larger data sets than before. These include finding groups of genes or people using genetic microarray data, finding groups and patterns in retail barcode data, finding groups of users and websites from Internet use data, and automatic document clustering for technical documents and websites.

Another major area of application has been image analysis. This includes medical image segmentation, for example for finding tumors in digital medical images such as X-rays, CAT scans, MRI scans and PET scans. In these applications, a cluster is typically a set of pixels in the image. Another application is image compression, using methods such as color image quantization, where a cluster would correspond to a set of color levels. For a history of cluster analysis to 1988, see Blashfeld and Aldenderfer (1988).

1.1.2 Model-based Clustering

Most of the earlier clustering methods were algorithmic and heuristic. The majority were based on a matrix of measures of similarity between objects, which were in turn derived from the objects' measured characteristics. The purpose was to divide or partition the data into groups such that objects in the same group were similar, and were dissimilar from objects in other groups. A range of automatic algorithms for doing this was proposed, starting in the 1930s.

These developments took place largely in isolation from mainstream statistics, much of which was based on a probability distribution for the data. At the same time, they left several practical questions unresolved, such as which of the many available clustering methods to use? How many clusters should we use? How should we treat objects that do not fall into any group, or outliers? How sure are we of a clustering partition, and how should we assess uncertainty about it?

The mainstream statistical approach of specifying a probability model for the full data set has the potential to answer these questions. The main statistical model for clustering is a finite mixture model, in which each group is modeled by its own probability distribution.

The first successful method of this kind was developed in sociology in the early 1950s for multivariate discrete data, where multiple characteristics are measured for each object, and each characteristic can take one of several values. Data of this kind are common in the social sciences, and are typical, for example, of surveys. The model proposed was called the latent class model, and it specified that within each group the characteristics were statistically independent (Lazarsfeld, 1950a,c). We discuss this model and its development in Chapter 6.

The dominant model for clustering continuous-valued data is the mixture of multivariate normal distributions. This seems to have been first mentioned by Wolfe (1963) in his Master's thesis at Berkeley. John Wolfe subsequently developed the first real software for estimating this model, called NORMIX, and also developed related theory (Wolfe, 1965, 1967, 1970), so he has a real claim to be called the inventor of model-based clustering for continuous data. Wolfe proposed estimating the model by maximum likelihood using the EM algorithm,

which is striking since he did so ten years before the article of Dempster et al. (1977) that popularized the EM algorithm. This remains the most used estimation approach in model-based clustering. We outline the early history of model-based clustering in Section 2.9, after we have introduced the main ideas.

Basing cluster analysis on a probability model has several advantages. In essence, this brings cluster analysis within the range of standard statistical methodology and makes it possible to carry out inference in a principled way. It turns out that many of the previous heuristic methods correspond approximately to particular clustering models, and so model-based clustering can provide a way of choosing between clustering methods, and encompasses many of them in its framework. In our experience, when a clustering method does not correspond to any probability model, it tends not to work very well. Conversely, understanding what probability model a clustering method corresponds to can give one an idea of when and why it will work well or badly.

It also provides a principled way to choose the number of clusters. In fact, the choice of clustering model and of number of clusters can be reduced to a single model selection problem. It turns out that there is a trade-off between these choices. Often, if a simpler clustering model is chosen, more clusters are needed to represent the data adequately.

Basing cluster analysis on a probability model also leads to a way of assessing uncertainty about the clustering. In addition, it provides a systematic way of dealing with outliers by expanding the model to account for them.

1.2 Classification

The problem of classification (also called discriminant analysis) involves classifying objects into classes when there is already information about the nature of the classes. This information often comes from a data set of objects that have already been classified by experts or by other means. Classification aims to determine which class new objects belong to, and develops automatic algorithms for doing so. Typically this involves assigning new observations to the class whose objects they most closely resemble in some sense.

Classification is said to be a "supervised" problem in the sense that it requires the supervision of experts to provide some examples of the classes. Clustering, in contrast, aims to divide a set of objects into groups without any examples of the "true" classes, and so is said to be an "unsupervised" problem.

1.2.1 From Taxonomy to Machine Learning

The history of classification is closely related to that of clustering. Indeed, the practical interest of taxonomies of animals or plants is to use them to recognize samples on the field. For centuries, the task of classification was carried out by humans, such as biologists, botanists or doctors, who learned to assign new observations to specific species or diseases. Until the twentieth century, this was done without automatic algorithms.

The first statistical method for classification is due to Ronald Fisher in his famous work on discriminant analysis (Fisher, 1936). Fisher asked what linear combination of several features best discriminates between two or more populations. He applied his methodology, known nowadays as Fisher's discriminant analysis or linear discriminant analysis, to a data set on irises that he had obtained from the botanist Edgar Anderson (Anderson, 1935).

In a following article (Fisher, 1938), he established the links between his discriminant analysis method and several existing methods, in particular analysis of variance (ANOVA), Hotelling's T-squared distribution (Hotelling, 1931) and the Mahalanobis generalized distance (Mahalanobis, 1930). In his 1936 paper, Fisher also acknowledged the use of a similar approach, without formalization, in craniometry for quantifying sex differences in measurements of the mandible.

Discriminant analysis rapidly expanded to other application fields, including medical diagnosis, fault detection, fraud detection, handwriting recognition, spam detection and computer vision. Fisher's linear discriminant analysis provided good solutions for many applications, but other applications required the development of specific methods.

Among the key methods for classification, logistic regression (Cox, 1958) extended the usual linear regression model to the case of a categorical dependent variable and thus made it possible to do binary classification. Logistic regression had a great success in medicine, marketing, political science and economics. It remains a routine method in many companies, for instance for mortgage default prediction within banks or for click-through rate prediction in marketing companies.

Another key early classification method was the perceptron (Rosenblatt, 1958). Originally designed as a machine for image recognition, the perceptron algorithm is supposed to mimic the behavior of neurons for making a decision. Although the first attempts were promising, the perceptron appeared not to be able to recognize many classes without adding several layers. The perceptron is recognized as one of the first artificial neural networks which recently revolutionized the classification field, partly because of the massive increase in computing capabilities. In particular, convolutional neural networks (LeCun et al., 1998) use a variation of multilayer perceptrons and display impressive results in specific cases.

Before the emergence of convolutional neural networks and deep learning, support vector machines also pushed forward the performances of classification at the end of the 1990s. The original support vector machine algorithm or SVM (Cortes and Vapnik, 1995), was invented in 1963 and it was not to see its first implementation until 1992, thanks to the "kernel trick" (Boser et al., 1992). SVM is a family of classifiers, defined by the choice of a kernel, which transform the original data in a high-dimensional space, through a nonlinear projection, where they are linearly separable with a hyperplane. One of the reasons for the popularity of SVMs was their ability to handle data of various types thanks to the notion of kernel.

As we will see in this book, statistical methods were able to follow the different revolutions in the performance of supervised classification. In addition, some of

the older methods remain reference methods because they perform well with low complexity.

1.2.2 Model-based Discriminant Analysis

Fisher discriminant analysis (FDA, Fisher (1936)) was the first classification method. Although Fisher did not describe his methodology within a statistical modeling framework, it is possible to recast FDA as a model-based method. Assuming normal distributions for the classes with a common covariance matrix yields a classification rule which is based on Fisher's discriminant function. This classification method, named linear discriminant analysis (LDA), also provides a way to calculate the optimal threshold to discriminate between the classes within Fisher's discriminant subspace (Fukunaga, 1999).

An early work considering class-conditional distributions in the case of discriminant analysis is due to Welch (1939). He gave the first version of a classification rule in the case of two classes with normal distributions, using either Bayes' theorem (if the prior probabilities of the classes are known) or the Neyman–Pearson lemma (if these prior probabilities have to be estimated). Wald (1939, 1949) developed the theory of decision functions which offers a sound statistical framework for further work in classification. Wald (1944) considered the problem of assigning an individual into one of two groups under normal distributions with a common covariance matrix, the solution of which involves Fisher's discriminant function. Von Mises (1945) addressed the problem of minimizing the classification error in the case of several classes and proposed a general solution to it. Rao (1948, 1952, 1954) extended this to consider the estimation of a classification rule from samples. See Das Gupta (1973) and McLachlan (1992) for reviews of the earlier development of this area.

Once the theory of statistical classification had been well established, researchers had to face new characteristics of the data, such as high-dimensional data, low sample sizes, partially supervised data and non-normality. Regarding high-dimensional data, McLachlan (1976) realized the importance of variable selection to avoid the curse of dimensionality in discriminant analysis. Banfield and Raftery (1993) and Bensmail and Celeux (1996) proposed alternative approaches using constrained Gaussian models. About partially supervised classification, McLachlan and Ganesalingam (1982) considered the use of unlabeled data to update a classification rule in order to reduce the classification error. Regarding non-normality, Celeux and Mkhadri (1992) proposed a regularized discriminant analysis technique for high-dimensional discrete data, while Hastie and Tibshirani (1996) considered the classification of non-normal data using mixtures of Gaussians. In Chapter 4 specific methods for classification with categorical data are presented. These topics will be developed in this book, in Chapters 4, 5 and 8.

1.3 Examples

We now briefly describe some examples of cluster analysis and discriminant analysis.

Example 1: Old Faithful geyser data

The Old Faithful geyser in Yellowstone National Park, Wyoming erupts every 35–120 minutes for about one to five minutes. It is useful for rangers to be able to predict the time to the next eruption. The time to the next eruption and its duration are related, in that the longer an eruption lasts, the longer the time until the next one. The data we will consider in this book consist of observations on 272 eruptions Azzalini and Bowman (1990). Data on two variables were measured: the time from one eruption to the next one, and the duration of the eruption. These data are often used to illustrate clustering methods.

Example 2: Diagnosing type of diabetes

Figure 1.1 shows measurements made on 145 subjects with the goal of diagnosing diabetes and, for diabetic patients, the type of diabetes present. The data consist of the area under a plasma glucose curve (glucose area), the area under a plasma insulin curve (insulin area) and the steady-state plasma glucose response (SSPG) for 145 subjects. The subjects were subsequently clinically classified into three groups: chemical diabetes (Type 1), overt diabetes (Type 2), and normal (non-diabetic). The goal of our analysis is either to develop a method for grouping patients into clusters corresponding to diagnostic categories, or to learn a classification rule able to predict the status of a new patient. These data were described and analyzed by Reaven and Miller (1979).

Example 3: Breast cancer diagnosis

In order to diagnose breast cancer, a fine needle aspirate of a breast mass was collected, and a digitized image of it was produced. The cells present in the image were identified, and for each cell nucleus the following characteristics were measured from the digital image: (a) radius (mean of distances from center to points on the perimeter); (b) texture (standard deviation of gray-scale values); (c) perimeter; (d) area; (e) smoothness (local variation in radius lengths); (f) compactness (perimeter^2 / area $-$ 1); (g) concavity (severity of concave portions of the contour); (h) concave points (number of concave portions of the contour); (i) symmetry; (j) fractal dimension. The mean, standard deviation and extreme values of the 10 characteristics across cell nuclei were then calculated, yielding 30 features of the image (Street et al., 1993; Mangasarian et al., 1995).

A pairs plot of three of the 30 variables is shown in Figure 1.2. These were selected as representing a substantial amount of the variability in the data, and in fact are the variables with the highest loadings on each of the first three principal components of the data, based on the correlation matrix.

This looks like a more challenging clustering/classification problem than the first two examples (Old Faithful data and diabetes diagnosis data), where clustering

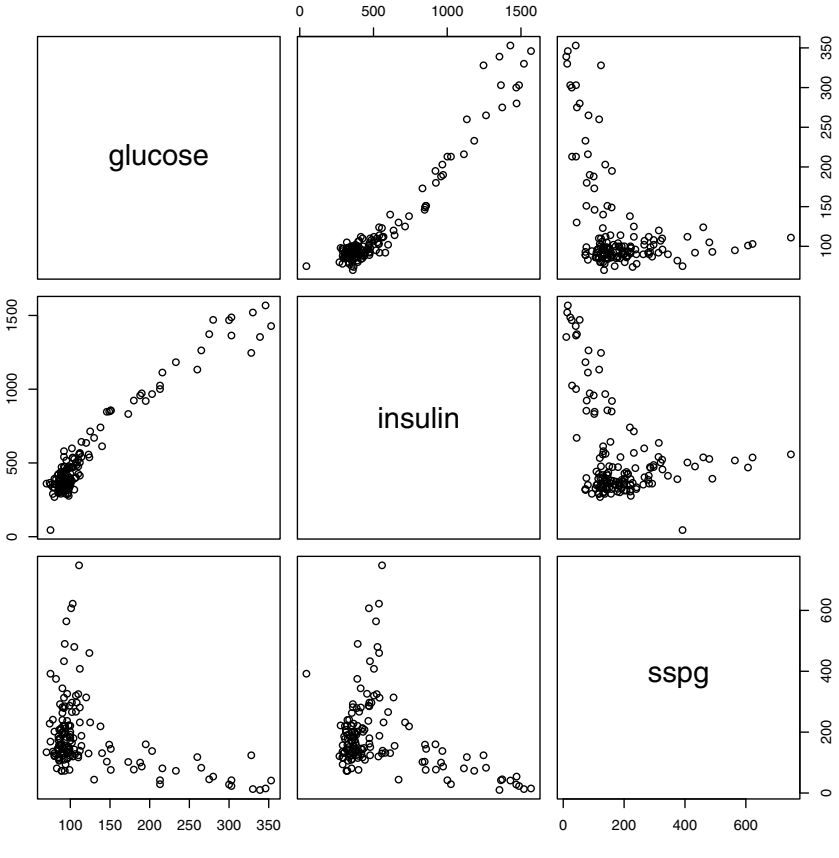

Figure 1.1 Diabetes data pairs plot: three measurements on 145 patients. *Source:* Reaven and Miller (1979).

was apparent visually from the plots of the data. Here it is hard to discern clustering in Figure 1.2. However, we will see in Chapter 2 that in this higher dimensional setting, model-based clustering can detect clusters that agree well with clinical criteria.

Example 4: Wine varieties

Classifying food and drink on the basis of characteristics is an important use of cluster analysis. We will illustrate this with a data set giving up to 27 physical and chemical measurements on 178 wine samples (Forina et al., 1986). The goal of an analysis like this is to partition the samples into types of wine, and potentially also by year of production. In this case we know the right answer: there are three types of wine, and the year in which each sample was produced is also known.

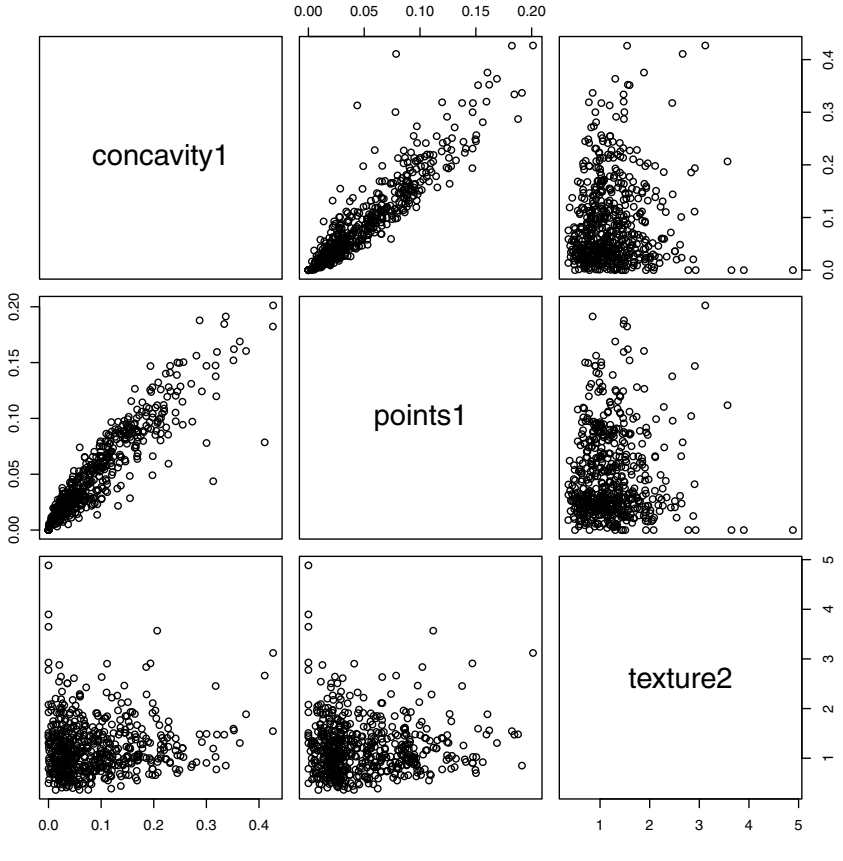

Figure 1.2 Pairs plot of three of the 30 measurements on breast cancer diagnosis images.

Thus we can assess how well various clustering methods perform. We will see in Chapter 2 that model-based clustering is successful at this task.

Example 5: Craniometric analysis

Here the task is to classify skulls according to the populations from which they came, using cranial measurements. We will analyze data with 57 cranial measurements on 2,524 skulls. As we will see in Chapter 2, a major issue is determining the number of clusters, or populations.

Example 6: Identifying minefields

We consider the problem of detecting surface-laid minefields on the basis of an image from a reconnaissance aircraft. After processing, such an image is reduced to a list of objects, some of which may be mines and some of which may be "clutter"

Figure 1.3 Minefield data. Left: observed data. Right: true classification into mines and clutter.

or noise, such as other metal objects or rocks. The objects are small and can be represented by points without losing much information. The analyst's task is to determine whether or not minefields are present, and where they are. A typical data set is shown in Figure 1.3.[1] The true classification of the data between mines and clutter is shown in the right panel of Figure 1.3. These data are available as the `chevron` data set in the `mclust` R package.

This problem is challenging because the clutter form over two-thirds of the data points and are not separated from the mines spatially, but rather by their density.

Example 7: Vélib data

This data set has been extracted from the Vélib large-scale bike sharing system of Paris, through the open-data API provided by the operator JCDecaux. The real time data are available at `https://developer.jcdecaux.com/` (with an api key) . The data set consists of information (occupancy, number of broken docks, ...) about bike stations collected on the Paris bike sharing system over five weeks, between February 24 and March 30, 2014. Figure 1.4 presents a map of the Vélib stations in Paris (left panel) and loading profiles of some Vélib stations (right panel). The red dots correspond to the stations for which the loading profiles are displayed on the right panel.

The information can be analyzed in different ways, depending on the objective or the data type. For instance, the data were first used in Bouveyron et al. (2015) in the context of functional time series clustering, in order to recover the temporal pattern of use of the bike stations. This data set will be analyzed in this book

[1] Actual minefield data were not available, but the data in Figure 1.3 were simulated according to specifications developed at the Naval Coastal Systems Station, Panama City, Florida, to represent minefield data encountered in practice (Muise and Smith, 1992).