Cambridge Elements

1
Introduction

1.1 Graphical Models

Graphical models are an elegant framework that combines uncertainty and graph theory to represent complex phenomena. A graph is a structure consisting of a set of objects, called vertices, and a set of connections between pairs of vertices, called edges. The vertices of the graph associated with a graphical model are the variables of the model and the edges describe how the variables interact with each other. A further fundamental component of graphical models is the notion of independence or, more generally, of conditional independence. The edges missing from the graph can be interpreted as absence of interaction, in the sense that variables are conditionally independent. An appealing characteristic of graphs is that they can be represented graphically and many of their features and properties can be understood from the visual inspection of their graphical representation. The graph greatly simplifies the interpretation of the model, making its independence structure more immediate and intuitive. This also facilitates the communication of the scientific contents of the model to researchers who are not familiar with the statistical formalism. In large models the information provided by the visual inspection of the graph may be less clear but, nevertheless, the graph is useful in many ways. To mention a few, the graph structure may imply an intrinsic modularity that allows one to split the model into submodels of smaller dimensions, so that analyses of interest and statistical inference procedures can be carried out locally on marginal distributions. Furthermore, the graph is a natural object to be dealt with in the implementation of algorithms and computational procedures. All in all, graphical models constitute a very versatile methodology that has proved useful in a wide range of domains and applications. An historical overview of graphical models, with a comprehensive list of early references, can be found in Cox and

Wermuth (1996, section 2.13) and Lauritzen (1996, chapter 1), whereas we refer to Drton *et al*. (2017) for a recent collection of reviews.

1.2 Outline of the Book

Graphical modeling has been a very active area of research in the last years, and nowadays a wide range of families of graphical models are available; see Sadeghi and Lauritzen (2014). This book focuses on categorical data and describes the theory of some of the most relevant classes of graphical models. In this context, the book aims at providing a unified view of the theory discussed in a largely scattered literature.

The different families of models can be distinguished for the kind of graph associated with the probability distribution. The models considered in Chapters 4 and 5 deal with symmetric relationships between variables represented by undirected and bidirected graphs, respectively. Undirected graph models, also known as Markov random fields, are characterized by collections of conditional independence relationships, whereas bidirected graph models are characterized by collections of marginal independencies. Asymmetric relationships between variables are introduced in Chapter 6. Firstly, we consider the family of models associated with directed acyclic graphs (DAGs), also known as Bayesian networks, and then models associated with regression graphs. The latter family includes each of the previous families of models as special case and, therefore, regression graph models constitute a general framework that unifies and extends the theory of undirected, bidirected and directed acyclic graph models.

1.2.1 Discrete Graphical Models and Their Parameterization

The graph associated with the model allows one to abstract the conditional independence relationships between variables from the details of their parametric forms. Accordingly, the interpretation of a graphical model and the inferential questions of interest usually do not depend on the kinds of variables involved in the analysis. On the other hand, the effective specification of the statistical model and the implementation of statistical techniques require the definition of a suitable parameterization that cannot prescind from the variable types. From this perspective, the attention is restricted to categorical variables, that are variables

taking one of a finite number of possible values. Our approach to the specification of parameterizations aims at building a common framework encompassing all the statistical models considered. Chapter 2 deals with the probability distribution of a random vector in the form of a probability table, and provides a set of rules for establishing conditional independence, based on cross-product ratios. Chapter 3 gives the theory of Möbius inversion that is then used to transform probabilities into more convenient log-linear parameterizations. Jointly, Chapters 2 and 3 provide a set of tools that can be applied, almost identically, to all the families of models we consider to obtain the corresponding parameterizations and derive their properties.

1.2.2 Binary vs Non-binary Variables

In its simplest from, a categorical variable takes one of two possible values and is thus called a binary variable. Dealing with binary variables is especially convenient because, in this setting, the notation is less involved and the presentation of the material is more immediate. Furthermore, it is often the case that the properties of the binary case can be used as building blocks for the derivation of the corresponding properties for the general case. Throughout this text, we keep a clear distinction between the general case of arbitrary categorical variables and the special case of binary variables. We deem that this approach may facilitate the comprehension of the material because the treatment of the general case can often be regarded as a mere technical generalization of the binary case. For this reason, when suitable, we will first discuss the binary case in detail, and then we will consider the general case in a separate subsection, which the reader may choose to skip on first reading.