11. Mixtures of X

Summary.

An important contribution of the neural network community to the study of mixture models is the insight that the components of mixture models can be more complicated probability models than those typically used in the conventional statistics literature. This chapter reviews several members of a family of models that we refer to as “mixtures of X.” We believe that the recent development of several of these models has usefully expanded the boundaries of mixture modelling. The chapter reviews (i) mixtures of distributions from the exponential family, (ii) hidden Markov models, (iii) Mixtures of Experts, (iv) mixtures of marginal models, (v) mixtures of Cox models, (vi) mixtures of factor models, and (vii) mixtures of trees.

11.1 Introduction

Consider the task of summarising the data shown in Figure 11.1. A common technique in the neural network literature for performing this task is to use a statistical model known as a mixture model. Relative to many other models for estimating densities, mixture models have a number of advantages [25] [39]. First, mixture models can summarise data that contain multiple modes. In this sense, they are more powerful than distributions from the exponential family. Second, mixture models are parametric models. Methods based on probability theory, such as maximum likelihood and Bayesian inference methods, are often easily applied to mixture models. Third, mixture models are parsimonious in the sense that they typically combine distributions that are simple and relatively well-understood. In the conventional statistics literature, the components of mixture models are nearly always members of the exponential family of distributions.

A mixture model summarising the data in Figure 11.1 might contain two mixture components, each a Gaussian distribution. The two Gaussians would have different mean vectors and covariance matrices. The mean of one Gaussian would roughly be the point (3, 3); the mean of the second Gaussian would be the point (7, 7). Mixture models provide a principled way of combining the two (uni-modal) Gaussian distributions into a single (multi-modal) distribution that summarises the entire data set. As this example illustrates,
Fig. 11.1. Two-dimensional data to be summarised.