

Learning with Mixture Models: Concepts and Applications

Padhraic Smyth

Information and Computer Science, University of California
Irvine, CA 92697-3425, USA
`{smyth}@ics.uci.edu`

Abstract. Probabilistic mixture models have been used in statistics for well over a century as flexible data models. More recently these techniques have been adopted by the machine learning and data mining communities in a variety of application settings. We begin this talk with a review of the basic concepts of finite mixture models: what can they represent? how can we learn them from data? and so on. We will then discuss how the traditional mixture model (defined in a fixed dimensional vector space) can be usefully generalized to model non-vector data, such as sets of sequences and sets of curves. A number of real-world applications will be used to illustrate how these techniques can be applied to large-scale real-world data exploration and prediction problems, including clustering of visitors to a Web site based on their sequences of page requests, modeling of sparse high-dimensional “market basket” data for retail forecasting, and clustering of storm trajectories in atmospheric science.