Unsupervised Learning

Zoubin Ghahramani*

Gatsby Computational Neuroscience Unit, University College London, UK
zoubin@gatsby.ucl.ac.uk
http://www.gatsby.ucl.ac.uk/~zoubin

Abstract. We give a tutorial and overview of the field of unsupervised learning from the perspective of statistical modeling. Unsupervised learning can be motivated from information theoretic and Bayesian principles. We briefly review basic models in unsupervised learning, including factor analysis, PCA, mixtures of Gaussians, ICA, hidden Markov models, state-space models, and many variants and extensions. We derive the EM algorithm and give an overview of fundamental concepts in graphical models, and inference algorithms on graphs. This is followed by a quick tour of approximate Bayesian inference, including Markov chain Monte Carlo (MCMC), Laplace approximation, BIC, variational approximations, and expectation propagation (EP). The aim of this chapter is to provide a high-level view of the field. Along the way, many state-of-the-art ideas and future directions are also reviewed.

1 Introduction

Machine learning is the field of research devoted to the formal study of learning systems. This is a highly interdisciplinary field which borrows and builds upon ideas from statistics, computer science, engineering, cognitive science, optimization theory and many other disciplines of science and mathematics. The purpose of this chapter is to introduce in a fairly concise manner the key ideas underlying the sub-field of machine learning known as *unsupervised learning*. This introduction is necessarily incomplete given the enormous range of topics under the rubric of unsupervised learning. The hope is that interested readers can delve more deeply into the many topics covered here by following some of the cited references. The chapter starts at a highly tutorial level but will touch upon state-of-the-art research in later sections. It is assumed that the reader is familiar with elementary linear algebra, probability theory, and calculus, but not much else.

1.1 What Is Unsupervised Learning?

Consider a machine (or living organism) which receives some sequence of inputs $x_1, x_2, x_3, \ldots$, where x_t is the sensory input at time t. This input, which we will

* The author is also at the Center for Automated Learning and Discovery, Carnegie Mellon University, USA.

O. Bousquet et al. (Eds.): Machine Learning 2003, LNAI 3176, pp. 72–112, 2004.

often call the *data*, could correspond to an image on the retina, the pixels in a camera, or a sound waveform. It could also correspond to less obviously sensory data, for example the words in a news story, or the list of items in a supermarket shopping basket.

One can distinguish between four different kinds of machine learning. In *supervised learning* the machine[1] is also given a sequence of desired outputs $y_1, y_2, \ldots,$ and the goal of the machine is to learn to produce the correct output given a new input. This output could be a class label (in classification) or a real number (in regression).

In *reinforcement learning* the machine interacts with its environment by producing actions $a_1, a_2, \ldots$. These actions affect the state of the environment, which in turn results in the machine receiving some scalar rewards (or punishments) $r_1, r_2, \ldots$. The goal of the machine is to learn to act in a way that maximizes the future rewards it receives (or minimizes the punishments) over its lifetime. Reinforcement learning is closely related to the fields of decision theory (in statistics and management science), and control theory (in engineering). The fundamental problems studied in these fields are often formally equivalent, and the solutions are the same, although different aspects of problem and solution are usually emphasized.

A third kind of machine learning is closely related to *game theory* and generalizes reinforcement learning. Here again the machine gets inputs, produces actions, and receives rewards. However, the environment the machine interacts with is not some static world, but rather it can contain other machines which can also sense, act, receive rewards, and learn. Thus the goal of the machine is to act so as to maximize rewards in light of the other machines' current and future actions. Although there is a great deal of work in game theory for simple systems, the dynamic case with multiple adapting machines remains an active and challenging area of research.

Finally, in *unsupervised learning* the machine simply receives inputs $x_1, x_2, \ldots$, but obtains neither supervised target outputs, nor rewards from its environment. It may seem somewhat mysterious to imagine what the machine could possibly learn given that it doesn't get any feedback from its environment. However, it is possible to develop of formal framework for unsupervised learning based on the notion that the machine's goal is to build representations of the input that can be used for decision making, predicting future inputs, efficiently communicating the inputs to another machine, etc. In a sense, unsupervised learning can be thought of as finding patterns in the data above and beyond what would be considered pure unstructured noise. Two very simple classic examples of unsupervised learning are clustering and dimensionality reduction. We discuss these in Section 2. The remainder of this chapter focuses on unsupervised learning,

[1] Henceforth, for succinctness I'll use the term machine to refer both to machines and living organisms. Some people prefer to call this a system or agent. The same mathematical theory of learning applies regardless of what we choose to call the learner, whether it is artificial or biological.