

Unsupervised Learning

Anil Jain
Michigan State University

Abstract:

A major dichotomy in statistical pattern recognition is between supervised learning (labeled training samples) and unsupervised learning (unlabeled training samples). The label on each training pattern represents the category to which that pattern belongs. In many applications of pattern recognition, it is extremely difficult or expensive, or even impossible, to reliably label a training sample with its true category. Unsupervised learning refers to situations where the objective is to construct decision boundaries based on unlabeled training data. Unsupervised learning is also known as data clustering which is a generic label for a variety of methods designed to find natural groupings, or clusters, in multidimensional data, based on measured or perceived similarities among the patterns. Clustering is useful in several exploratory pattern analysis, grouping, decision making, and machine learning situations, including data mining, information retrieval, signal compression and coding, and image segmentation. As a consequence, hundreds of clustering algorithms have been proposed in the literature and new clustering algorithms continue to appear. This talk will address the difficulty of data clustering problem, major approaches to clustering, finding a suitable data representation (unsupervised feature selection/extraction), and the issue of cluster validation.