Multiple Cause Vector Quantization

David Ross and Richard Zemel
Department of Computer Science
University of Toronto
Toronto, ON M5S 1A4
{dross,zemel}@cs.toronto.edu

The aim of *unsupervised learning* is to develop more useful representations of data in situations in which correct labels are not provided. This extends learning to domains in which specific supervisory signals are not available. One of the dominant approaches to unsupervised learning involves *stochastic generative models*, in which each data item is characterized as having been generated from a model of random variables and their relationships. Adopting this approach permits the application of principled, statistical methods to optimize and interpret the parameters of the model. Most standard unsupervised learning algorithms can be re-formulated as a generative model, with particular assumptions about the underlying random variables.

We propose a generative model for collections of high-dimensional data, such as images and text, that improves on several previous models, including naive Bayes, probabilistic latent semantic indexing, and non-negative matrix factorization. Our key assumption is that the dimensions of the data can be separated into several disjoint subsets, or *multiple causes*, which take on values independently of each other. Given a set of training examples, our system learns the association of data dimensions with the causes. We also assume each cause is a *vector quantizer*, with a small number of discrete states. The system also learns these particular states from the training examples. For example, given a set of face images, the causes could correspond to eyes, nose, and mouth, and the values within each cause could represent different examples of these facial features. As opposed to standard vector quantization, which posits many values of a single cause, the use of multiple causes to account for a single input allows for a richer repertoire of inputs through the combination of different values of each cause.

Problem domains where this model could prove effective include text categorization, collaborative filtering, and object detection/recognition. In particular, recent work by Tomaso Poggio and colleagues [1] has demonstrated the efficacy of a parts-based approach for detect-

ing highly variable objects, such as the human body, in images. The drawback with current techniques is that they require the experimenter to specify the segmentation of the object into parts, as well as to manually locate these parts in the training images. Experiments show that multiple cause vector quantization can be used to automate this segmentation, potentially generating a more discriminative set of parts on which to train supervised detectors.

References

[1] A. Mohan, C. Papageorgiou, and T. Poggio, "Example-Based Object Detection in Images by Components," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 4, pp. 349-361, April 2001.