Appeared in Wilson, RA & Keil, F, editors. *The MIT Encyclopedia of the Cognitive Sciences.*

Unsupervised Learning

Peter Dayan

MIT

Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns. By contrast with SUPERVISED LEARNING or REINFORCEMENT LEARNING, there are no explicit target outputs or environmental evaluations associated with each input; rather the unsupervised learner brings to bear prior biases as to what aspects of the structure of the input should be captured in the output.

Unsupervised learning is important since it is likely to be much more common in the brain than supervised learning. For instance there are around 10^6 photoreceptors in each eye whose activities are constantly changing with the visual world and which provide all the information that is available to indicate what objects there are in the world, how they are presented, what the lighting conditions are, *etc.* Developmental and adult plasticity are critical in animal vision (see VISION AND LEARNING) – indeed structural and physiological properties of synapses in the neocortex are known to be substantially influenced by the patterns of activity in sensory neurons that occur. However, essentially none of the information about the contents of scenes is available during learning. This makes unsupervised methods essential, and, equally, allows them to be used as computational models for synaptic adaptation.

The only things that unsupervised learning methods have to work with are the observed input patterns $\mathbf{x}_i$, which are often assumed to be independent samples from an underlying unknown probability distribution $\mathcal{P}_I[\mathbf{x}]$, and some explicit or implicit *a priori* information as to what is important. One key notion is that input, such as the image of

a scene, has distal independent *causes,* such as objects at given locations illuminated by particular lighting. Since it is on those independent causes that we normally must act, the best representation for an input is in their terms. Two classes of method have been suggested for unsupervised learning. Density estimation techniques explicitly build statistical models (such as BAYESIAN NETWORKS) of how underlying causes could create the input. Feature extraction techniques try to extract statistical regularities (or sometimes irregularities) directly from the inputs.

Unsupervised learning in general has a long and distinguished history. Some early influences were Horace Barlow (see Barlow, 1992), who sought ways of characterising neural codes, Donald MacKay (1956), who adopted a cybernetic-theoretic approach, and David Marr (1970), who made an early unsupervised learning postulate about the goal of learning in his model of the neocortex. The Hebb rule (Hebb, 1949), which links statistical methods to neurophysiological experiments on plasticity, has also cast a long shadow. Geoffrey Hinton and Terrence Sejnowski in inventing a model of learning called the Boltzmann machine (1986), imported many of the concepts from statistics that now dominate the density estimation methods (Grenander, 1976-1981). Feature extraction methods have generally been less extensively explored.

Clustering provides a convenient example. Consider the case in which the inputs are the photoreceptor activities created by various images of an apple or an orange. In the space of all possible activities, these particular inputs form two clusters, with many fewer degrees of variation than 10^6, *ie* lower *dimension*. One natural task for unsupervised learning is to find and characterise these separate, low dimensional clusters.

The larger class of unsupervised learning methods consists of maximum likelihood (ML) density estimation methods. All of these are based on building parameterised *models*

$\mathcal{P}[\mathbf{x};\mathcal{G}]$ (with parameters $\mathcal{G}$) of the probability distribution $\mathcal{P}_I[\mathbf{x}]$, where the forms of the models (and possibly prior distributions over the parameters $\mathcal{G}$) are constrained by *a priori* information in the form of the representational goals. These are called *synthetic* or *generative* models, since, given a particular value of $\mathcal{G}$, they specify how to synthesise or generate samples $\mathbf{x}$ from $\mathcal{P}[\mathbf{x};\mathcal{G}]$, whose statistics should match $\mathcal{P}_I[\mathbf{x}]$. A typical model has the structure:

$$\mathcal{P}[\mathbf{x};\mathcal{G}] = \sum_{\mathbf{y}} \mathcal{P}[\mathbf{y};\mathcal{G}]\mathcal{P}[\mathbf{x}|\mathbf{y};\mathcal{G}]$$

where $\mathbf{y}$ represents all the potential causes of the input $\mathbf{x}$. The typical measure of the degree of mismatch is called the Kullback-Leibler divergence:

$$KL\left[\mathcal{P}_I[\mathbf{x}], \mathcal{P}[\mathbf{x};\mathcal{G}]\right] = \sum_{\mathbf{x}} \mathcal{P}_I[\mathbf{x}] \log\left[\frac{\mathcal{P}_I[\mathbf{x}]}{\mathcal{P}[\mathbf{x};\mathcal{G}]}\right] \geq 0$$

with equality if and only if $\mathcal{P}_I[\mathbf{x}] = \mathcal{P}[\mathbf{x};\mathcal{G}]$.

Given an input pattern $\mathbf{x}$, the most general output of this model is the posterior, *analytical*, or *recognition* distribution $\mathcal{P}[\mathbf{y}|\mathbf{x};\mathcal{G}]$, which recognises which particular causes might underlie $\mathbf{x}$. This analytical distribution is the statistical inverse of the synthetic distribution.

A very simple model can be used in the example of clustering (Nowlan, 1990). Consider the case in which there are two values for y (1 and 2), with $\mathcal{P}[y=1] = \pi; \mathcal{P}[y=2] = 1-\pi$, where π is called a mixing proportion, and two different Gaussian distributions for the activities $\mathbf{x}$ of the photoreceptors depending on which y is chosen: $\mathcal{P}[\mathbf{x}|y=1] \sim \mathcal{N}[\boldsymbol{\mu}^1, \Sigma^1]$ and $\mathcal{P}[\mathbf{x}|y=2] \sim \mathcal{N}[\boldsymbol{\mu}^2, \Sigma^2]$, where $\boldsymbol{\mu}^*$ are means and Σ^* are covariance matrices. Unsupervised learning of the means determines the clusters. Unsupervised learning of the mixing proportions and the covariances characterises the size and (rather coarsely) the shape of the clusters. The posterior distribution $\mathcal{P}[y=1|\mathbf{x};\pi,\boldsymbol{\mu}^*,\Sigma^*]$ reports how likely it is that a new image $\mathbf{x}$ was generated from the first cluster, *ie* that $y=1$ is the true

hidden cause. Clustering can occur (with or) without any supervision information about the different classes. This model is called a *mixture of Gaussians.*

Maximum likelihood density estimation, and approximations to it, cover a very wide spectrum of the principles that have been suggested for unsupervised learning. This includes versions of the notion that the outputs should convey most of the information in the input; that they should be able to reconstruct the inputs well, perhaps subject to constraints such as being independent or sparse; and that they should report on the underlying causes of the input. Many different mechanisms apart from clustering have been suggested for each of these, including forms of Hebbian learning, the Boltzmann and Helmholtz machines, sparse-coding, various other mixture models, and independent components analysis.

Density estimation is just a heuristic for learning good representations. It can be too stringent — making it necessary to build a model of all the irrelevant richness in sensory input. It can also be too lax — a look-up table that reported $\mathcal{P}_I[\mathbf{x}]$ for each $\mathbf{x}$ might be an excellent way of modeling the distribution, but provides no way to represent particular examples $\mathbf{x}$.

The smaller class of unsupervised learning methods seeks to discover how to represent the inputs $\mathbf{x}$ by defining some quality that good 6features have, and then searching for those features in the inputs. For instance, consider the case that the output $y(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$ is a linear projection of the input onto a weight vector $\mathbf{w}$. The central limit theorem implies that most such linear projections will have Gaussian statistics. Therefore if one can find weights $\mathbf{w}$ such that the projection has a highly non-Gaussian (for instance, multi-modal) distribution, then the output is likely to reflect some interesting aspect of the input. This is the intuition behind a statistical method called projection pursuit. It has been shown

that projection pursuit can be implemented using a modified form of Hebbian learning (Intrator & Cooper, 1992). Arranging that different outputs should represent different aspects of the input turns out to be surprisingly tricky.

Projection pursuit can also execute a form of clustering in the example. Consider projecting the photoreceptor activities onto the line joining the centers of the clusters. The distribution of all activities will be bimodal – one mode for each cluster – and therefore highly non-Gaussian. Note that this single projection does not characterise well the nature or shape of the clusters.

Another example of a heuristic underlying good features is that causes are often somewhat global. For instance, consider the visual input from an object observed in depth. Different parts of the object may share few features, except that they are at the same depth, *ie* one aspect of the disparity in the information from the two eyes at the separate locations is similar. This is the global underlying feature. By maximising the mutual information between outputs y_a and $y_{a'}$ that are calculated on the basis of the separate input, one can find this disparity. This technique was invented by Becker & Hinton (1992) and is called IMAX.

References

Barlow, HB (1989). Unsupervised learning. *Neural Computation*, **1**, 295-311.

Becker, S & Hinton, GE (1992). A self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, **355**, 161-163.

Grenander, U (1976-1981). *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures.* Berlin: Springer-Verlag., Berlin, 1976-1981).

Hebb, DO (1949) *The Organization of Behavior*. New York: Wiley.

Hinton, GE & Sejnowski, TJ (1986). Learning and relearning in Boltzmann machines. In DE Rumelhart, JL McClelland and the PDP research group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations,* Cambridge, MA: MIT Press, 282-317.

Intrator, N & Cooper, LN (1992). Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability conditions. *Neural Networks*, **5**, 3-17.

MacKay, DM (1956). The epistemological problem for automata. In CE Shannon & J McCarthy, editors, *Automata Studies,* Princeton, NJ: Princeton University Press, 235-251.

Marr, D (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London, Series B*, **176**, 161-234.

Nowlan, SJ (1990). Maximum likelihood competitive learning. In DS Touretzky, editor, *Advances in Neural Information Processing Systems, 2.* San Mateo, CA: Morgan Kaufmann.

Further Readings

Becker, S & Plumbley, M (1996). Unsupervised neural network learning procedures for feature extraction and classification. *International Journal of Applied Intelligence*, **6**, 185-203.

Dayan, P, Hinton, GE, Neal, RM & Zemel, RS (1995). The Helmholtz machine. *Neural Computation*, **7**, 889-904.

Hinton, GE (1989a). Connectionist learning procedures. *Artificial Intelligence*, **40**, 185-234.

Linsker, R (1988). Self-organization in a perceptual network. *Computer*, **21**, 105-128.

Mumford, D (1994). Neuronal architectures for pattern-theoretic problems. In C Koch and J Davis, editors, *Large-Scale Theories of the Cortex*. Cambridge, MA: MIT Press, 125-152.

Olshausen, BA & Field, DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, **381**, 607-609.