The mixture of Gaussians model

Bruno A. Olshausen

March 4, 2004

The mixture of Gaussians model is probably the simplest (interesting) example of a generative model that illustrates the principles of inference and learning. It is particularly well suited to describe data containing clusters. The probability density over data vectors $\mathbf{x}$ is specified as

$$p(\mathbf{x}) = \sum_{\alpha=1}^{K} p(\mathbf{x}|\alpha)\ P(\alpha) \tag{1}$$

where $p(\mathbf{x}|\alpha)$ is a Gaussian distribution with mean μ_α and variance σ_α (in this case, isotropic), and $P(\alpha)$ specifies the probability of drawing from that distribution. Thus, a data vector $\mathbf{x}$ is generated by first selecting a Gaussian to draw from with probability $P(\alpha)$, and then drawing $\mathbf{x}$ from $p(\mathbf{x}|\alpha)$.

The problem of *inference* is to determine α given a data vector $\mathbf{x}$ drawn from the world (e.g., sensory data). Thus, we are essentially asking, "what cluster does this data vector belong to?" The problem of *learning* is to determine the parameters μ_α and σ_α that best fit the data.

Inference

In order to determine which cluster a given data vector $\mathbf{x}$ belongs to, we need to compute the posterior distribution:

$$P(\alpha|\mathbf{x}) = \frac{p(\mathbf{x}|\alpha)\ P(\alpha)}{p(\mathbf{x})} \tag{2}$$

It makes the most sense then to choose the cluster α that maximizes this distribution.

Learning

Learning the parameters is accomplished by doing gradient ascent on the log-likelihood of the data:

$$\langle \log p(\mathbf{x}) \rangle \tag{3}$$

Computing derivatives of this function with respect to μ_α, σ_α, and $P(\alpha)$ yields the following learning rules:

$$\Delta\mu_\alpha \ \propto\ \sum_\alpha P(\alpha|\mathbf{x})\lambda_\alpha(\mathbf{x} - \mu_\alpha) \tag{4}$$

$$\Delta\lambda_\alpha \propto -\frac{1}{2}\sum_\alpha P(\alpha|\mathbf{x})|\mathbf{x}-\mu_\alpha|^2 \quad (5)$$

$$\Delta\gamma_\alpha \propto P(\alpha|\mathbf{x}) - P(\alpha) \quad (6)$$

where $\sigma^2 = 1/\lambda$ and $P(\alpha) = e^{\gamma_\alpha}/\sum_\beta e^{\gamma_\beta}$.

Note that all of the above expressions utilize $P(\alpha|\mathbf{x})$. Thus, learning draws upon inference. That is, in order to update the parameters, you need to infer causes (α) given the current model. This process is iteratively repeated until the parameters converge to best model the distribution of the data.