Machine Learning (CS 5350/CS 6350) *27 Oct 2009*

Gaussian Mixture Models

We can think of k-means clustering in a probabilistic framework. Suppose that we have a Gaussian centered at each of the means, then we can get the probability of the data set as:

$$p(x_{1:N} \mid c_{1:N}) = \prod_n \mathcal{N}or(x_n \mid \mu_{c_n}, \sigma^2 I)$$

Here, μ_k is the the mean of cluster k. (For now, we assume common variance.)

We can think of the clustering problem as trying to find good μ_ks.

Change of notation: Instead of c_n being the cluster for data point n, let $z_n \in \{0,1\}^K$ be an indicator vector for data point n. I.e., $z_{n,k} = 1$ if x_n is in k and $= 0$, otherwise.

Model: Generate each data point by first choosing among one of k clusters, each with probability π_k. Then generate the data point by a Gaussian centered at μ_k. In equations:

$$\begin{aligned} p(x_{1:N}, z_{1:N,1:K} \mid \mu_{1:K}, \sigma^2, \pi) &= \prod_n \prod_k \left\{\pi_k \mathcal{N}or(x_n \mid \mu_k, \sigma^2 I)\right\}^{z_{n,k}} \\ &= \prod_n \prod_k \left\{\pi_k (2\pi(\sigma^2)^d)^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2\right]\right\}^{z_{n,k}} \end{aligned}$$

From this, we get the likelihood of the data by summing over the unknown zs:

$$\begin{aligned} p(x \mid \mu, \sigma^2, \pi) &= \sum_{z_{1:N,1:K}} \prod_n \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2\right]\right\}^{z_{n,k}} \\ &= \prod_n \sum_{z_{n,1:K}} \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2\right]\right\}^{z_{n,k}} \end{aligned}$$

So now we follow our standard recipe of taking logs and derivatives...

$$\log p(x \mid \mu, \sigma^2, \pi) = \sum_n \log \sum_{z_n} \prod_k \left\{\pi_k (2\pi\sigma^{2d})^{-1/2} \exp\left[-\frac{1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2\right]\right\}^{z_{n,k}}$$

But at this point we get stuck!

If we *knew* z, we could do this easily:

$$\log p(x, z \mid \mu, \sigma^2, \pi) = \sum_n \sum_k z_{n,k} \left\{\log \pi_k + \log \mathcal{N}or(x_n \mid \mu_k, \sigma^2 I)\right\}$$

We call the value *with* z the "complete log likelihood" and the value *without* z the "incomplete log likelihood."

The idea for clustering with GMMs is the same as for k-means. We will make an initial guess at z, and then try to iteratively refine it. This turns out to be a special case of the "expectation maximization" algorithm, which we will discuss shortly in more generality.

In k-means, we made "hard" guesses at the clusters: the z vector we considered had a one in a single location and zeros everywhere else. In Gaussian mixture models, we make "soft" guesses. The z vector will satisfy $z_{n,k} \geq 0$ for all n, k and $\sum_k z_{n,k} = 1$ for all n. Thus, it's a probabilistic guess at the clustering.

Given some setting of μ and σ^2, we can make guesses at z by just looking at their expectations:

$$\begin{aligned}
\mathbb{E}_{p(z \mid x,\mu,\sigma^2,\pi)} z_{n,k} &= 1 \times p(z_{n,k} = 1 \mid x_n, \mu, \sigma^2, \pi) + 0 \times p(z_{n,k} = 0 \mid x_n, \mu, \sigma^2, \pi) \\
&= p(z_{n,k} = 1 \mid x_n, \mu, \sigma^2, \pi) \\
&= \frac{p(x_n \mid z_{n,k} = 1, \mu, \sigma^2) p(z_{n,k} = 1 \mid \pi)}{\sum_{k'} p(x_n \mid z_{n,k'} = 1, \mu, \sigma^2) p(z_{n,k'} = 1 \mid \pi)} \\
&= \frac{\mathcal{N}or(x_n \mid \mu_k, \sigma^2) \pi_k}{\sum_{k'} \mathcal{N}or(x_n \mid \mu_{k'}, \sigma^2) \pi_{k'}}
\end{aligned}$$

These expectations give us a soft clustering for each data point into each of the k clusters.

Now, using these "guesses", we want to maximize the *complete* data log likelihood with respect to μ and σ^2. To do this, we take the gradient of the complete likelihood with respect to π, μ and σ^2.

We do π first. For this, we actually need to introduce a Lagrange multiplier to ensure that the constraints on π are satisfied ($\pi_k \geq 0$ and $\sum_k \pi_k = 1$). This gives us an augmented likelihood function of:

$$\sum_n \sum_k z_{n,k} \log \pi_k - \lambda \left(\sum_k \pi_k - 1 \right)$$

We differentiate this with respect to π_k to get:

$$\sum_n \frac{z_{n,k}}{\pi_k} - \lambda = \sum_n z_{n,k} - \lambda \pi_k = 0$$

Summing over all K, we get that $\lambda = \sum_n \sum_k z_{n,k} = N$, so:

$$\pi_k = \frac{1}{N} \sum_n z_{n,k}$$

This makes intuitive sense!

Next, we'll take care of μ_k. These are somewhat easier since we don't need to worry about constraints, so there are no Lagrange multipliers.

Here, we take the gradient of the complete log likelihood with respect to μ_k:

$$\begin{aligned}\nabla_{\mu_k} \log p(x, z \mid \mu, \sigma^2, \pi) &= \sum_n z_{n,k} \nabla_{\mu_k} \log \mathcal{N}or(x_n \mid \mu_k, \sigma^2 I) \\ &= \sum_n z_{n,k} \nabla_{\mu_k} \frac{-1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2 \\ &= -\sum_n z_{n,k} \frac{1}{\sigma^2}(x_n - \mu_k)\end{aligned}$$

We equate this to zero to give:

$$\begin{aligned}&\sum_n z_{n,k} \frac{1}{\sigma^2}(\mu_k - x_n) = 0 \\ \Longrightarrow &\sum_n z_{n,k} \frac{1}{\sigma^2} \mu_k = \sum_n z_{n,k} \frac{1}{\sigma^2} x_n \\ \Longrightarrow &\mu_k \sum_n z_{n,k} = \sum_n z_{n,k} x_n \\ \Longrightarrow &\mu_k = \sum_n \frac{z_{n,k}}{\sum_{n'} z_{n',k}} x_n\end{aligned}$$

Again, this result is intuitive!

Finally, we deal with σ^2.

$$\begin{aligned}\nabla_{\sigma^2} \log p(x, z \mid \mu, \sigma^2, \pi) &= \sum_n \sum_k z_{n,k} \nabla_{\sigma^2} \log \mathcal{N}or(x_n \mid \mu_k, \sigma^2) \\ &= \sum_n \sum_k z_{n,k} \nabla_{\sigma^2} \left[\frac{-d}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \left|\left|x_n - \mu_k\right|\right|^2\right] \\ &= \sum_n \sum_k z_{n,k} \left[\frac{-d}{2\sigma^2} + \frac{1}{2\sigma^4} \left|\left|x_n - \mu_k\right|\right|^2\right]\end{aligned}$$

We set this equal to zero to obtain:

$$\begin{aligned}&\sum_n \sum_k z_{n,k} \frac{d}{2\sigma^2} = \sum_n \sum_k z_{n,k} \frac{1}{2\sigma^4} \left|\left|x_n - \mu_k\right|\right|^2 \\ \Longrightarrow &\frac{d}{\sigma^2} \sum_n \sum_k z_{n,k} = \frac{1}{\sigma^4} \sum_n \sum_k z_{n,k} \left|\left|x_n - \mu_k\right|\right|^2 \\ \Longrightarrow &dN\sigma^2 = \sum_n \sum_k z_{n,k} \left|\left|x_n - \mu_k\right|\right|^2 \\ \Longrightarrow &\sigma^2 = \frac{1}{dN} \sum_n \sum_k z_{n,k} \left|\left|x_n - \mu_k\right|\right|^2\end{aligned}$$

Putting this all together, we obtain the following algorithm:

- Initialize cluster centers $\mu_{1:K}$, π and σ^2
- Iterate T times . . .
    - Compute expectation of z variables by:

$$\mathbb{E}_{p(z \mid x,\mu,\sigma^2,\pi)} z_{n,k} = \frac{\mathcal{N}or(x_n \mid \mu_k, \sigma^2)\pi_k}{\sum_{k'} \mathcal{N}or(x_n \mid \mu_{k'}, \sigma^2)\pi_{k'}}$$

    - Compute new values of π, μ, σ^2 by:

$$\begin{aligned} \pi_k &= \frac{1}{N}\sum_n z_{n,k} \\ \mu_k &= \sum_n \frac{z_{n,k}}{\sum_{n'} z_{n',k}} x_n \\ \sigma^2 &= \frac{1}{dN}\sum_n \sum_k z_{n,k} \left|\left|x_n - \mu_k\right|\right|^2 \end{aligned}$$

One can obtain a more general solution, where we use full covariance matrices and/or cluster-specific covariance matrices.