Gaussian Mixture Models

Rebecca C. Steorts

Simple EM Algorithm

Notation and Setup

We know the following:

- Observations $x_{1:n}$.
- K total classes
- $P(Z_i = k) = \pi_k$ (for $i = 1, \dots, K$)
- Common variance σ^2.

We do not know $\mu_1, \dots, \mu_K$ and want to learn these.

This is a very unrealistic setting, however, it hopefully provides intuition regarding the algorithm itself (and the math is simplified).

EM Algorithm

$\propto$ will drop any constants (and I will make sure to include them back in later). Common trick in Bayesian statistics.

$$
\begin{aligned}
&p(x_1, \dots, x_n \mid \mu_1, \dots, \mu_K) && (1)\\
&= \prod_{i=1}^{n} p(x_i \mid \mu_1, \dots, \mu_K) \text{ independent data} && (2)\\
&= \prod_{i=1}^{n} \sum_{k=1}^{K} p(x_i, z_i = k \mid \mu_1, \dots, \mu_K) \text{ marg. over labels} && (3)\\
&= \prod_{i=1}^{n} \sum_{k=1}^{K} p(x_i \mid z_i = k, \mu_1, \dots, \mu_K) p(z_i = k) && (4)\\
&\propto \prod_{i=1}^{n} \sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k)^2)\pi_k \;\text{ dropped normal constants} && (5)
\end{aligned}
$$

EM Algorithm

Let $\theta^{(t)} = (\mu_1^{(t)}, \dots, \mu_k^{(t)})$ at some iteration t.

At iteration t consider the function:

$$
\begin{aligned}
Q(\theta^{(t)} \mid \theta^{(t-1)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) && (6) \\
\times \log P(x_i, z_i = k \mid \theta^{(t-1)}) && (7)
\end{aligned}
$$

E-step

$$
\begin{aligned}
& P(z_i = k \mid x_i, \theta^{t-1}) && (8) \\
& = P(z_i = k \mid x_i, \mu_1^{(t-1)}, \ldots, \mu_K^{(t-1)}) && (9) \\
& \propto P(x_i \mid z_i = k, \mu_1^{(t-1)}, \ldots, \mu_K^{(t-1)}) P(z_i = k) && (10) \\
& \propto \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k && (11) \\
& = \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k}{\sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k} && (12)
\end{aligned}
$$

This is equivalent to assigning clusters to each data point in a soft-way (clusters can overlap).

M-step

Recall that in the E-step, we calculated $R_{ik}^{(t-1)} = P(z_i = k \mid x_i, \theta^{(t-1)})$

$$
\begin{aligned}
& Q(\theta^{(t)} \mid \theta^{(t-1)}) && (13) \\
& = \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) \times \log P(x_i, z_i = k \mid \theta^{(t-1)}) && (14) \\
& = \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) && (15) \\
& \times [\log P(x_i \mid z_i = k, \theta^{(t-1)}) + \log P(z_i = k \mid \theta^{(t-1)})] && (16) \\
& = \sum_{i=1}^{n} \sum_{k=1}^{K} R_{ik}^{(t-1)} [-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2) + \log \pi_k] && (17)
\end{aligned}
$$

M-step

At each iteration t, maximize Q in term of $\theta^{(t)}$.

$$
\begin{aligned}
Q(\mu_k^{(t)} \mid \theta^{(t-1)}) \propto \sum_{i=1}^{n} R_{ik}^{(t-1)} (-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2), \implies && (18) \\
\frac{\partial Q(\mu_k^{(t)} \mid \theta^{(t-1)})}{\partial \mu_k^{(t)}} = \sum_{i=1}^{n} R_{ik}^{(t-1)} (x_i - \mu_k^{(t-1)})) = 0 \implies && (19)
\end{aligned}
$$

$$
\mu_k^{(t)} = \sum_{i=1}^{n} w_i x_i \quad \text{where}
$$

$$w_i = \frac{R_{ik}^{t-1}}{\sum_{i=1}^{n} R_{ik}^{t-1}} = \frac{P(z_i = k \mid x_i, \theta^{(t-1)})}{\sum_{i=1}^{n} P(z_i = k \mid x_i, \theta^{(t-1)})}$$

This is equivalent to updating the cluster centers.

Summarize EM Algorithm

1. E-step

Compute the expected classes of all data points for each class:

$$P(z_i = k \mid x_i, \theta^{(t-1)}) = \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k^{(t-1)}}{\sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k^{(t-1)}}$$

2. M-step

Then compute the maximum value given our data's class membership.

$$\mu_i^{(t)} = \sum_{i=1}^{n} w_i x_i.$$

In this case, it's the MLE but with weighted data.