Lecture 22,23	Nov. 8,10 -2006
Data Visualization	
STAT 442 / 890, CM 462	
Lecture: Ali Ghodsi	Scribes: Stefan Pintilie

1 Mixture Models

In this section the data points come from a density function. Lets assume that the data comes from a Gaussian, or a mixture of Gaussian. We have chosen the Gaussian distribution so that it is easier to find the mean and variance. The task of clustering is the task of finding the parameter estimates for each of the Gaussian distributions for the given data set.

1.1 Reminder

According to Bayes rule:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Where:

1. $P(x|y)$ is called to posterior. In order to find the posterior we need to know a prior.
2. $P(x)$ is called the prior and its values must be assumed based on some prior knowledge.
3. $P(y|x)$ is called the likelihood and in order to calculate it we need to assume a distribution for the data.

1.2 Mixture of Gaussian

Suppose we have a mixture of two Gaussian.

$$P(x|\theta) = \alpha N(x; \mu_1, \sigma_1) + (1 - \alpha)N(x; \mu_2, \sigma_2)$$

Where $\theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \alpha\}$. We can assume that there is a hidden variable Z which picks the first normal with probability α and it picks the second normal with probability $(1 - \alpha)$. We can then redefine the set of input data as $Data = \{(x_1, z_1), (x_2, z_2), \ldots (x_n, z_n)\}$ where the z_i are unknown. Then,

$$\begin{aligned} P(Data|\theta) &= P(x, z|\theta) \\ &= P(x|z, \theta)P(z|\theta) \end{aligned}$$

If $z_i = 1$ then $P(X|Z, \theta) = \phi_1(x) = N(x_i; \mu_1, \sigma_1)$.

If $z_i = 0$ then $P(X|Z, \theta) = \phi_2(x) = N(x_i; \mu_2, \sigma_2)$.

Therefore, $P(x_i|z_i, \theta) = \phi_1(x)^{z_i}\phi_2(x)^{1-z_i}$.

We can also rewrite the other term in the product as: $P(z_i|\theta) = \alpha^{z_i}\alpha^{1-z_i}$. And therefore we get

$$P(X|Z, \theta) = \phi_1(x)^{z_i}\phi_2(x)^{1-z_i}\alpha^{z_i}\alpha^{1-z_i}$$

The complete likelihood can be written as follows.

$$L_c(X, \theta) = \prod_{i=1}^{n} P(x_i|z_i, \theta)$$

It would be useful to take the expectation of the log likelihood but for this we need to assume that the expectation of z_i is known. So assume that $E[z_i] = w_i$.

$$\begin{aligned} L_c(X,\theta) &= \phi_1(x)^{z_i}\phi_2(x)^{1-z_i}\alpha^{z_i}\alpha^{1-z_i} \\ l_c(X,\theta) &= z_i log(\phi_1(x)) + (1-z_i)log(\phi_2(x)) + z_i log(\alpha) + (1-z_i)log(1-\alpha) \\ E[l_c(X,\theta)] &= w_i log(\phi_1(x)) + (1-w_i)log(\phi_2(x)) + w_i log(\alpha) + (1-w_i)log(1-\alpha) \end{aligned}$$

1.2.1 M-Step

Assume $E[z_i] = w_i$. We need to maximise the expectation of the log likelihood. Take the derivative with respect to each parameter in θ and set it equal to zero.

$$\begin{aligned} \frac{dl}{d\alpha} = 0 &\Rightarrow \alpha = \frac{\sum_{i=1}^n w_i}{n} \\ \frac{dl}{d\mu_1} = 0 &\Rightarrow \mu_1 = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} \\ \frac{dl}{d\mu_2} = 0 &\Rightarrow \mu_2 = \frac{\sum_{i=1}^n (1-w_i) x_i}{\sum_{i=1}^n (1-w_i)} \\ \frac{dl}{d\sigma_1} = 0 &\Rightarrow \sigma_1 = \frac{\sum_{i=1}^n w_i (x_i-\mu_1)^2}{\sum_{i=1}^n w_i} \\ \frac{dl}{d\sigma_2} = 0 &\Rightarrow \sigma_2 = \frac{\sum_{i=1}^n (1-w_i)(x_i-\mu_2)^2}{\sum_{i=1}^n (1-w_i)} \end{aligned}$$

The M-Step can estimate the parameters θ if we have w_i.

1.2.2 E-Step

To get $E[z_i]$ we need to know the parameters θ. The purpose of the E-Step is to find w_i if we know θ.

$$\begin{aligned}
E[z_i] &= E_{z_i|x_i,\theta^{(t)}}(z_i) \\
&= \sum_z z_i P(z_i|x_i\theta^{(t)}) \\
&= 1P(z_i = 1|x_i\theta^{(t)}) + 0P(z_i = 0|x_i\theta^{(t)}) \\
&= 1P(z_i = 1|x_i\theta^{(t)}) \\
&= \frac{P(z_i, x_i|\theta^{(t)})}{P(x_i|\theta^{(t)})} \\
&= \frac{P(x_i|z_i\theta^{(t)})P(z_i|\theta^{(t)})}{\sum_{z_i} P(z_i, x_i|\theta^{(t)})} \\
&= \frac{\alpha N(x_i, \mu_1, \sigma_1)}{\alpha N(x_i, \mu_1, \sigma_1) + (1-\alpha)N(x_i, \mu_2, \sigma_2)} = w_i
\end{aligned}$$

The EM-Algorithm employs alternating the E-Step and the M-Step so that at each iteration we get closer to a solution of w_i and θ. We stop integrating when we see that the estimates of the parameters have converged.