Clustering

An example of Gaussian mixture model-based clustering

EM Clustering Example

Consider a one-dimensional clustering problem in which the data given are:

$x_1 = -4$
$x_2 = -3$
$x_3 = -1$
$x_4 = 3$
$x_5 = 5$

We will cluster these data into two clusters (K = 2)

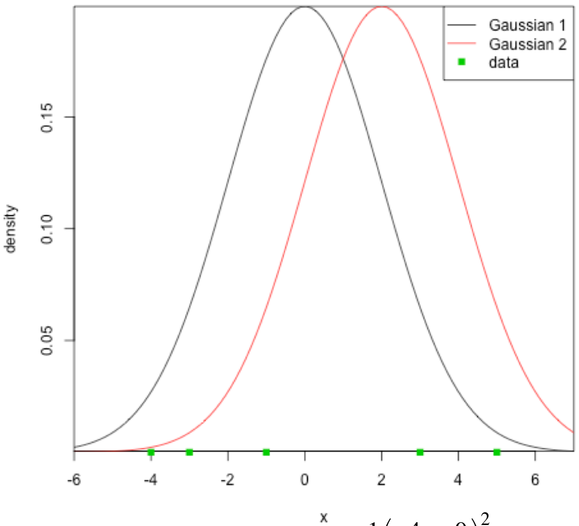
The initial mean of the first Gaussian is 0 and the initial mean of the second is 2. The Gaussians both have variance = 4; their density function is:

$$f(x \mid \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{2}\right)^2}$$

where μ denotes the mean (center) of the Gaussian.

Initially, we set $P_1 = P_2 = 0.5$

EM Clustering Example



$$f(x \mid \mu) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{2}\right)^2}$$

$$f(-4 \mid \mu_1) = \frac{1}{\sqrt{8\pi}} e^{-\frac{1}{2}\left(\frac{-4\ -\ 0}{2}\right)^2} = .0269$$

$$f(-4 \mid \mu_2) = .0022$$

$$f(-3 \mid \mu_1) = .0646$$

$$f(-3 \mid \mu_2) = .00874$$

$$f(-1 \mid \mu_1) = .176$$

$$f(-1 \mid \mu_2) = .0646$$

$$f(3 \mid \mu_1) = .0646$$

$$f(3 \mid \mu_2) = .176$$

$$f(5 \mid \mu_1) = .00874$$

$$f(5 \mid \mu_2) = .0646$$

EM Clustering Example: E Step

$$h_{11} = \frac{\frac{1}{2} f(x_1 | \mu_1)}{\frac{1}{2} f(x_1 | \mu_1) + \frac{1}{2} f(x_1 | \mu_2)} = \frac{.0269}{.0269 + .0022}$$
$$= 0.924$$

$$h_{12} = \frac{\frac{1}{2} f(x_1 | \mu_2)}{\frac{1}{2} f(x_1 | \mu_1) + \frac{1}{2} f(x_1 | \mu_2)} = \frac{.0022}{.0269 + .0022}$$
$$= 0.076$$

$$h_{21} = \frac{\frac{1}{2} f(x_2 | \mu_1)}{\frac{1}{2} f(x_2 | \mu_1) + \frac{1}{2} f(x_2 | \mu_2)} = \frac{.0646}{.0646 + .00874}$$
$$= 0.881$$

$$h_{22} = \frac{.00874}{.0646 + .00874} = 0.119$$

$$h_{32} = \frac{.0646}{.176 + .0646} = 0.268$$

$$h_{31} = \frac{.176}{.176 + .0646} = 0.732$$

$$h_{42} = \frac{.176}{.0646 + .176} = 0.732$$

$$h_{41} = \frac{.0646}{.0646 + .176} = 0.268$$

$$h_{52} = \frac{.0646}{.00874 + .0646} = 0.881$$

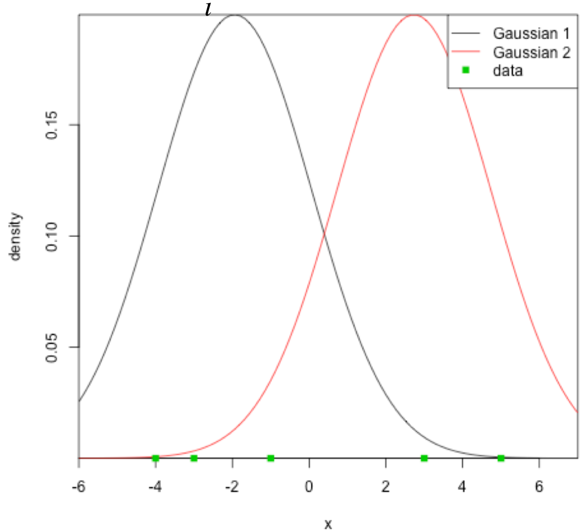
$$h_{51} = \frac{.00874}{.00874 + .0646} = 0.119$$

$$h_{ij} = E(Z_{ij} | \vec{x}_i) = \Pr(Z_{ij} = 1 | \vec{x}_i) = \frac{P_j N_j(\vec{x}_i)}{\sum_{l=1}^{K} P_l N_l(\vec{x}_i)}$$

EM Clustering Example: M-step

$$\mu_1 = \frac{\sum_i x_i \times h_{i1}}{\sum_i h_{i1}} = \frac{-4 \times .924 + -3 \times .881 + -1 \times .732 + 3 \times .268 + 5 \times .119}{.924 + .881 + .732 + .268 + .119} = -1.94$$

$$\mu_2 = \frac{\sum_i x_i \times h_{i2}}{\sum_i h_{i2}} = \frac{-4 \times .076 + -3 \times .119 + -1 \times .268 + 3 \times .732 + 5 \times .881}{.076 + .119 + .268 + .732 + .881} = 2.73$$

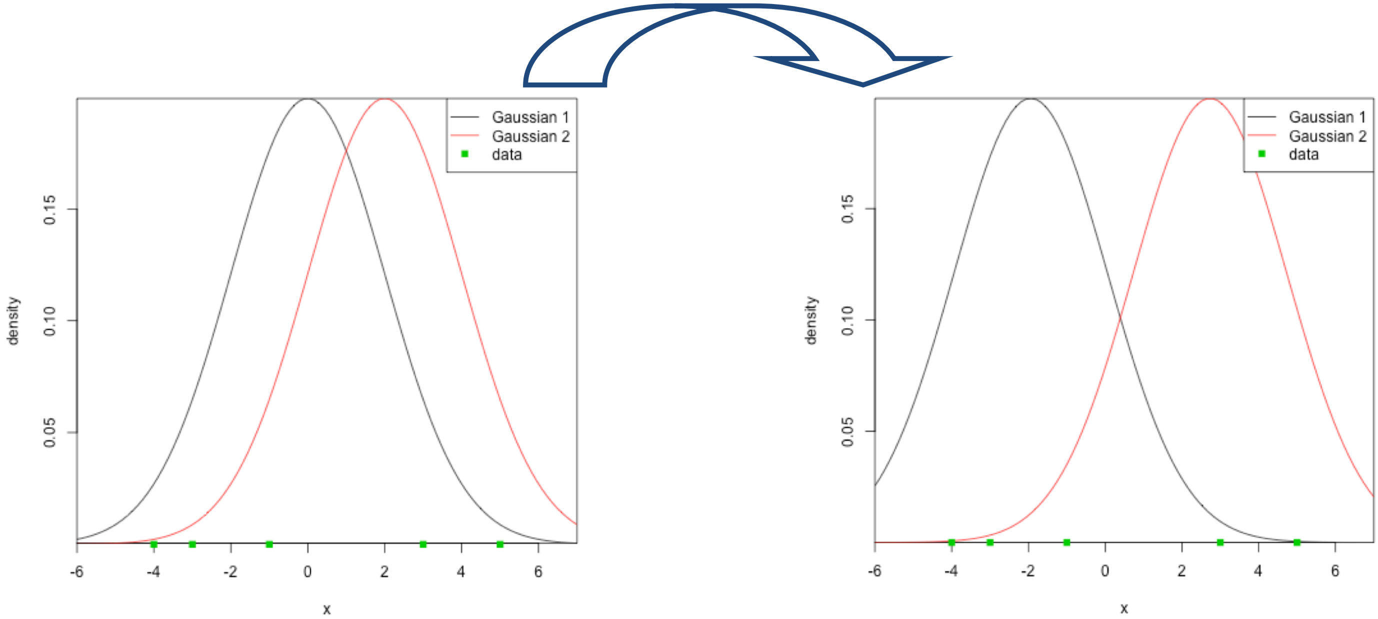


$$P_1 = \frac{\sum_i h_{i1}}{n} = \frac{.924 + .881 + .732 + .268 + .119}{5} = 0.58$$

$$P_2 = \frac{\sum_i h_{i2}}{n} = \frac{.076 + .119 + .268 + .732 + .881}{5} = 0.42$$

EM Clustering Example

- here we’ve shown just one step of the EM procedure



- we would continue the E- and M-steps until convergence