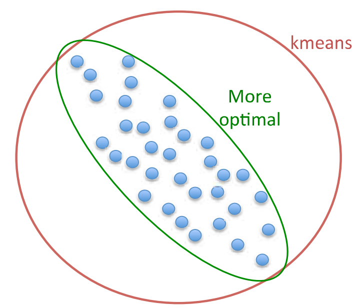
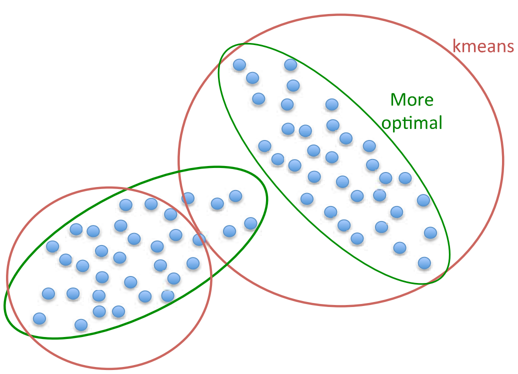
Expectation Maximization

Disadvantages of Kmeans

- Assumes spherical variance



Easy to see how kmeans could make a mistake here

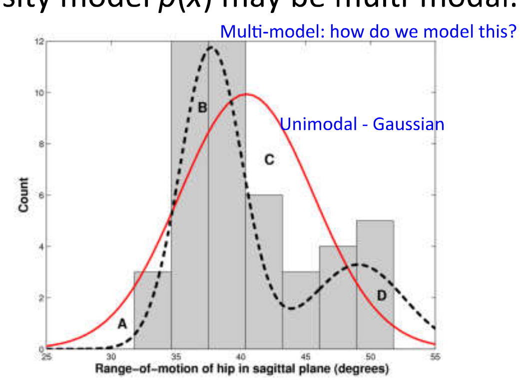


Soft Clustering

- K means does a hard assignment of points to clusters.
- Might also like to know the probability of belonging to a cluster
- Model each cluster with a probability distribution
  - Normal with params μ,Σ

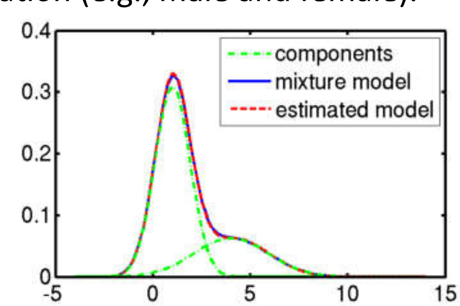
Mixture Model

- A density model $p(x)$ may be multi-modal.



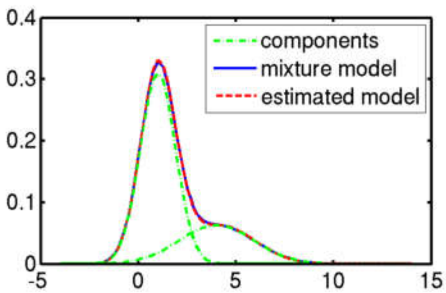
Mixture Model

- We may be able to model it as a mixture of uni-modal distributions (e.g., Gaussians).
- Each mode may correspond to a different sub-population (e.g., male and female).



Mixture Model

- We observe the mixture
- Can we recover the components?



The Model

- Probability of a point x given

$$\Theta = \{\theta_1 \dots \theta_K\}$$
$$\theta_j = \{\mu_j, \Sigma_j\}$$
$$p(x|\Theta) = \sum_{j=1}^{K} w_j p_j(x|\theta_j)$$

- w_j is probability any x belongs to cluster j
  - Note: does not depend on i

The Model

- Extend this to all points

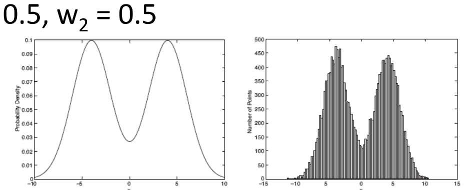
$$\begin{aligned} p(X|\Theta) &= \prod_{i=1}^{N} p(x_i|\Theta) \\ &= \prod_{i=1}^{N} \sum_{j=1}^{K} w_j \; p_j(x|\theta_j) \end{aligned}$$

Univariate Normal Case

$$\begin{aligned} p_j(x|\theta) &= p_j(x|\mu, \sigma) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \end{aligned}$$

Example Mixture

- Two univariate gaussians with
  - μ_1=4, μ_2=-4,
  - σ_1 = 2, σ_1= 2,
  - w_1 = 0.5, w_2 = 0.5



(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Figure 9.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

How to Estimate the Params?

- Calculate the MLE!
- Turns out to be the sample mean and sample standard deviation

$$\mu = \frac{1}{m} \sum_{i=1}^{m} x_i$$
$$\sigma = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2}$$

We're Missing Some Info

- Can't calculate the mean and std without knowing which m points belong to which cluster
- But we can't assign points to clusters without knowing the mean and std of the clusters
- EM handles this circularity

EM Algorithm

- Select initial set of parameters
  - i.e. Set μ and σ randomly, set all w = 1/K
- Repeat:
  - E-step: for each object, calculate the probability that it belongs to each distribution p(dist j | x, Θ)
  - M-step: given probs from e-step, calculate new estimates of params that maximize the expected likelihood
- Until the params don't change too much

E-step (example with K=2 clusters)

- Find probability for belonging to each cluster
  - e.g. with two clusters:

$$p(dist\ j|x_i,\theta) = \frac{w_j\ p(x_i|\theta_j)}{w_1\ p(x_i|\theta_1) + w_2\ p(x_i|\theta_2)}$$

- (by Bayes rule)

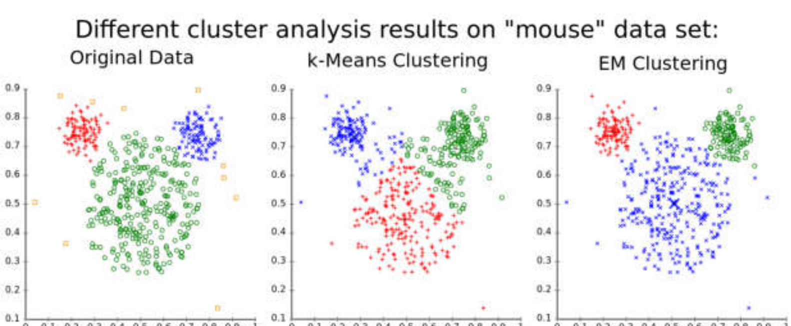
M-step

$$w_j = \frac{1}{N}\sum_{i=1}^{N} p(dist\ j|x_i,\theta)$$

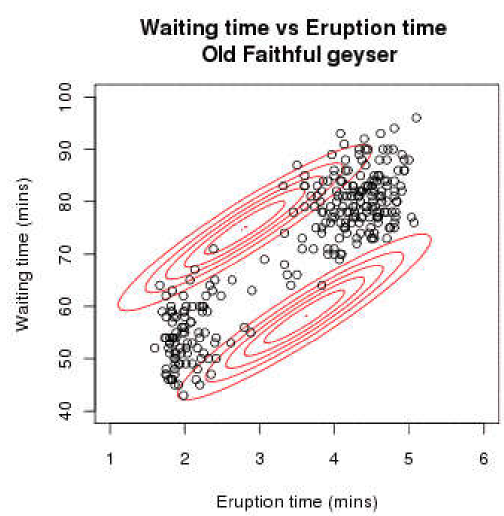
$$\mu_j = \frac{\sum_{i=1}^{N} p(dist\ j|x_i,\theta)x_i}{\sum_{i=1}^{N} p(dist\ j|x_i,\theta)}$$

$$\sigma_j = \frac{\sum_{i=1}^{N} p(dist\ j|x_i,\theta)(x_i-\mu_j)^2}{\sum_{i=1}^{N} p(dist\ j|x_i,\theta)}$$

K Means vs EM



from https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm



from https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Differences in Density

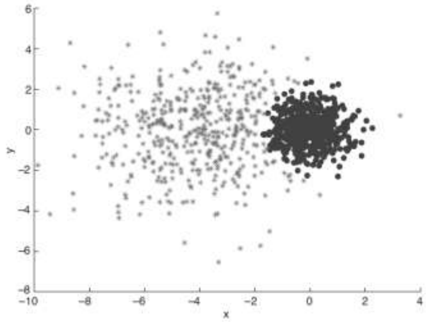


Figure 9.5. EM clustering of a two-dimensional point set with two clusters of differing density.

Non-spherical data

(a) Clusters produced by mixture model clustering.

(b) Clusters produced by K-means clustering.