- Mixture models with incomplete data
  - Expectation Maximization Algorithm
  - EM for Gaussian distribution
  - Application to images and disadvantages of the method

1 Mixture Models with incomplete data

In last lecture, we discussed how to estimate the parameters $\theta_k = (\pi_k, \phi_k), k = 1 \dots K$ of the probability distribution of the mixture model by assuming complete data knowledge, that is, we assumed we had measurements for both x and q. In this lecture, we will discuss how to estimate the parameters when we have incomplete data knowledge. We will concentrate in the case where q is not known.

1.1 Expectation Maximization Algorithm

In the complete data case, we computed the parameters of the model by maximizing the log-likelihood of the joint probability density of x and q. Now q is unobserved, thus we might try to estimate the marginal density of x

$$\sum_q P(x, q; \theta) \tag{1}$$

and then apply the maximum likelihood method. In order to do that, assume we have a guess θ' of the value of the parameters, and consider a series of measurements of x, $x_1, \dots, x_N$. Thus if we are able to construct a function $Q(\theta, \theta', (x_1, \dots x_N))$ that estimates the log-likelihood of x given the guess of the parameters and the measurements on x, then we can find a new estimate of the parameters that maximizes Q. This procedure can be repeated to find a sequence of better approximations of θ. Such an iterative algorithm is called Expectation Maximization (EM) and can be summarized as follows

E-step: Construct a function $Q(\theta, \theta', (x_1, \dots x_N))$, the so-called "Expected complete log-likelihood" based on the previous guess θ'

M-step: Find θ that maximizes Q

Iterate Repeat E and M steps until the difference between subsequent estimates is small

For the E-step, we can choose Q as the expected value of the log-likelihood of x given the measurements, that is:

$$
\begin{aligned}
Q(\theta, \theta', (x_1, \ldots, x_N)) &= E\left[\sum_{i=1}^{N}\sum_{k=1}^{K} \log\left(P(q_i^k, x_i; \theta)\right) \mid x = (x_1, \ldots, x_N)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{q} \log\left(P(q_i^k, x_i; \theta)\right) P(q_i^k | x_i; \theta') \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{q} q_i^k \log\left(\pi_k f_k(x_i, \phi_k)\right) P(q_i^k | x_i; \theta') \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} < q_i^k > \log\left(\pi_k f_k(x_i, \phi_k)\right)
\end{aligned}
$$

where

$$
\begin{aligned}
< q_i^k > &= \sum_{q} q_i^k P(q_i^k | x_i; \theta') \\
&= \frac{\sum_q P(x_i | q_i^k; \theta') P(q_i^k) q_i^k}{P(x_i; \theta')} \\
&= \frac{\pi'_k f_k(x_i; \phi'_k)}{\sum_{k=1}^{K} P(x_i | q_i^k; \theta') P(q_i^k)} \\
&= \frac{\pi'_k f_k(x_i; \phi'_k)}{\sum_{k=1}^{K} \pi'_k f_k(x_i; \phi'_k)}
\end{aligned}
$$

1.2 EM for Gaussian Distribution

If the distribution of the images x conditioned on q is Gaussian, that is:

$$
f_k = \frac{1}{\sqrt{2\pi}\,\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right), \tag{2}
$$

then we have to optimize

$$
\mathcal{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} < q_i^k > \log\left(\pi_k f_k(x_i, \phi_k)\right) + \lambda(1 - \sum_{k=1}^{K} \pi_k)
$$

where λ is the Lagrange multiplier.

Taking derivatives with respect to π_k and $\phi_k = (\mu_k, \sigma_k)$, and recalling that $\sum_k < q_i^k >= 1$, we get

$$
\begin{aligned}
\frac{\partial Q}{\partial \pi_k} &= \sum_{i=1}^{N} \frac{< q_i^k >}{\pi_k} - \lambda = 0 && \Rightarrow \widehat{\pi_k} = \frac{\sum_{i=1}^{N} < q_i^k >}{N} \\
\frac{\partial Q}{\partial \mu_k} &= \sum_{i=1}^{N} \frac{< q_i^k > (x_i - \mu_k)}{\sigma_k^2} = 0 && \Rightarrow \widehat{\mu_k} = \frac{\sum_{i=1}^{N} < q_i^k > x_i}{\sum_{i=1}^{N} < q_i^k >} \\
\frac{\partial Q}{\partial \sigma_k} &= \sum_{i=1}^{N} \frac{< q_i^k > [(x_i - \mu_k)^2 - \sigma_k^2]}{\sigma_k^3} = 0 && \Rightarrow \widehat{\sigma_k}^2 = \frac{\sum_{i=1}^{N} < q_i^k > (x_i - \widehat{\mu_k})^2}{\sum_{i=1}^{N} < q_i^k >}
\end{aligned}
$$

This formulae give us new estimation for θ given the current estimation θ' and the measurements $(x_1, \dots x_N)$. Recall that $< q_i^k >$ is a function of θ' and $(x_1, \dots x_N)$.

More generally, we might assume that the distribution is a Multivariate Gaussian. In this case the joint density is given by:

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)}{2}\right),$$

where $\mathbf{x}$ is a vector in R^d, $\mu \in R^d$ is the mean, and $\Sigma \in R^{d \times d}$ is the covariance matrix.

1.3 Application to images

Based on the EM philosophy, the following scheme (see Figure 1) can be used for image segmentation

- Assume we know to which class each pixel belongs to (pixel classes) and find class characteristics
- Assume we know class characteristics and find pixel classes
- Iterate

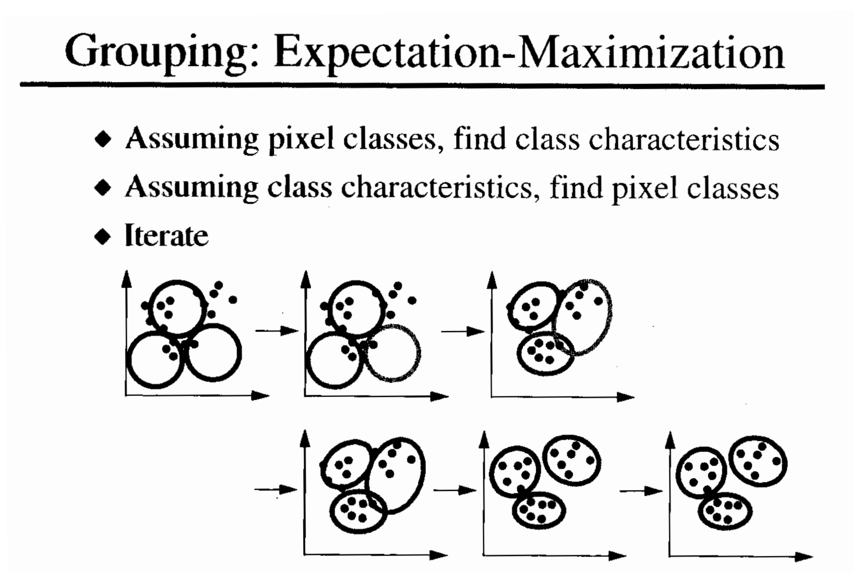

Figure 1: Grouping: Expectation-Maximization

Figure 2 shows an example of the application of EM to image segmentation

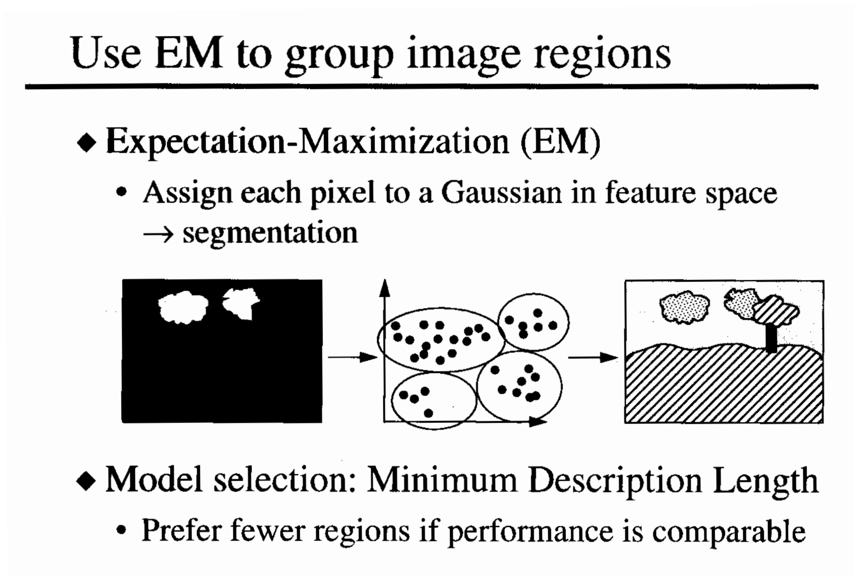

Figure 2: Using EM to group image regions

1.4 Disadvantages of the method

- Can converge to a local minima
- Sensitive to initialization
- We do not know how to choose K.