EECS 545: Machine Learning

Lecture 16. Learning in Graphical Models

Honglak Lee

3/9/2011

Outline

- Maximum Likelihood parameter estimation
- Expectation maximization

Overview: Graphical Models

- Representation
  - Which joint probability distributions does a graphical model represent?
  - Directed and Undirected Graphical models
  - Conditional Independence
- Inference
  - How to answer questions about the joint probability distribution?
  - Marginal distribution of a node variable (or subset of nodes)
  - Most likely assignment of node variables
  - Sum-product algorithm
- Learning (today’s lecture)
  - How to learn the parameters and structure of a graphical model?

Learning

- Learn parameters or structure from data
- Parameter learning: find maximum likelihood estimates of parameters
- Structure learning: find correct connectivity between existing nodes

Overview: Learning Graphical Models

Structure	Observation	Method
Known	Full	Maximum Likelihood (ML) estimation
Known	Partial	Expectation Maximization algorithm (EM)
Unknown	Full	Model selection
Unknown	Partial	EM + model selection

Covered today

Maximum Likelihood
(for Bayes Nets)

Example: Coin Toss

- We have a coin, with a probability of head p_H
- Suppose that we have tossed the coin 5 times, and got 3 heads and 2 tails. What is the most likely value of $p_{H?}$

Example: Coin Toss

- We have a coin, with a probability of head p_H
- Suppose that we have tossed the coin 5 times, and got 3 heads and 2 tails. What is the most likely value of $p_{H?}$
- Answer: 3/(3+2) = 0.6
- In fact, this is maximum likelihood estimation!

$$\mathrm{P}(D) = p_H^3(1 - P_H)^2$$
$$\log P(D) = 3\log p_H + 2\log(1 - p_H)$$

Taking partial derivative:

$$\frac{\partial \log P(D)}{\partial p_H} = \frac{3}{p_H} - \frac{2}{1 - p_H} = 0$$

We get p_H = 0.6!

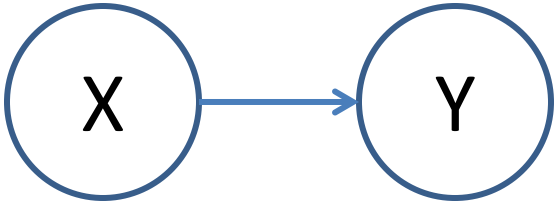
Example: Coin Toss

- Generalization: For a Bernoulli IID random variable (coin) with probability θ, and given H 1's (heads) and T 0's (tails), the maximum likelihood estimate is:

$$\theta_{ML} = \frac{H}{H + T}$$

Two variable case in BN

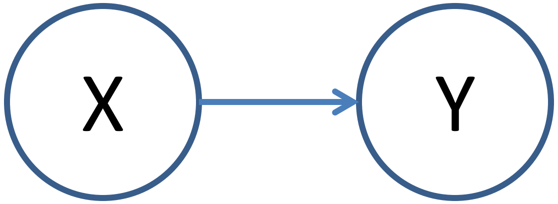
- Given a Bayes Net: X-> Y
  - X, Y are both binary
- What are the parameters of the model?



$P(X=1)$	$P(X=0)=1-P(X=1)$
$P(Y=1 \mid X=1)$	$P(Y=0 \mid X=1)$
$P(Y=1 \mid X=0)$	$P(Y=0 \mid X=0)$

Two variable case in BN

- Given a Bayes Net: X-> Y
  - X, Y are both binary



- What are the parameters?

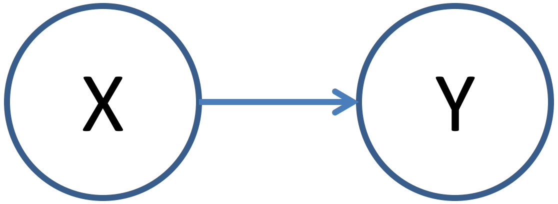
$$\theta_X = P(X = 1)$$
$$\theta_{Y|X=0} = P(Y = 1|X = 0)$$
$$\theta_{Y|X=1} = P(Y = 1|X = 1)$$

- What is the maximum likelihood?

$$\left\{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)}) \right\}$$

Two variable case in BN

- Given a Bayes Net: X-> Y
  - X, Y are both binary
- What are the parameters?



$$\theta_X = P(X=1)$$
$$\theta_{Y|X=0} = P(Y=1|X=0)$$
$$\theta_{Y|X=1} = P(Y=1|X=1)$$

$x^{(i)} = 1, \quad y^{(i)} = 0$

$P(x=1)\,P(Y=0|X=1)$

- What is the maximum likelihood?

$$P(D) = \prod_{i}^{N} P(x^{(i)}, y^{(i)}; \theta)$$

$$P(x^{(i)}, y^{(i)}; \theta) = \theta_X^{\overset{1}{I[x^{(i)}=1]}} \cancel{(1-\theta_X)^{I[x^{(i)}=0]}}$$
$$\cancel{\theta_{Y|X=1}^{I[x^{(i)}=1, y^{(i)}=1]}} (1-\theta_{Y|X=1})^{\overset{1}{\boxed{I[x^{(i)}=1, y^{(i)}=0]}}} = \theta_X(1-\theta_{Y|X=1})$$
$$\cancel{\theta_{Y|X=0}^{I[x^{(i)}=0, y^{(i)}=1]} (1-\theta_{Y|X=0})^{I[x^{(i)}=0, y^{(i)}=0]}}$$

- Overall:

$$P(D) = \theta_X^{Counts[x^{(i)}=1]} (1-\theta_X)^{Counts[x^{(i)}=0]}$$
$$\theta_{Y|X=1}^{Counts[x^{(i)}=1, y^{(i)}=1]} (1-\theta_{Y|X=1})^{Counts[x^{(i)}=1, y^{(i)}=0]}$$
$$\theta_{Y|X=0}^{Counts[x^{(i)}=0, y^{(i)}=1]} (1-\theta_{Y|X=0})^{Counts[x^{(i)}=0, y^{(i)}=0]}$$

Two variable case in BN

- Taking derivatives with respect to the parameters and setting it to zero, we have:

$$\hat{P}^{ML}(x=1) = \theta_X^{ML} = \frac{Counts[X=1]}{Counts[X=1]+Counts[X=0]} = \frac{Counts[X=1]}{Total\ counts}$$

$$\hat{P}^{ML}(y=1|x=1) = \theta_{Y|X=1} = \frac{Counts[X=1,Y=1]}{Counts[X=1]}$$

$$\theta_{Y|X=0} = \frac{Counts[X=0,Y=1]}{Counts[X=0]}$$

Q. Verify this (or earlier cases)

MLE in Bayesian Nets

- The likelihood term decomposes with respect to local CPTs

$P(X_i | PaX_i)$

- Overall, the MLE parameter estimation will be

$$\theta_{X_i=val|PaX_i=valPa} = \frac{Counts[X_i = val|PaX_i = valPa] + \alpha'}{Counts[PaX_i = valPa] + \alpha}$$

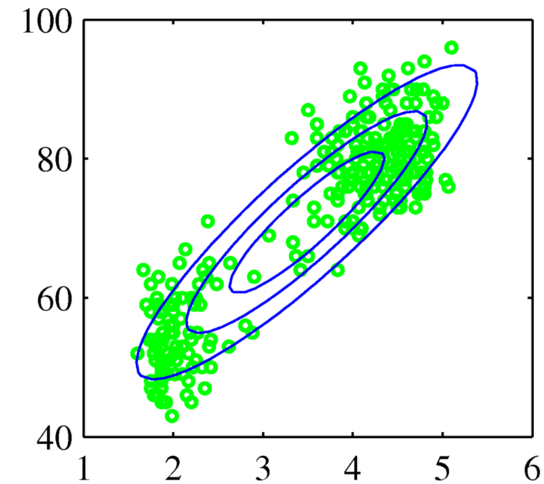
Expectation Maximization

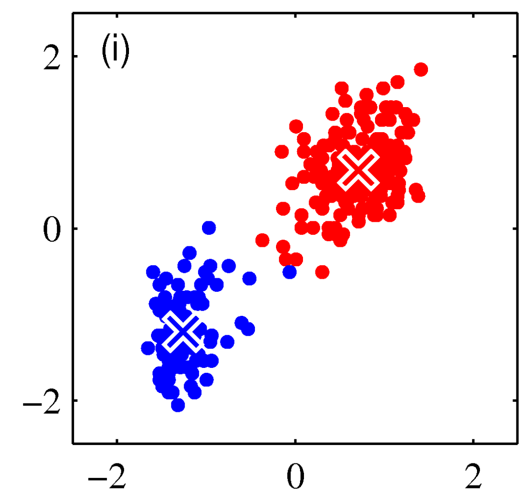
Expectation Maximization

- Parameter learning when the data is not fully observed.
  - Suppose that we have observed varaibles X, and hidden variables Z
- Main idea:
  - Run inference about Z given X: $Q=P(Z|X)$
  - Update parameters by treating Q as observation!
- Example:
  - Gaussian mixtures
  - (We will start with Kmeans which is a special case of Gaussian mixtures)

The K-Means Algorithm

- Given unlabeled data x_n, $(n=1,\ldots,N)$,
- And believing it belongs in *K* clusters,
- How do we find the clusters?





The K-Means Algorithm

- We need indicator variables r_{nk} in {0,1}.
  - $\boldsymbol{r_{nk}}$ = 1 if $\mathbf{x}_n$ is in cluster k.
  - and $\boldsymbol{r_{nj}}$ = 0 for all j other than k.

$$r \quad \begin{array}{c|c} & 1 \cdots K \\ \hline 1 & 0001000 \\ \vdots & 0100000 \\ \vdots & 0000010 \\ N & \end{array}$$

- Minimize the distortion measure J: sum of squared distance of points from the center of its own cluster.

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \mu_k||^2$$

The K-Means Algorithm

- Set the cluster centers arbitrarily.
- Repeat until quiescence:
  - **E Step: assign each point to closest center.**

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j ||\mathbf{x}_n - \mu_j||^2 \\ 0 & \text{otherwise} \end{cases}$$

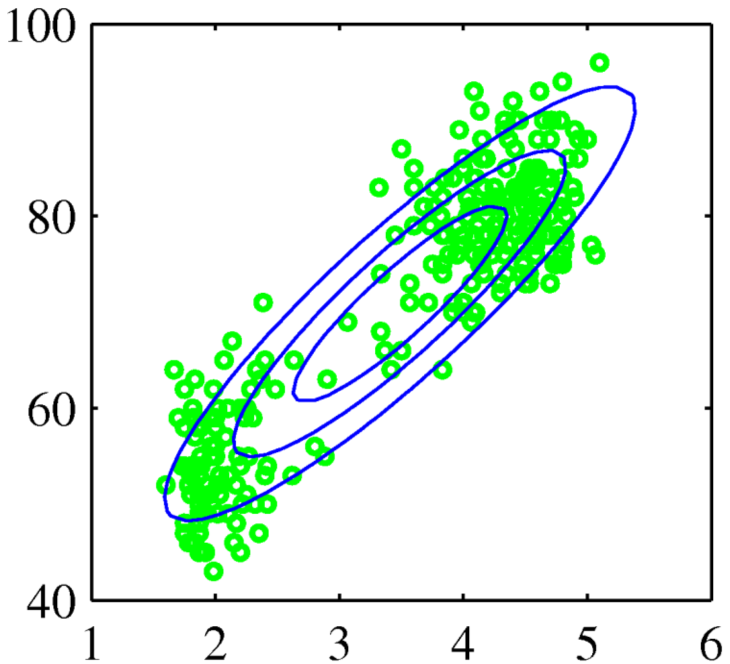
  - **M Step: update the centers**

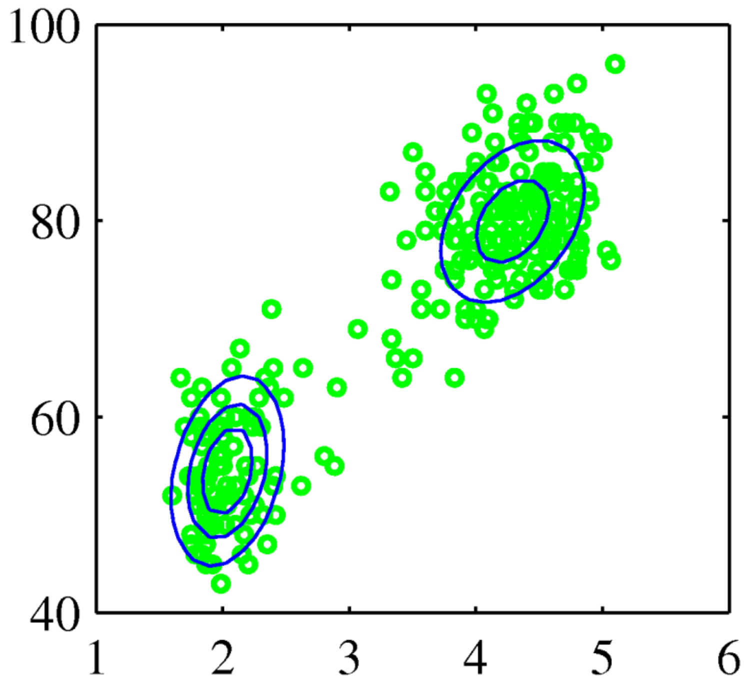
$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

Q. Verify this

Clustering Pixels

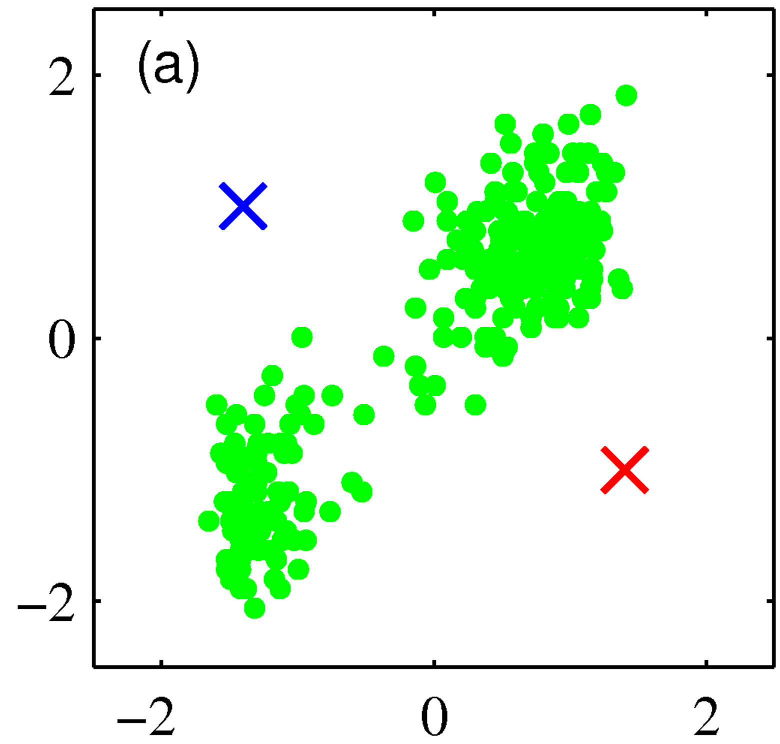
- How do we find clusters of pixels?





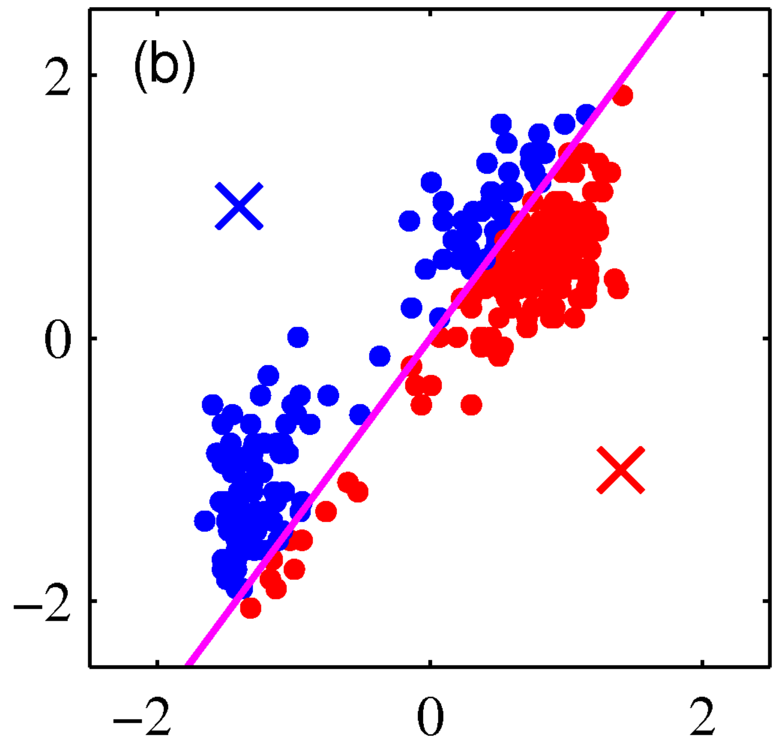
K-Means Clustering

- Select K. Pick random means.
  – Here K=2.



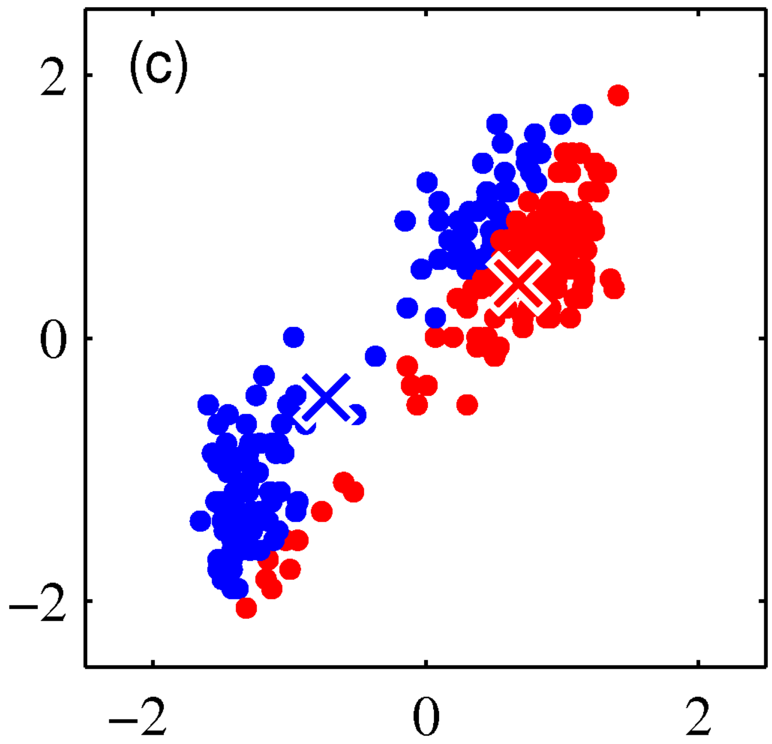
The E Step

- Assign each point to the nearest center.



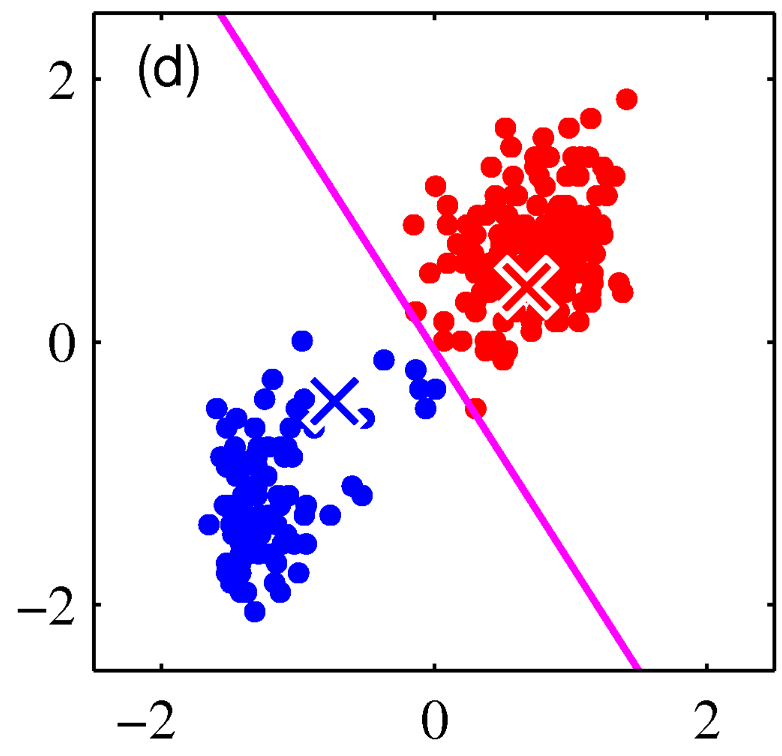
The M-Step

- Compute new centers for each cluster.



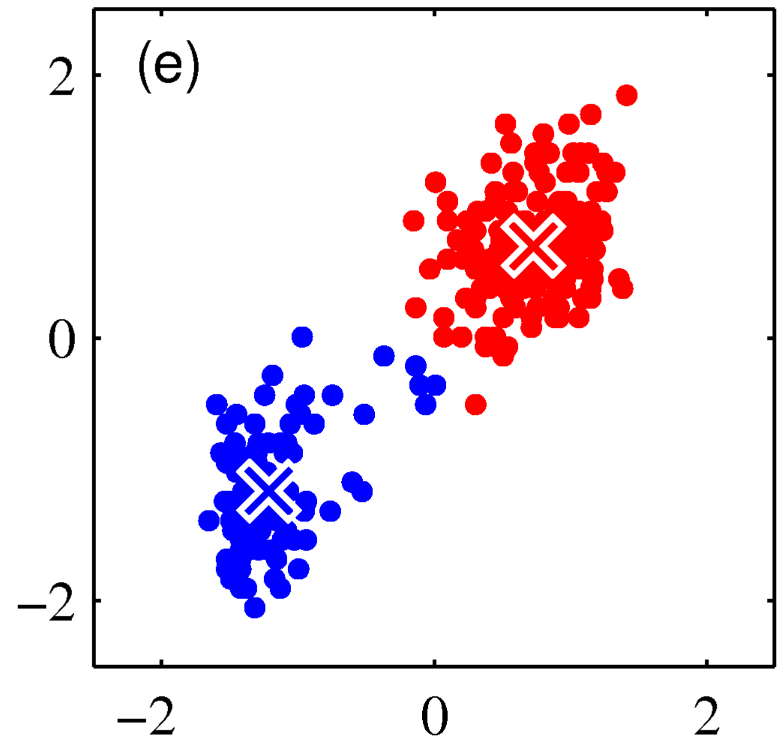
The E-Step Again

- Re-assign points to the now-nearest center.



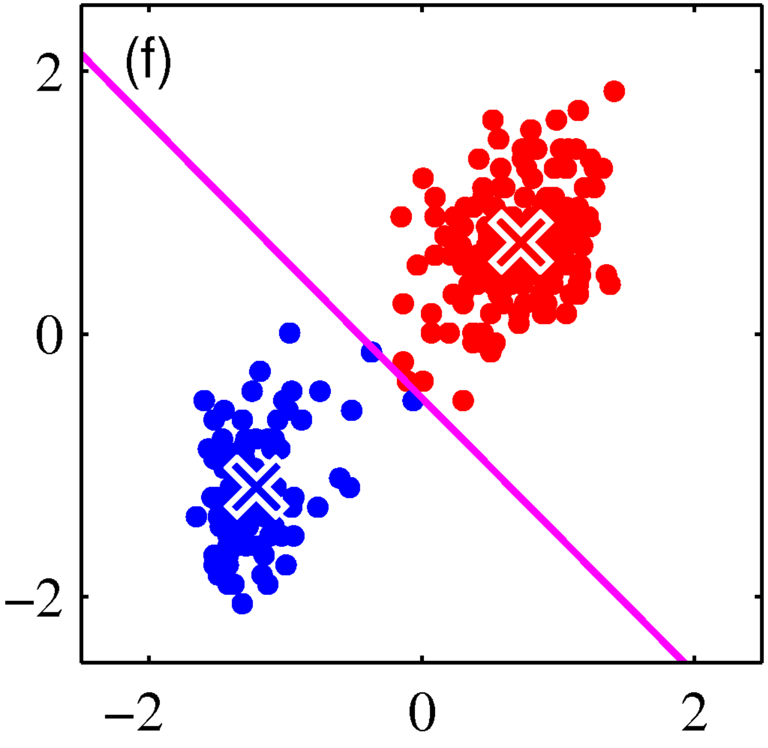
The M-Step Again

- Compute centers for the new clusters.



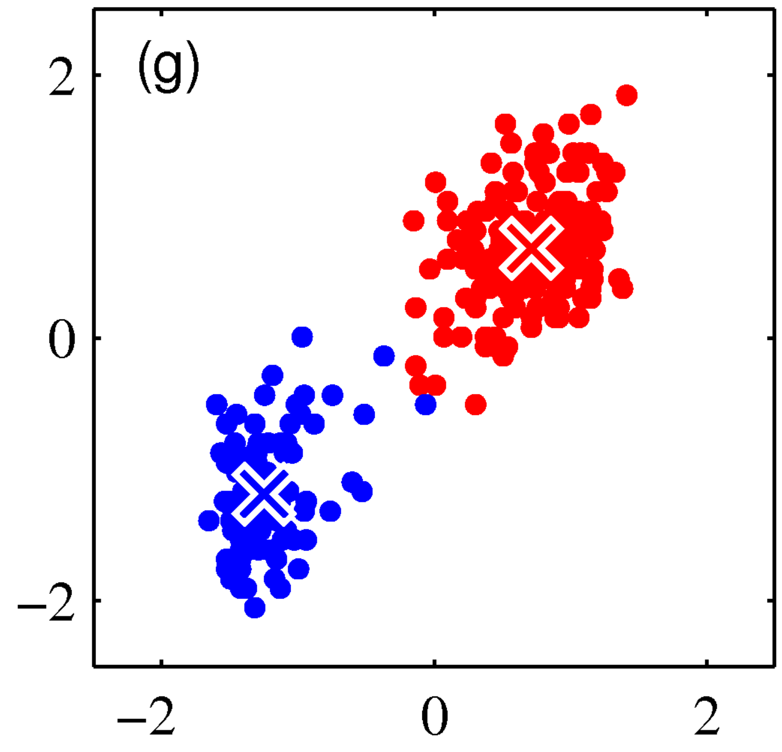
Another E-Step

- Reassign the pixels to centers.



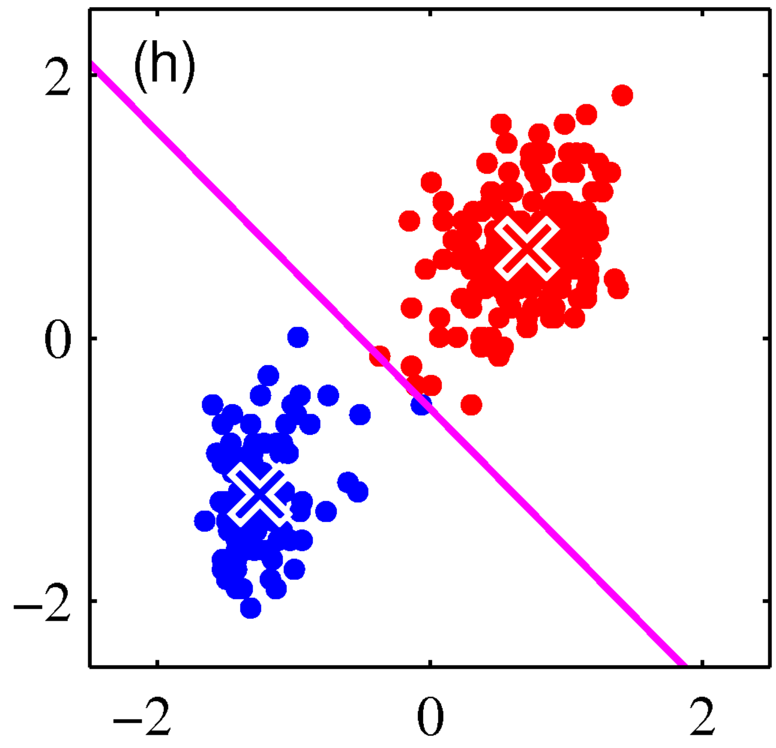
Another M-Step

- New centers.



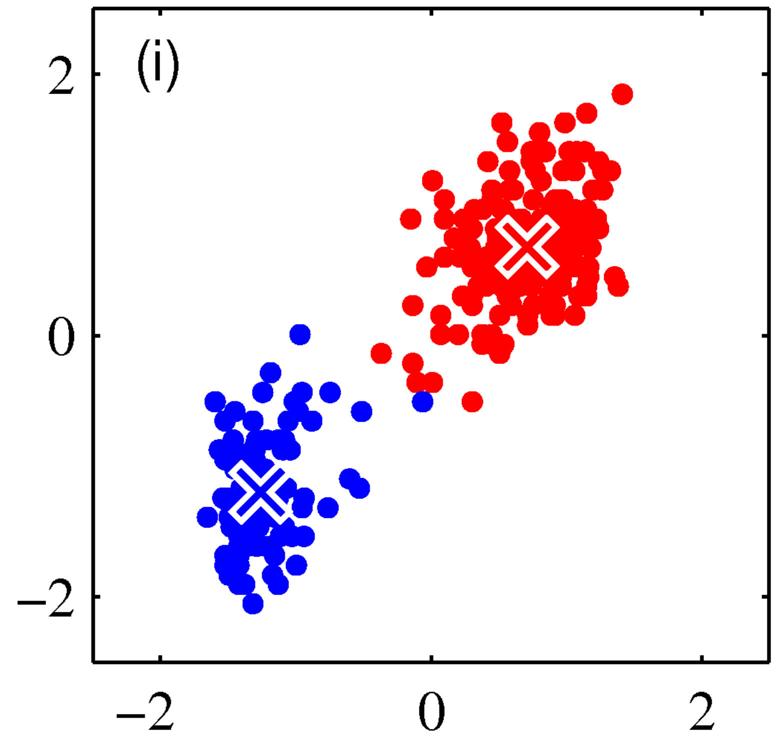
Another E-Step.

- New cluster assignments.



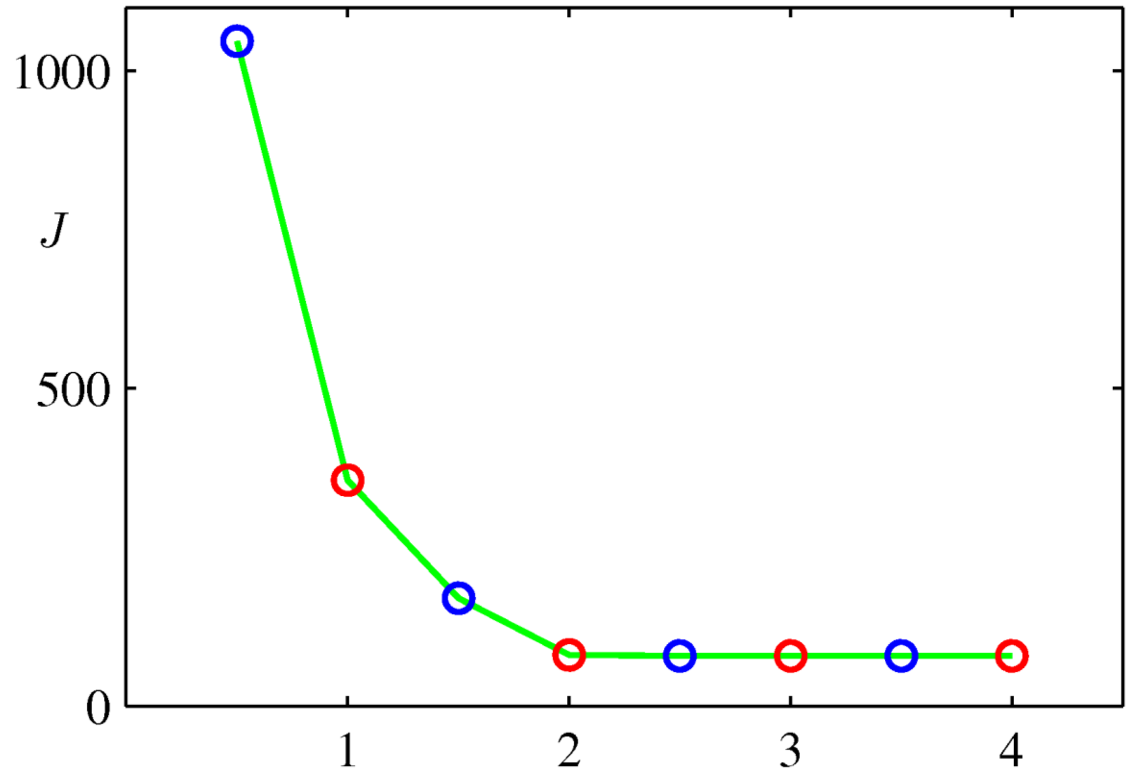
M-Step again.

- The cluster centers have stopped changing.



Convergence

- Convergence is relatively quick, in steps.
  - But: all those distance computations are expensive.



Hard and Soft Clusters

- K-Means uses hard clustering.
  - A point belongs to exactly one cluster.

- Mixture of Gaussians uses soft clustering.
  - A point could be explained by any cluster.
  - Different clusters take different levels of responsibility for that point.
  - (It was actually generated by only one cluster, but we don’t know which one.)

Expectation Maximization

- Parameter learning when the data is not fully observed.
  – Suppose that we have observed varaibles X, and hidden variables Z
- Main idea:
  – E-step: Run inference about Z given X: $Q=P(Z|X)$
  – M-step: Update parameters by treating Q as observation!
- Example:
  – Gaussian mixtures
  – (We will start with Kmeans which is a special case of Gaussian mixtures)

One page-derivation of EM

- Given the observed input data x, latent variable z, and parameter θ:

$$
\begin{aligned}
\log P_\theta(x) &= \log \sum_z P_\theta(x,z) \\
&= \log \sum_z Q(z) \frac{P_\theta(x,z)}{Q(z)} \quad (Set\, Q(z) \geq 0, \sum_z Q(z) = 1) \\
&\geq \sum_z Q(z) \log \frac{P_\theta(x,z)}{Q(z)} \quad (Jensen's\, inequality)
\end{aligned}
$$

One page-derivation of EM

- Given the observed input data x, latent variable z, and parameter θ:

$$\begin{aligned}\log P_\theta(x) &= \log \sum_z P_\theta(x,z) \\ &= \log \sum_z Q(z) \frac{P_\theta(x,z)}{Q(z)} \quad (Set\, Q(z) \geq 0, \sum_z Q(z) = 1) \\ &\geq \sum_z Q(z) \log \frac{P_\theta(x,z)}{Q(z)} \quad (Jensen's\, inequality)\end{aligned}$$

- Equality holds when $\mathrm{Q(z)} \propto P_\theta(x,z) = P_\theta(z|x)$
  - (E-step) Compute the posterior of z given x
- Fix Q, update θ that maximize the “data completion” log-likelihood (M-step)

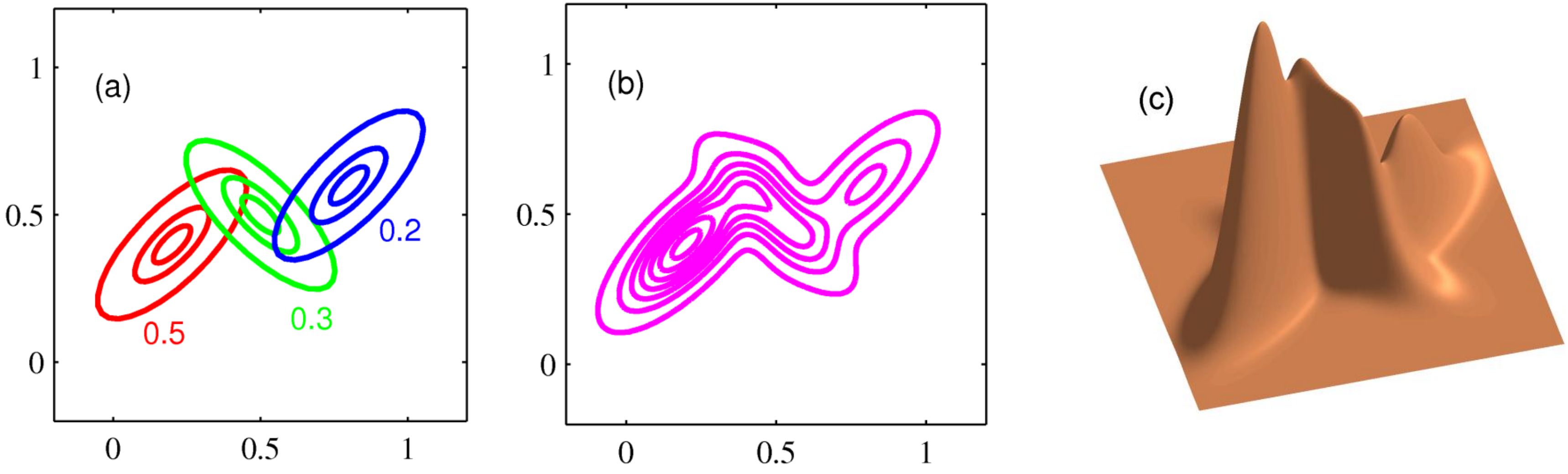
$$\sum_i \sum_{z^{(i)}} Q\left(z^{(i)}\right) \log P_\theta(x^{(i)}, z^{(i)})$$

Q. Verify this!

Mixtures of Gaussians

- Mixtures of Gaussians make it possible to describe much richer distributions.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$



Mixtures of Gaussians

- Note the mixing coefficients in

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \qquad \sum_{k=1}^{K} \pi_k = 1$$

- Let z in $\{0,1\}^K$ be a 1-of-*K* random variable;

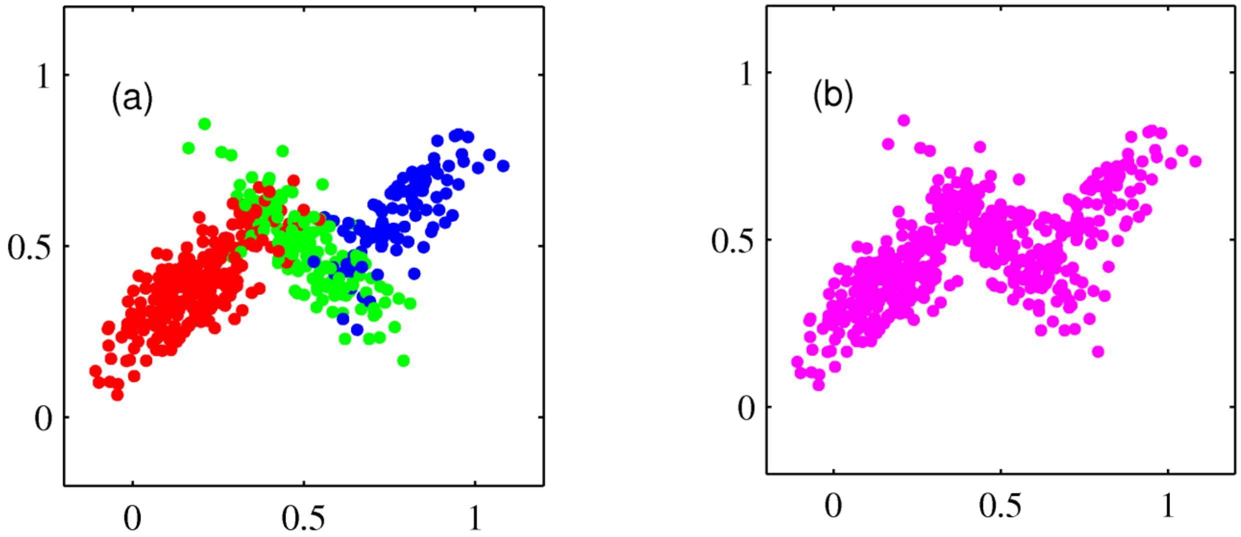
$$p(z_k = 1) = \pi_k \qquad p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)$$

Mixtures of Gaussians

- To generate samples from a Gaussian mixture distribution *p*(x), use *p*(x,z):
  - Select a value **z** from the marginal *p*(**z**);
  - Then select a value **x** from *p*(**x** | **z**) for that **z**.

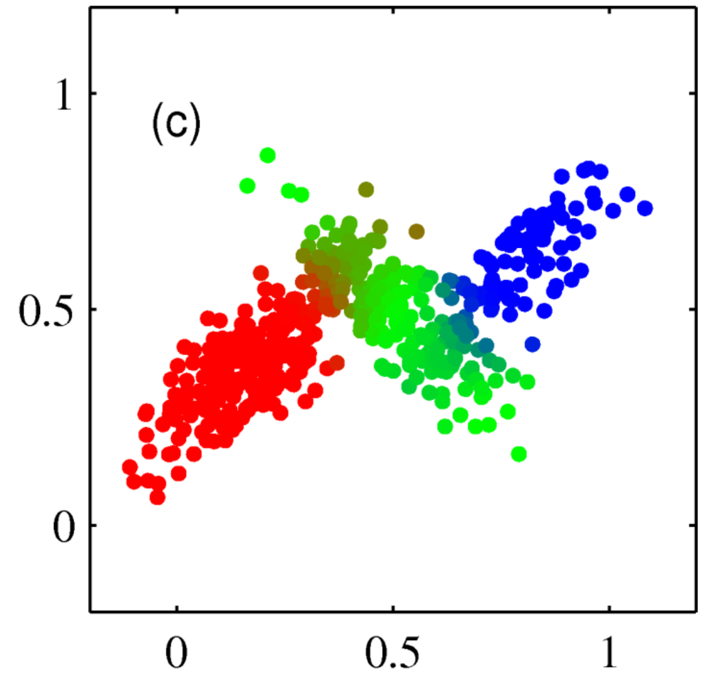


Mixtures of Gaussians

- Responsibility is the degree to which each Gaussian explains an observation x.

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x})$$

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\mu_j, \mathbf{\Sigma}_j)}$$



Q. Verify this!

Mixtures of Gaussians

- The mean of a cluster is the weighted mean, weighted by the responsibilities.

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

  – N_k is the effective number of points in cluster k

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \qquad \pi_k = \frac{N_k}{N}$$

- Likewise for covariance:

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

EM for Gaussian Mixtures

- Initialize means, covariances, and mixing coefficients for the K Gaussians.

- E Step: Given the coefficients, evaluate the responsibilities.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \mathbf{\Sigma}_j)}$$

EM for Gaussian Mixtures

- M Step: Given the responsibilities, re-evaluate the coefficients.

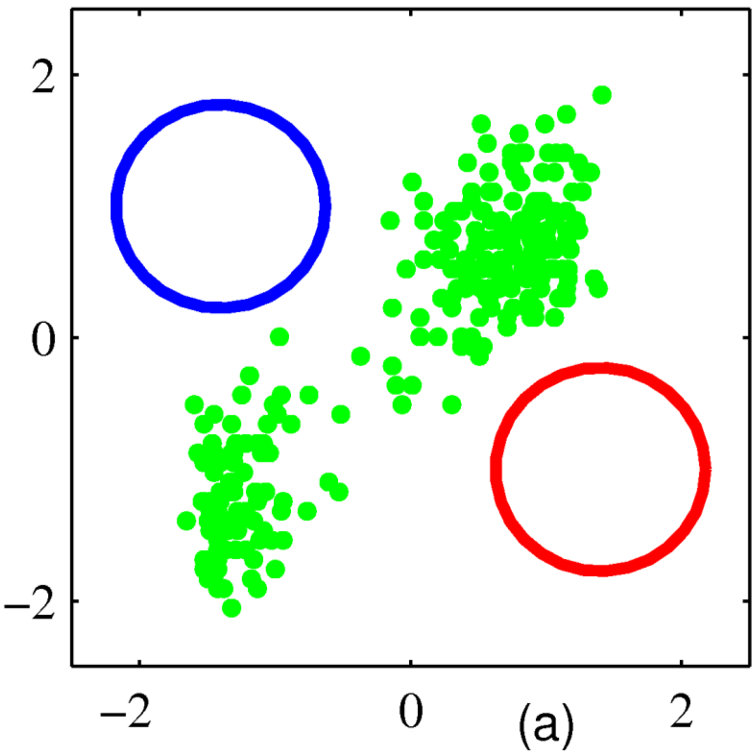
$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \pi_k^{\text{new}} = \frac{N_k}{N}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^T$$

- Stop when either coefficients or log likelihood converges.

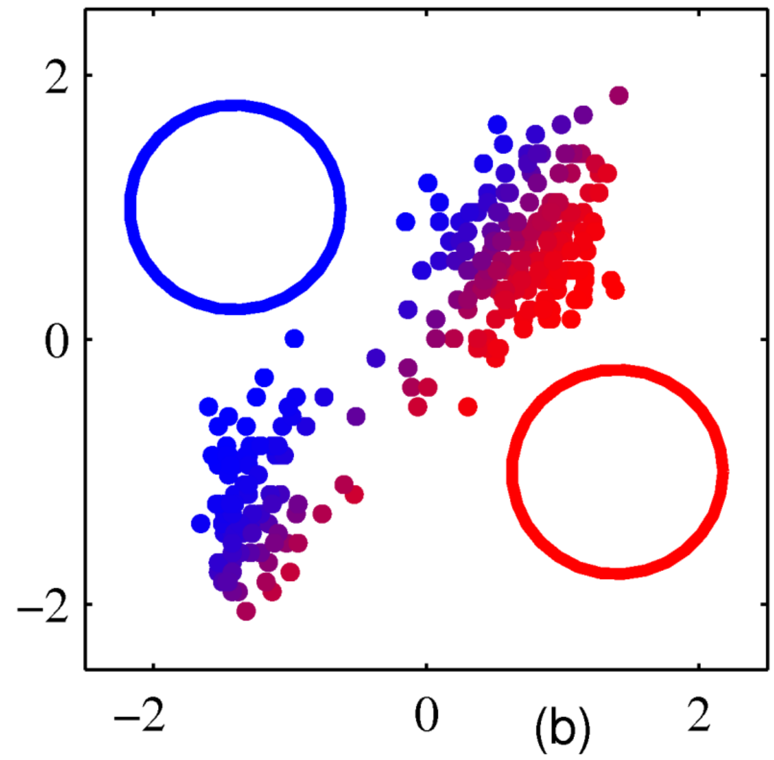
EM Example

- Initialize parameters: means, covariances, and mixing coefficients.



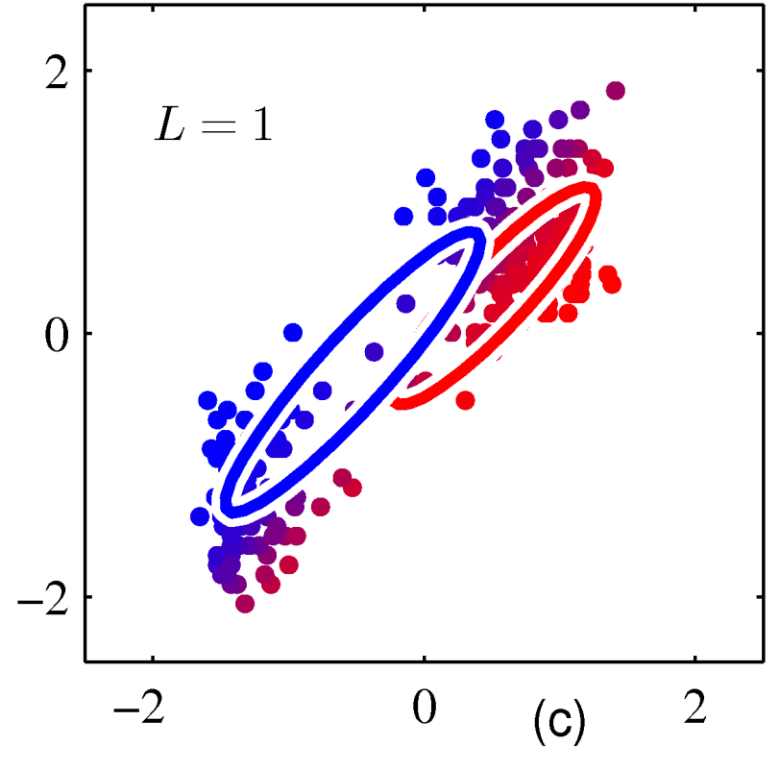
EM Example

- First E Step



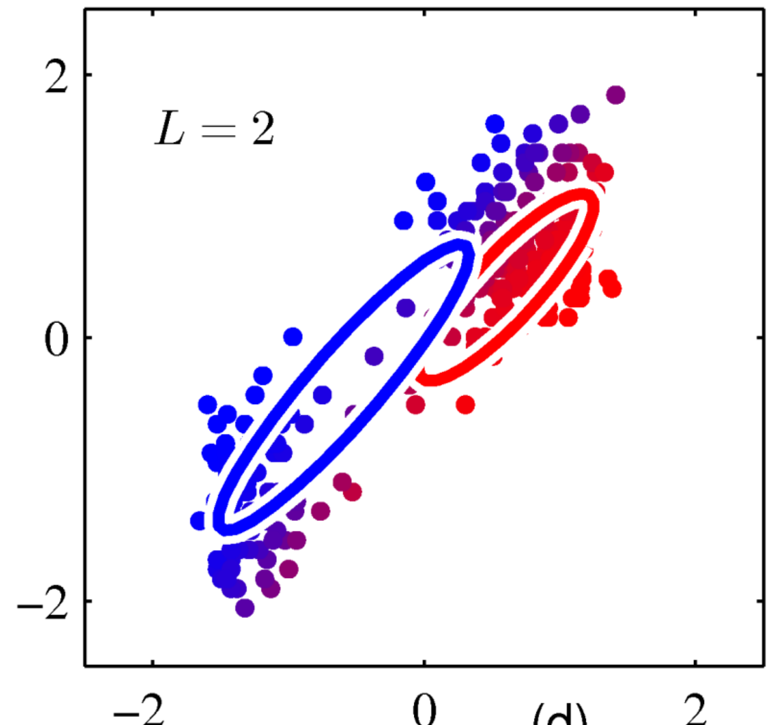
EM Example

- First M Step



EM Example

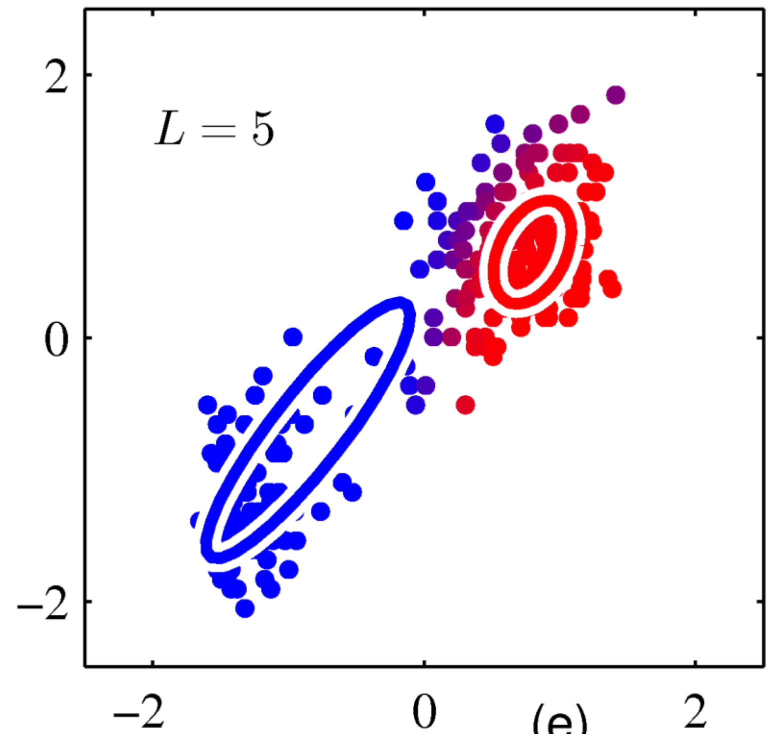
- Second E and M Steps



(d)

EM Example

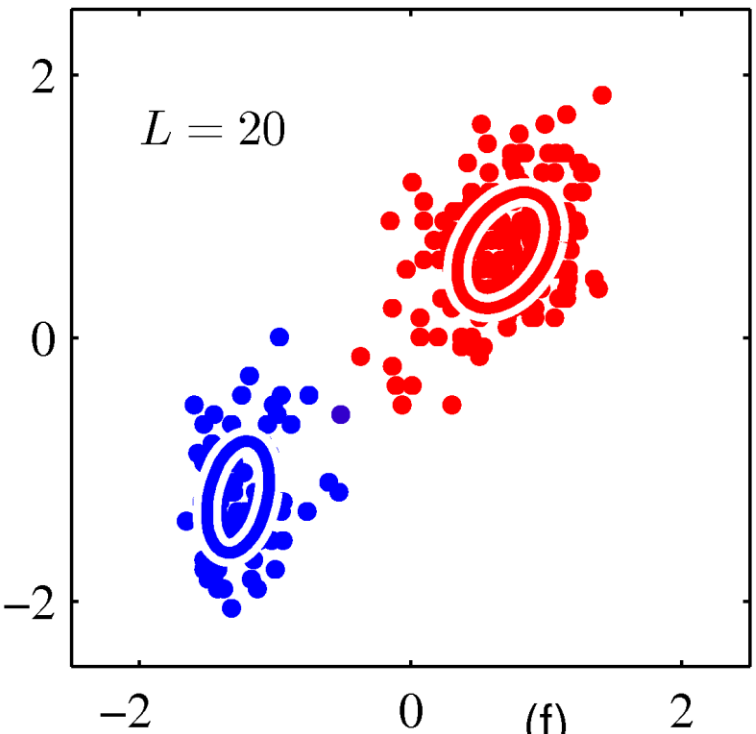
- Three more E-M cycles



(e)

EM Example

- Fifteen E-M cycles later



(f)

Abstract view of EM

Latent Variables

- A system with observed variables X,
  - may be far easier to understand in terms of additional variables **Z**,
  - but they are not observed (latent).

- For example, in a mixture of Gaussians,
  - The latent variable **z** specifies which Gaussian generated the sample **x**.
  - The *responsibility* is essentially p(**z** | **x**).

Latent Variables

- We find model parameters by maximizing log likelihood of observed data.
- If we had complete data {X,Z}, we could easily maximize likelihood $p(\mathbf{X}, \mathbf{Z}|\theta)$
- Unfortunately, with incomplete data (X only), we must marginalize over Z, so

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}$$

- (The sum inside the log makes it hard.)

Expectation, then Maximization

- E-Step:
  - Given current parameter values, find the *distribution* $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$
  - This lets us define the *expectation*

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

- M-Step:
  - Maximize the expectation of log likelihood, over the distribution $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$

$$\theta^{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{\text{old}})$$

The E-M Algorithm

- Choose initial values for the parameters.
- Repeat:
  – **E-Step:** $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$
  – **M-Step**

$$\theta^{\text{new}} = \arg\max_{\theta} \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$$

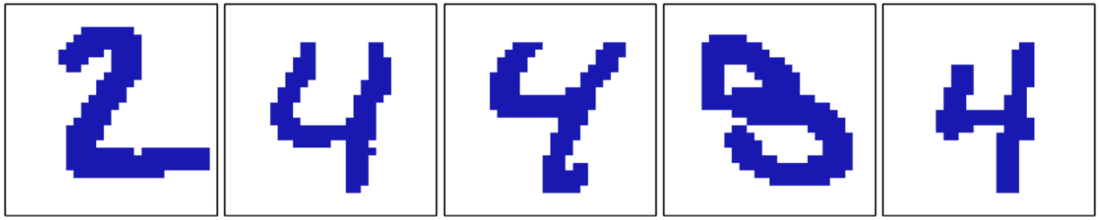
- Until convergence
  – of parameters or log likelihood

K-Means and E-M

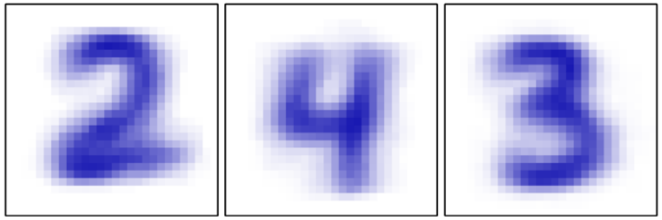
- Consider E-M over Gaussian models with fixed covariance matrix $\epsilon\mathbf{I}$
- In the limit as $\epsilon \rightarrow 0$ the responsibility goes to 1 for the closest Gaussian, and 0 elsewhere.
- This gives hard assignment to clusters, and the K-Means algorithm.

More Clustering

- These images are points in $\{0,1\}^D$.



- We find three clusters:



- The clusters are (very large) mixtures of Bernoulli distributions. These images show the latent responsibilities.

The EM Algorithm in General

- Our goal is to maximize $p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)$

- For any distribution *q*(Z) over latent variables

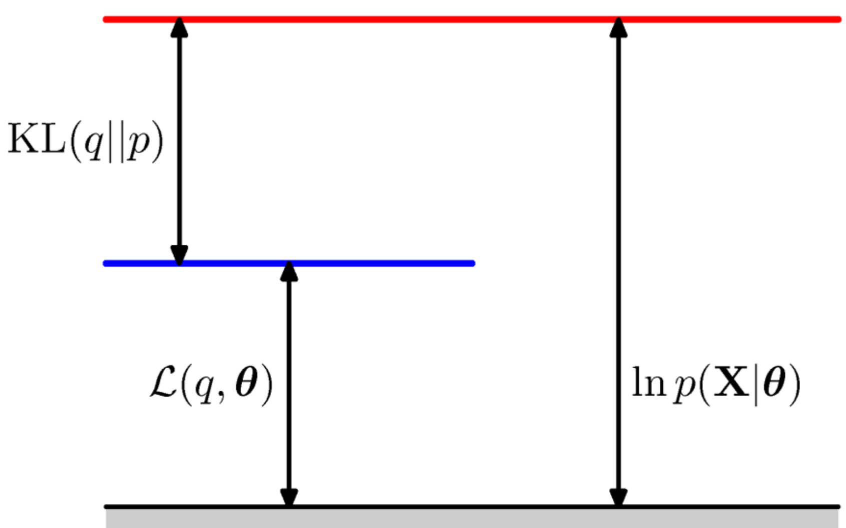
$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + KL(q||p)$$

- where

$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \right\}$$

$$KL(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

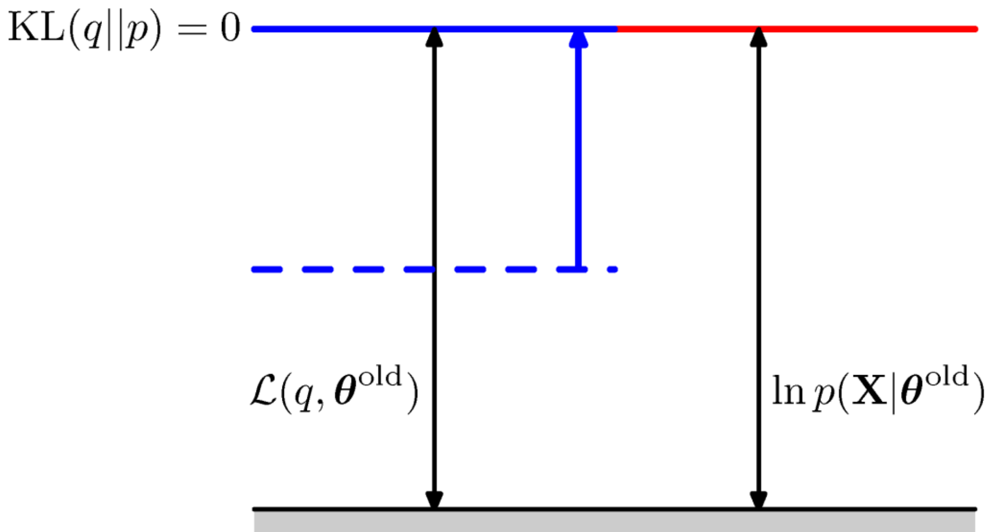
Visualize the Decomposition



$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + KL(q||p)$$

- Recall: $KL(q||p) \geq 0$
  - with equality only when *q*=*p*.
- Thus, $\mathcal{L}(q, \theta)$
  - is a lower bound on $\ln p(\mathbf{X}|\theta)$
- which EM tries to maximize.

Visualize the E-Step

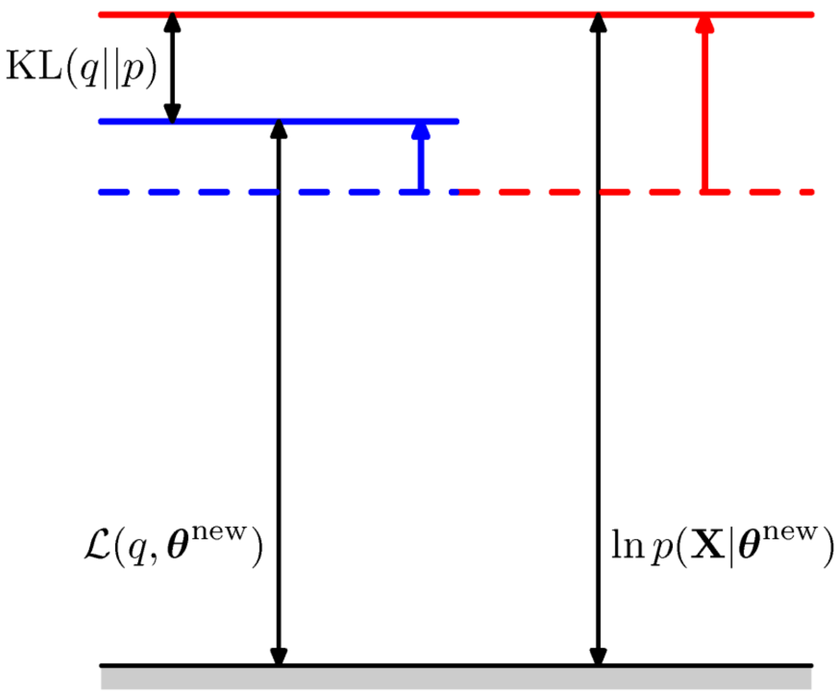


$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + KL(q||p)$$

- E-Step changes q(Z) to maximize $\mathcal{L}(q, \theta)$
- q has no effect on $\ln p(\mathbf{X}|\theta)$
- So maximizes when $KL(q||p) = 0$

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$$

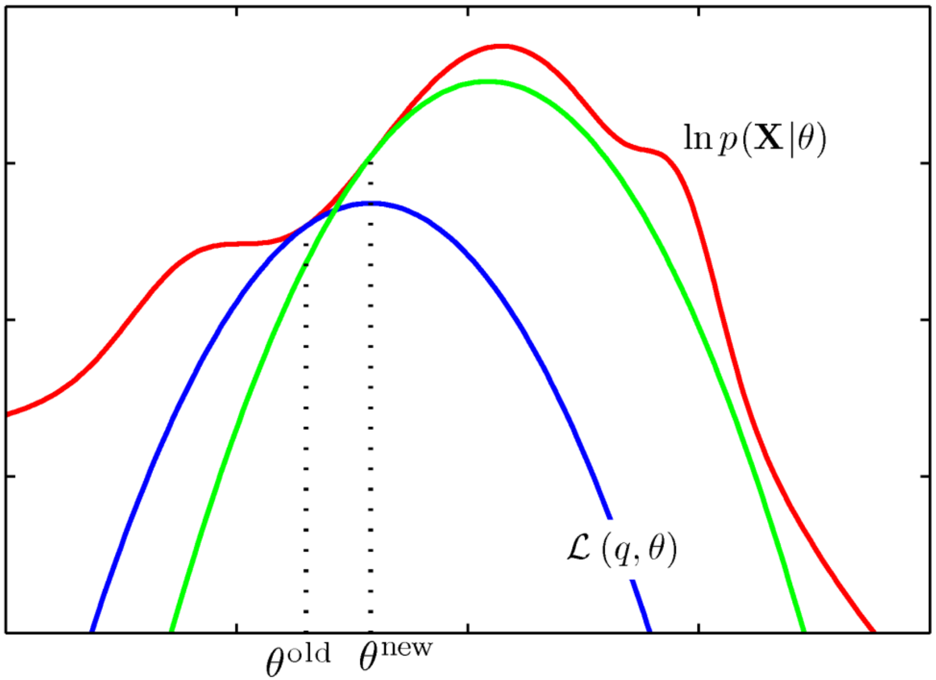
Visualize the M-Step



$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(q, \theta) + KL(q||p)$$

- Holding *q*(Z) constant increase $\mathcal{L}(q, \theta)$
- This increases $\ln p(\mathbf{X}|\theta)$
- But now $p \neq q$
- so $KL(q||p) > 0$

Another view of E-M



- Given old params, find *q* so that
- $\mathcal{L}(q, \theta)$ is tangent to $\ln p(\mathbf{X}|\theta)$
- Find new params to maximize $\mathcal{L}(q, \theta)$
- Then find new *q* to be tangent at a higher point.

Next

- Unsupervised Learing