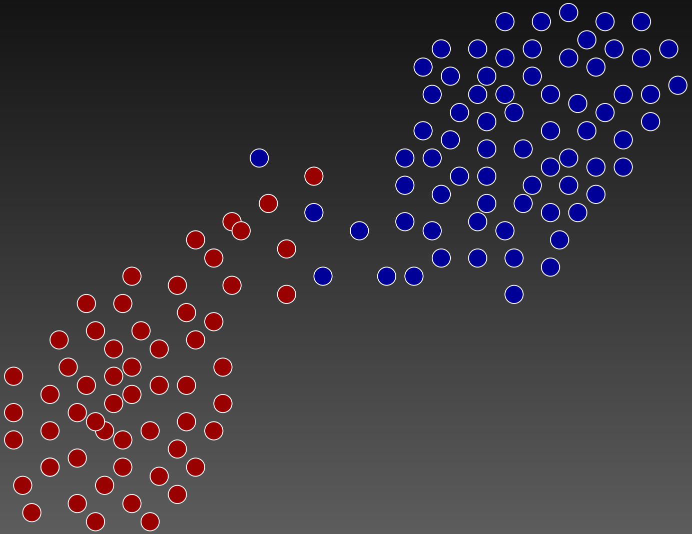
Mixture Models and EM

Yoan Miche

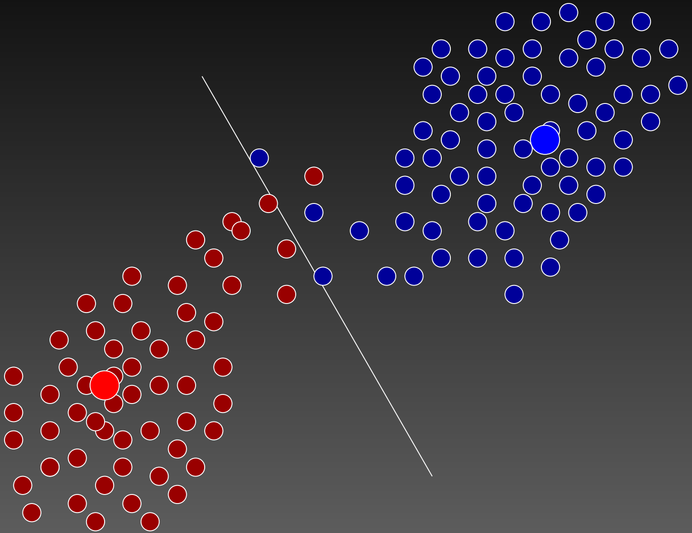
CIS, HUT

March 17, 2007

Mise en Bouche



Mise en Bouche



Outline

1. Starters: K-Means
2. Main Course: EM Algorithm
3. Dessert: General Algorithm

Setting

Data set of N observations of D-dimensional variable $\mathbf{x}$

$$\{\mathbf{x}_1, ..., \mathbf{x}_N\}$$

with each vector D-dimensional

$$\mathbf{x}_1 = (x_1^1, ..., x_1^D)$$

K-Means Idea: Clustering

Idea

- Define the K *prototype vectors* (code-book vectors) D-dimensional

$$\boldsymbol{\mu}_k$$

- All *prototypes* of the corresponding cluster: Sum of squares of distances from each data point to its closest prototype $\boldsymbol{\mu}_k$ is minimum.
- This is the condition to define the $\boldsymbol{\mu}_k$

K-Means Idea: Clustering

Idea

- Define the K *prototype vectors* (code-book vectors) D-dimensional

$$\boldsymbol{\mu}_k$$

- All *prototypes* of the corresponding cluster: Sum of squares of distances from each data point to its closest prototype $\boldsymbol{\mu}_k$ is minimum.
- This is the condition to define the $\boldsymbol{\mu}_k$

K-Means Idea: Clustering

Idea

- Define the K *prototype vectors* (code-book vectors) D-dimensional

$$\boldsymbol{\mu}_k$$

- All *prototypes* of the corresponding cluster: Sum of squares of distances from each data point to its closest prototype $\boldsymbol{\mu}_k$ is minimum.
- This is the condition to define the $\boldsymbol{\mu}_k$

Clustering idea (2)

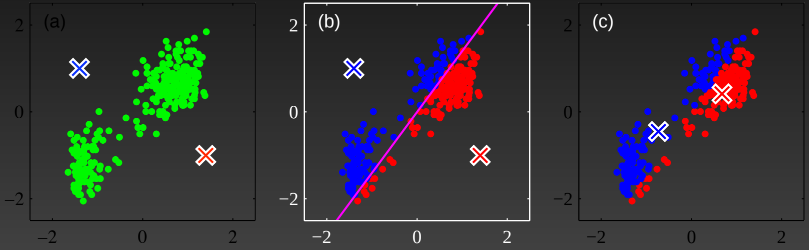
Minimize the *distortion measure J*

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 \quad (1)$$

Sum of squares of distances of all data points to their prototype.

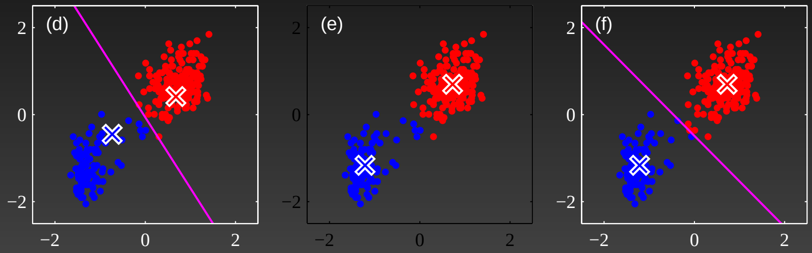
$$r_{nk} = \delta_{n,k} \quad \text{(Kronecker)}$$

The Old Faithful data set



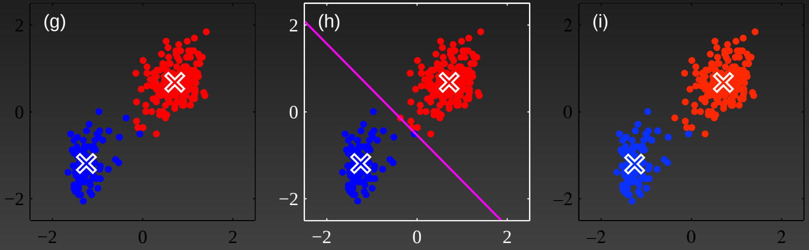
(a) (b) (c)

The Old Faithful data set (2)



(d) (e) (f)

The Old Faithful data set (3)



(g) (h) (i)

How to determine r_{nk} and $\boldsymbol{\mu}_k$

Iterative procedure in two steps

- E Step: Fix initial values for the $\{\boldsymbol{\mu}_k\}$ and minimize J with respect to the r_{nk}
- M Step: Fix the r_{nk} and minimize J with respect to the $\{\boldsymbol{\mu}_k\}$

In practice...

$$r_{nk} = \begin{cases} 1 & \text{if } k = \min_j ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 \\ 0 & else \end{cases} \quad (2)$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} \quad (3)$$

How to determine r_{nk} and $\boldsymbol{\mu}_k$

Iterative procedure in two steps

- E Step: Fix initial values for the $\{\boldsymbol{\mu}_k\}$ and minimize J with respect to the r_{nk}
- M Step: Fix the r_{nk} and minimize J with respect to the $\{\boldsymbol{\mu}_k\}$

In practice...

$$r_{nk} = \begin{cases} 1 & \text{if } k = \min_j ||\mathbf{x}_n - \boldsymbol{\mu}_k||^2 \\ 0 & else \end{cases} \tag{2}$$

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} \tag{3}$$

Generalizing to other "distances"

Generalizing to the K-medoids algorithm.

$$\widetilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \nu(\mathbf{x}_n - \boldsymbol{\mu}_k) \tag{4}$$

Image Clustering

Pixels considered as 3-dimensional vectors

(j) $K = 2$

(k) $K = 3$

(l) $K = 5$

(m) $K = 10$

Reminders and Notations

Gaussian mixture distribution:
Recall that a Gaussian mixture can be written as a sum of Gaussians:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \quad (5)$$

And let us introduce

$$\mathbf{z} = (\underset{z_1}{0}, \underset{z_2}{0}, \ldots, 0, \underset{z_k}{1}, 0, \ldots, \underset{z_K}{0}) \quad (6)$$

whose distribution, considering $p(z_k = 1) = \pi_k$ can be written in the form

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \quad (7)$$

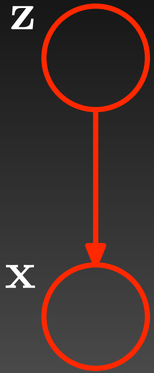
Reminders and Notations (2)

We also have

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k)^{zk} \qquad (8)$$

giving finally, by summing the joint distribution

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \Sigma_k) \qquad (9)$$



Only interest in these development in the introduction of the latent variable $\mathbf{z}_n$

Reminders and Notations (3)

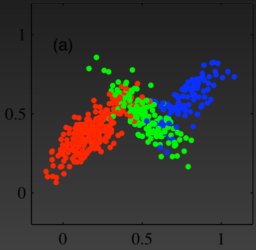
Lets us note, further, $\gamma(z_k) = p(z_k = 1|\mathbf{x})$

$$\gamma(z_k) = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{10}$$

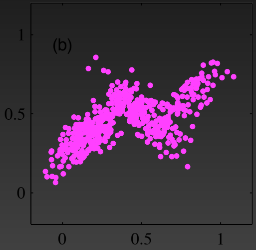
the conditional probability of $\mathbf{z}$ given $\mathbf{x}$.

$\gamma(z_k)$ is in fact the "responsibility" that k has in the observation $\mathbf{x}$

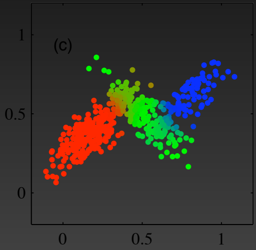
Example for mixture of Gaussians



(n)



(o)



(p)

And again...

This time, instead of codebook vectors, we use Gaussians:

Now...

- Maximizing the log-likelihood instead of the distortion measure
- $\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$ with $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$
- $\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$

And again...

This time, instead of codebook vectors, we use Gaussians:

Now...

- Maximizing the log-likelihood instead of the distortion measure
- $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$ with $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$
- $\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$

And again...

This time, instead of codebook vectors, we use Gaussians:

Now...

- Maximizing the log-likelihood instead of the distortion measure
- $\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$ with $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$
- $\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$

And again...

This time, instead of codebook vectors, we use Gaussians:

Now...

- Maximizing the log-likelihood instead of the distortion measure
- $\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$ with $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$
- $\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$

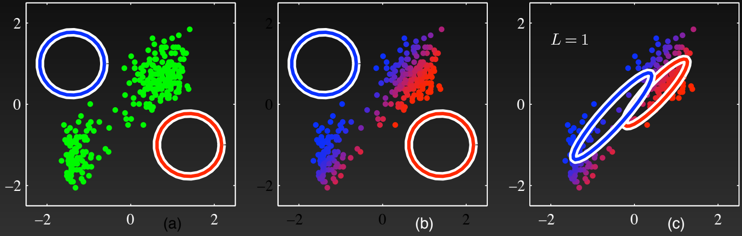
And again...

This time, instead of codebook vectors, we use Gaussians:

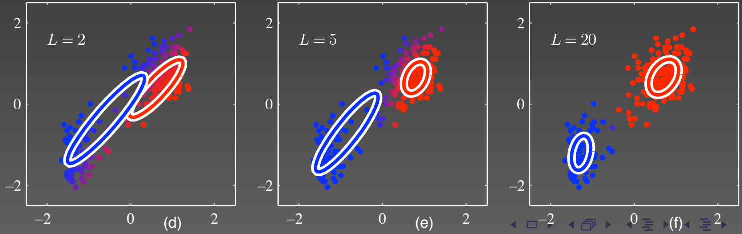
Now...

- Maximizing the log-likelihood instead of the distortion measure
- $\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$ with $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$
- $\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$

In images...



(q) (r) (s)



EM for Gaussian Mixtures (1)

EM Algorithm (1)

- Initialize $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k and evaluate log-likelihood with these
- E Step: $\gamma(z_{nk}) = \dfrac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$

EM for Gaussian Mixtures (1)

EM Algorithm (1)

- Initialize $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and π_k and evaluate log-likelihood with these
- E Step: $\gamma(z_{nk}) = \dfrac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\displaystyle\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$

EM for Gaussian Mixtures (2)

EM Algorithm (2)

- M Step: $\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n,$

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

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

EM for Gaussian Mixtures (3)

EM Algorithm (3)

- Evaluate log-likehood

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{j=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Kidding...

The proposed version of the general algorithm in the book is way too heavy to be presented on slides. . .

Meanwhile, I invite you to read section 9.4 about this general algorithm since it gives a general framework that is given in Chapter 10

Or not. . .

Had fun ?

Hope you enjoyed. . .