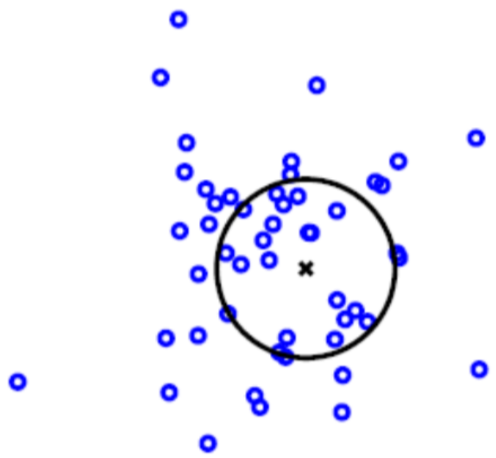
50.007
Machine Learning

Lu, Wei

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Mixture Models and Expectation Maximization

Generative Process
Spherical Gaussian



The likelihood for each point:

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}||x-\mu||^2)$$

Point Mean Variance

Generative Process
Spherical Gaussian

Our training set has the points

$$S_n = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$$

The likelihood for each point is:

$$p(x^{(t)}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}||x^{(t)} - \mu||^2)$$

What is the overall objective that we would like to optimize for this training set?

Generative Process
Spherical Gaussian

$$p(x^{(t)}|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}||x^{(t)} - \mu||^2)$$

Overall objective:

$$\ell(S_n|\mu,\sigma^2) = \sum_{t=1}^{n} \log p(x^{(t)}|\mu,\sigma^2)$$

Generative Process
Spherical Gaussian

$$p(x^{(t)}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp(-\frac{1}{2\sigma^2}||x^{(t)} - \mu||^2)$$

Overall objective:

$$\ell(S_n|\mu, \sigma^2) = \sum_{t=1}^{n} \log p(x^{(t)}|\mu, \sigma^2)$$

$$= \sum_{t=1}^{n} \left[-\frac{d}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||x^{(t)} - \mu||^2\right]$$

$$= -\frac{dn}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}||x^{(t)} - \mu||^2$$

Generative Process
Spherical Gaussian

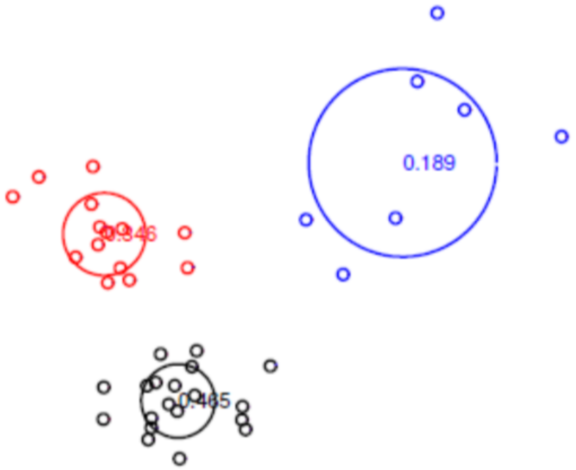
$$\ell(S_n|\mu,\sigma^2) = -\frac{dn}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{n}||x^{(t)} - \mu||^2$$

$$\frac{\partial \ell(S_n|\mu,\sigma^2)}{\partial \mu} = 0 \qquad \hat{\mu} = \frac{1}{n}\sum_{t=1}^{n} x^{(t)}$$

$$\frac{\partial \ell(S_n|\mu,\sigma^2)}{\partial \sigma^2} = 0 \qquad \hat{\sigma}^2 = \frac{1}{dn}\sum_{t=1}^{n}||x^{(t)} - \hat{\mu}||^2$$

Maximum Likelihood Estimators

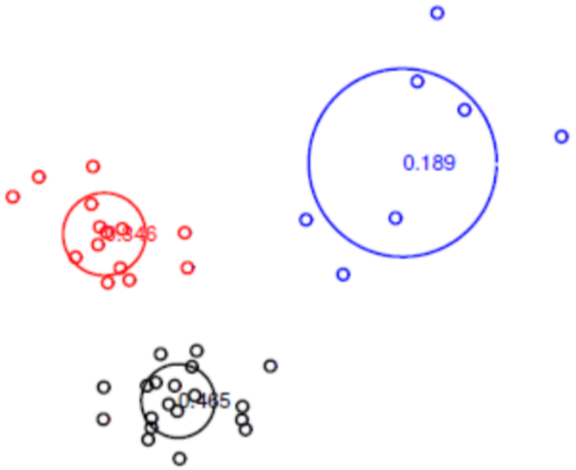
Mixture of Gaussians
Labeled Case



How do we model the generation of each point in this case?

Mixture of Gaussians

Labeled Case



$$i \sim \text{Multinomial}(p_1, \ldots, p_k)$$

$$x \sim p(x|\mu^{(i)}, \sigma_i^2)$$

Mixture of Gaussians
Labeled Case

$$\delta(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{t=1}^{n} \Big[\underbrace{\log\left(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2)\right)}_{\text{Only one term, as specified by } \delta(i|t)=1} \Big]$$

Mixture of Gaussians
Labeled Case

$$\delta(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{t=1}^{n} \left[\underbrace{\sum_{i=1}^{k} \delta(i|t) \log \left(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2) \right)}_{\text{Only one term is non-zero, as specified by } \delta(i|t)} \right]$$

Mixture of Gaussians
Labeled Case

$$\delta(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{t=1}^{n} \left[\sum_{i=1}^{k} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right) \right]$$

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right) \right]$$

Mixture of Gaussians

Labeled Case

$$\sum_{i=1}^{k} \Big[\underbrace{\sum_{t=1}^{n} \delta(i|t) \log \Big(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2) \Big)}_{\text{The objective for all points in } i\text{-th cluster}} \Big]$$

Mixture of Gaussians
Labeled Case

$$\underbrace{\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log \left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right)\right]}_{\text{The objective for all points in } i\text{-th cluster}}$$

$$\hat{n}_i = \sum_{t=1}^{n} p(i|t) \quad \hat{p}_i = \frac{\hat{n}_i}{n}$$

(fraction of points in cluster i)

$$\hat{\mu}^{(i)} = \frac{1}{\hat{n}_i} \sum_{t=1}^{n} \delta(i|t) x^{(t)}$$

(mean of points in cluster i)

$$\hat{\sigma_i^2} = \frac{1}{d\hat{n}_i} \sum_{t=1}^{n} \delta(i|t) ||x^{(t)} - \hat{\mu}^{(i)}||^2$$

(mean squared spread in cluster i)

Mixture of Gaussians

Labeled Case

$$\underbrace{\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right)\right]}_{\text{The objective for all points in } i\text{-th cluster}}$$

$\hat{n}_i = \sum_{t=1}^{n} p(i|t)$ $\hat{p}_i = \frac{\hat{n}_i}{\ldots}$

(fraction of

Parameter Estimation (Supervised Learning)

in cluster i)

$$\sigma_i^2 = \frac{1}{d\hat{n}_i} \sum_{t=1}^{n} \delta(i|t) ||x^{(t)} - \hat{\mu}^{(i)}||^2$$

(mean squared spread in cluster i)

Mixture of Gaussians

Labeled Case

Now, assume we have finished learning.

For any point $x^{(t')}$

which cluster it belongs to?

Mixture of Gaussians
Labeled Case

Now, assume we have finished learning.

For any point $x^{(t')}$

which cluster it belongs to?

$$\delta(i|t') = \begin{cases} 1 & \text{if } i = \arg\max_i p_j \cdot p(x^{(t')}|\mu^{(j)}, \sigma_j^2) \\ 0 & \text{otherwise} \end{cases}$$

Mixture of Gaussians
Labeled Case

Now, assume we have finished learning.

For any point $x^{(t')}$

which cluster it belongs to?

Evaluation (Testing)

$\delta(i|t')$ … $p(x^{(t')}|\mu^{(j)}, \sigma_j^2)$

otherwise

Mixture of Gaussians

Unlabeled Case

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log \left(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2) \right) \right]$$

These guys are now not given to you!

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right) \right]$$

$$\delta(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

$$\boxed{\sum_{i}^{k} \delta(i|t) = 1}$$

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right)\right]$$

Initialization

Randomly initialize the model parameters

(after this, we know the values for $p_i, \mu^{(i)}, \sigma_i^2$)

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k}\left[\sum_{t=1}^{n}\delta(i|t)\log\left(p_i\cdot p(x^{(t)}|\mu^{(i)},\sigma_i^2)\right)\right]$$

Expectation

find new assignments

(after this, we know the values for $\delta(i|t)$)

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k}\left[\sum_{t=1}^{n}\delta(i|t)\log\left(p_i\cdot p(x^{(t)}|\mu^{(i)},\sigma_i^2)\right)\right]$$

Expectation

Evaluation (Testing)

find new assignments

(after this, we know the values for $\delta(i|t)$)

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k}\left[\sum_{t=1}^{n}\delta(i|t)\log\left(p_i\cdot p(x^{(t)}|\mu^{(i)},\sigma_i^2)\right)\right]$$

Maximization

update the model parameters

(after this, we know the updated model parameters: $p_i, \mu^{(i)}, \sigma_i^2$)

Mixture of Gaussians
Unlabeled Case

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} \delta(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right)\right]$$

Maximization

Parameter Estimation (Supervised Learning)

update the model parameters

(after this, we know the updated model parameters: $p_i, \mu^{(i)}, \sigma_i^2$)

Mixture of Gaussians

Hard EM

$$\sum_{i=1}^{k}\left[\sum_{t=1}^{n} \color{green}{\delta(i|t)} \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)},\sigma_i^2)\right)\right]$$

These are
binary variables

$$\delta(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

This can be understood as a *collapsed* distribution!

Mixture of Gaussians

Hard EM

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} p(i|t) \log \left(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2) \right) \right]$$

$$p(i|t) = \begin{cases} 1 & \text{if } x^{(t)} \text{ is assigned to } i \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{i}^{k} p(i|t) = 1$$

Mixture of Gaussians

Soft EM

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} p(i|t) \log \left(p_i \cdot p(x^{(t)} | \mu^{(i)}, \sigma_i^2) \right) \right]$$

$$p(i|t) = \text{probability that } x^{(t)} \text{ is assigned to } i$$

$$\boxed{\sum_{i}^{k} p(i|t) = 1}$$

Mixture of Gaussians
Soft EM

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} p(i|t) \log \left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2) \right) \right]$$

Expectation

find new soft assignments

$$p(i|t) = \frac{p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)}{\sum_j p_j \cdot p(x^{(t)}|\mu^{(j)}, \sigma_j^2)}$$

Mixture of Gaussians

Soft EM

$$\sum_{i=1}^{k} \left[\sum_{t=1}^{n} p(i|t) \log\left(p_i \cdot p(x^{(t)}|\mu^{(i)}, \sigma_i^2)\right)\right]$$

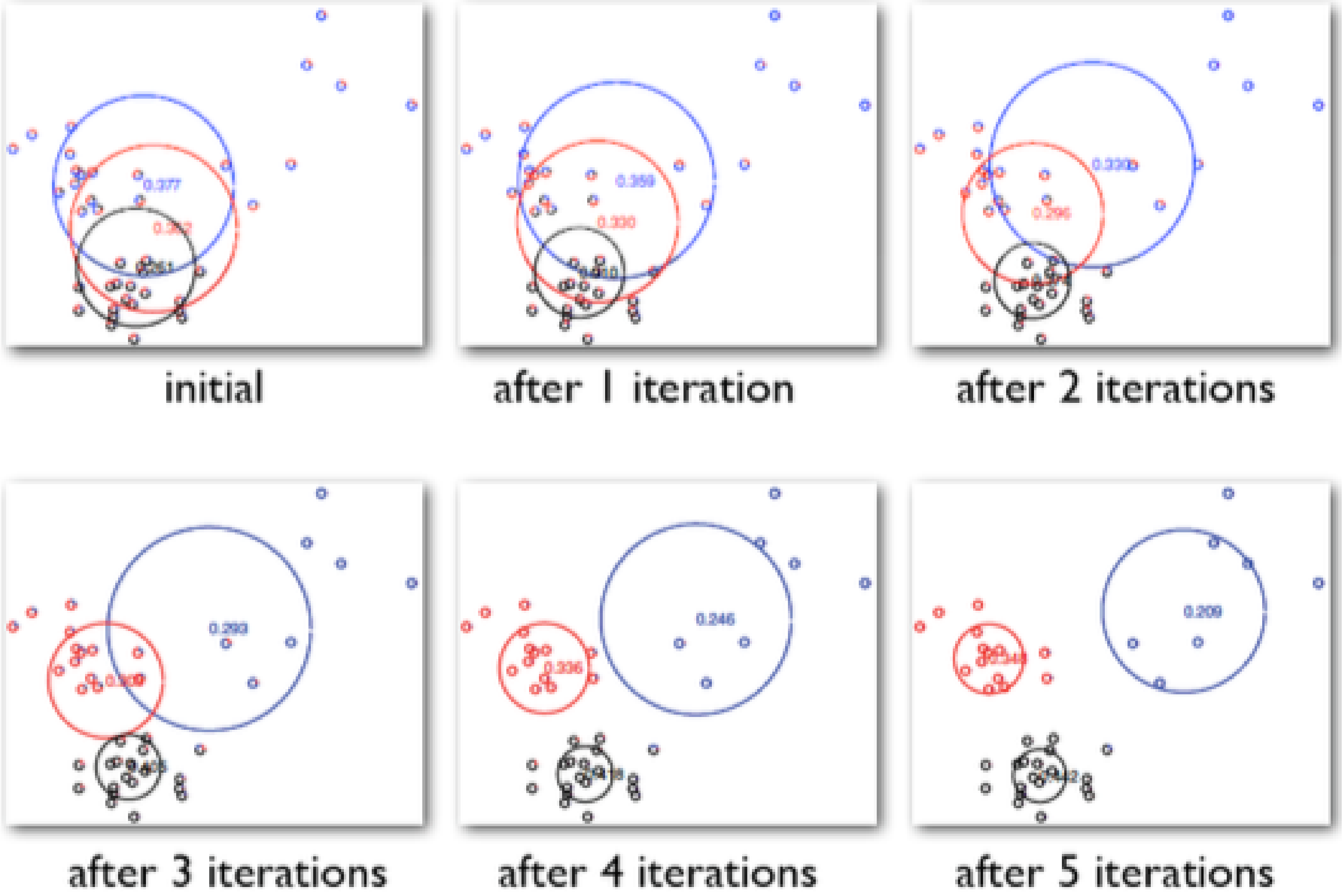
Maximization

update the model parameters

$$\hat{n}_i = \sum_{t=1}^{n} p(i|t) \qquad \hat{p}_i = \frac{\hat{n}_i}{n} \qquad \hat{\mu}^{(i)} = \frac{1}{\hat{n}_i}\sum_{t=1}^{n} p(i|t) x^{(t)}$$

$$\hat{\sigma_i^2} = \frac{1}{d\hat{n}_i}\sum_{t=1}^{n} p(i|t)||x^{(t)} - \hat{\mu}^{(i)}||^2$$

Mixture of Gaussians
Soft EM



initial | after 1 iteration | after 2 iterations

after 3 iterations | after 4 iterations | after 5 iterations

Question

Is there any guarantee on the soft EM?

Soft EM Algorithm

There is a guarantee that after each iteration, the objective does not decrease.

However, there is no guarantee that it will reach the global optimal value (similar to k-means)

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

What is the Objective?

$$\log \prod_t p(x^{(t)})$$

$$\log p(x)$$

OPTIONAL FROM HERE ON

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$\mathcal{L}(\theta) = \log p_\theta(x)$$

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$\begin{aligned}\mathcal{L}(\theta) &= \log p_\theta(x) \\ &= \log \sum_y p_\theta(x, y)\end{aligned}$$

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$\begin{aligned}
\mathcal{L}(\theta) &= \log p_\theta(x) \\
&= \log \textstyle\sum_y p_\theta(x, y) \\
&= \log \textstyle\sum_y q(y) \frac{p_\theta(x,y)}{q(y)}
\end{aligned}$$

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$
\begin{aligned}
\mathcal{L}(\theta) &= \log p_\theta(x) \\
&= \log \sum_y p_\theta(x, y) \\
&= \log \sum_y q(y) \frac{p_\theta(x,y)}{q(y)} \\
&\geq \sum_y q(y) \log \frac{p_\theta(x,y)}{q(y)}
\end{aligned}
$$

See the whiteboard to understand why this step makes sense.

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$
\begin{aligned}
\mathcal{L}(\theta) &= \log p_\theta(x) \\
&= \log \textstyle\sum_y p_\theta(x, y) \\
&= \log \textstyle\sum_y q(y) \frac{p_\theta(x,y)}{q(y)} \\
&\geq \textstyle\sum_y q(y) \log \frac{p_\theta(x,y)}{q(y)} = F(q, \theta)
\end{aligned}
$$

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$F(q, \theta)$$

$$= \sum_y q(y) \log \frac{p_\theta(x,y)}{q(y)}$$

$$= \sum_y q(y) \log \frac{p_\theta(x) p_\theta(y|x)}{q(y)}$$

$$= \sum_y q(y) \log p_\theta(x) - \sum_y q(y) \log \frac{q(y)}{p_\theta(x) p_\theta(y|x)}$$

$$= \mathcal{L}(\theta) - \mathbf{KL}(q(y) || p_\theta(y|x))$$

Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$F(q,\theta) = \mathcal{L}(\theta) - \mathbf{KL}(q(y)||p_\theta(y|x))$$

$$\begin{aligned} q^{t+1} &= \arg\max_q F(q,\theta^t) \\ \text{(E-step)} \quad &= \arg\min_q \mathbf{KL}(q(y)||p_{\theta^t}(y|x)) \end{aligned}$$

$$\begin{aligned} \theta^{t+1} &= \arg\max_\theta F(q^{t+1},\theta) \\ \text{(M-step)} \quad &= \arg\max_\theta \mathbf{E}_{q^{t+1}}[\log p_\theta(x,y)] \end{aligned}$$

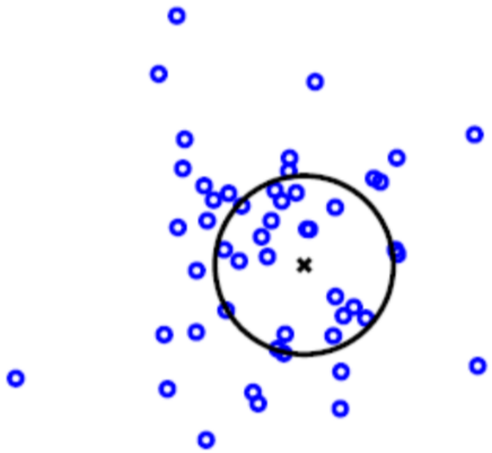
Soft EM Algorithm

Training set: $(x^{(1)}, x^{(2)}, \ldots, x^{(n)}) = x$

$$F(q,\theta) = \mathcal{L}(\theta) - \mathbf{KL}(q(y)||p_\theta(y|x))$$

$$\begin{aligned}\mathcal{L}(\theta^{t+1}) &= F(q^{t+2}, \theta^{t+1}) \\ &\geq F(q^{t+1}, \theta^{t+1}) \\ &\geq F(q^{t+1}, \theta^{t}) = \mathcal{L}(\theta^{t})\end{aligned}$$

Generative Models



Continuous

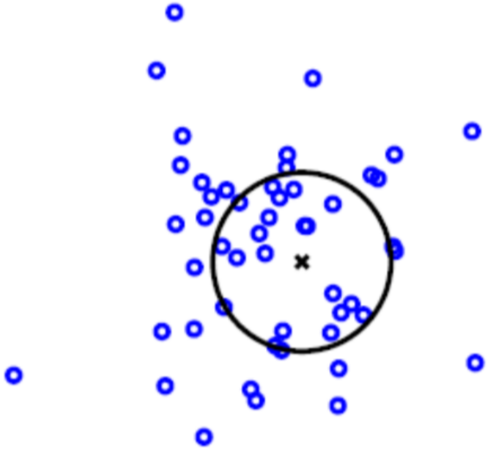
$$a\ b\ c$$

Discrete

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5$$

Sequential

Generative Models



$a\ b\ c$

Continuous

Discrete

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5$$

Hidden Markov Model

Sequential