Elective Module:

Advanced ML Techniques

Chapter 9

EM Algorithm

Expectation Maximization

EM Algorithm

- EM stands for **Expectation-Maximization**.
- This can be used for **probabilistic-clustering** or **soft-clustering**

K-Means : Hard clustering

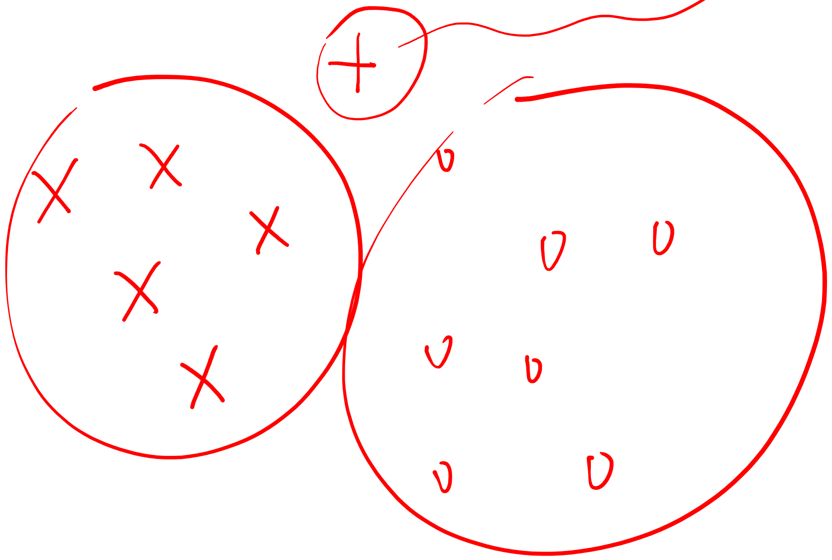
Probabilistic clustering

- What is **probabilistic-clustering**?

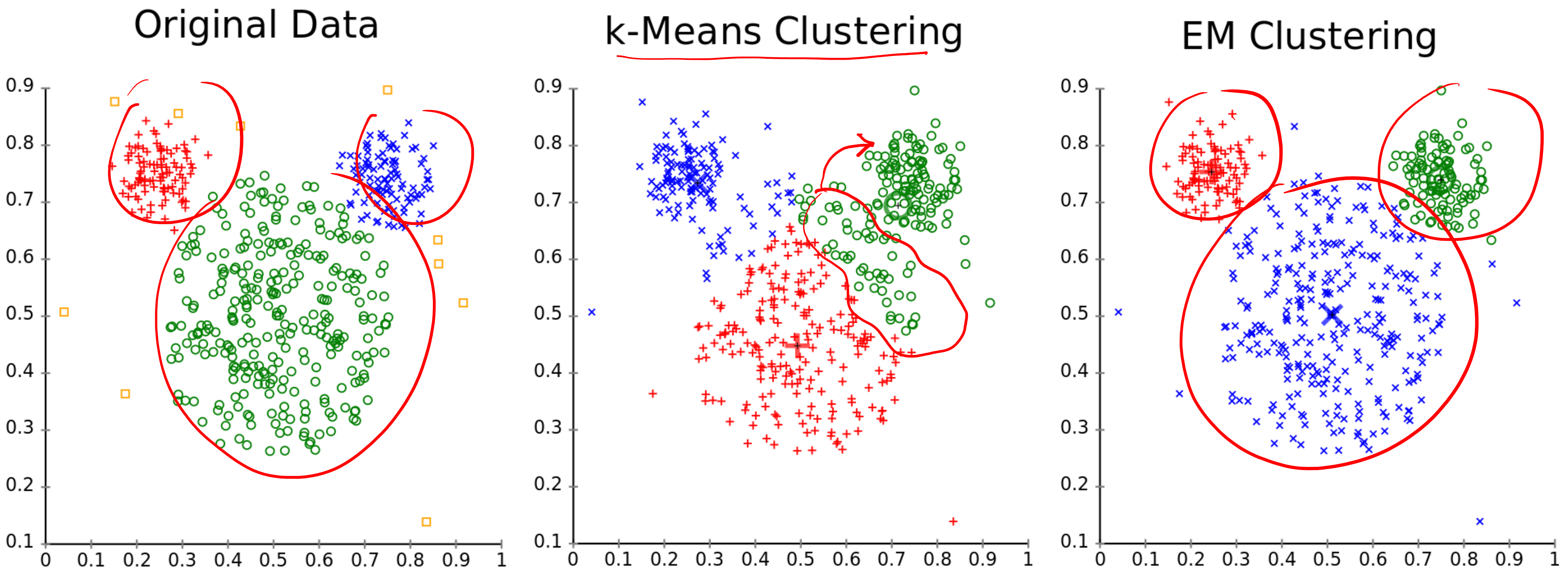
Probabilistic clustering

- Previously we assigned each point to a **unique cluster**.
- Now we calculate the **probability** that a data point belongs to a cluster!

probability belongs to a cluster.



Different cluster analysis results on "mouse" data set:



EM Algorithm

Prior probability of i^{th} cluster

- Consider a **Gaussian cluster model**.
- With probability p_i generate a sample $\bar{\mathbf{x}}$ from **Gaussian** ***cluster*** i i.e.

$$\mathcal{N}(\bar{\mu}_i, \sigma^2 I) \to \Sigma_i$$

centroid cluster

covariance matrix

EM Algorithm

K clusters.

- Consider a **Gaussian cluster model**.
- With probability p_i generate a sample $\bar{\mathbf{x}}$ from **Gaussian** *cluster* i i.e.

$$N(\bar{\boldsymbol{\mu}}_i, \sigma^2 \mathbf{I})$$

$$\underbrace{N(\bar{\mu}_1, \sigma^2 I),\ N(\bar{\mu}_2, \sigma^2 I),\ \ldots,\ N(\bar{\mu}_K, \sigma^2 I)}_{\text{K clusters.}}$$

EM Algorithm

$p_1, p_2, \ldots, p_K$: Prior probabilities of clusters.

- The PDF is given as

$$f(\bar{x}) = \sum_{i=1}^{K} p_i \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\frac{1}{2\sigma^2}\|\bar{x}-\bar{\mu}_i\|^2}}_{\text{Probability density function}}$$

Gaussian mixture model.

EM Algorithm

- The PDF is given as

$$f_X(\bar{\mathbf{x}}) = \sum_{i=1}^{K} p_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{\mathbf{x}}-\bar{\boldsymbol{\mu}}_i\|^2}$$

EM Algorithm

$$f_X(\bar{\mathbf{x}}) = \sum_{i=1}^{K} p_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{\mathbf{x}}-\bar{\boldsymbol{\mu}}_i\|^2}$$

- This is termed as a Gaussian mixture

K Gaussians:
One for each cluster

EM Algorithm

$$f_X(\bar{\mathbf{x}}) = \sum_{i=1}^{K} p_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{\mathbf{x}}-\bar{\boldsymbol{\mu}}_i\|^2}$$

- This is termed as a ***Gaussian mixture***.

EM Algorithm

- Consider now M data points

cluster?

$$\underbrace{\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)}_{\text{Data}}$$

EM Algorithm

- Consider now M data points

$$\bar{\mathbf{x}}(1), \bar{\mathbf{x}}(2), \dots, \bar{\mathbf{x}}(M)$$

EM Algorithm

Centroids.

- We desire to ***estimate*** $\overline{\boldsymbol{\mu}}_i$

$\alpha_i(j)$.

- As well as the **cluster assignments**
- Performing direct ML estimation is **mathematically intractable**.

NOT possible.

Maximum Likelihood.

cluster assignments.

Cluster information

- However, if ***cluster assignment*** is known, problem is ***simple***!

Cluster assignment $M=8$

- For example
- Cluster 1: $\bar{\mathbf{x}}(1), \bar{\mathbf{x}}(3), \bar{\mathbf{x}}(5), \bar{\mathbf{x}}(8)$: cluster 1.
- Cluster 2: $\bar{\mathbf{x}}(2), \bar{\mathbf{x}}(4), \bar{\mathbf{x}}(6), \bar{\mathbf{x}}(7)$: cluster 2

Cluster assignment

- For example

$$\hat{\boldsymbol{\mu}}_1 = \frac{\bar{x}(1) + \bar{x}(3) + \bar{x}(5) + \bar{x}(8)}{4}$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{\bar{x}(2) + \bar{x}(4) + \bar{x}(6) + \bar{x}(7)}{4}$$

Average of points in cluster 1.

Average of points in cluster 2

NO Labels!

Unsupervised Learning

Cluster assignment

- For example

$$\hat{\boldsymbol{\mu}}_1 = \frac{\bar{\mathbf{x}}(1) + \bar{\mathbf{x}}(3) + \bar{\mathbf{x}}(5) + \bar{\mathbf{x}}(8)}{4}$$

$$\hat{\boldsymbol{\mu}}_2 = \frac{\bar{\mathbf{x}}(2) + \bar{\mathbf{x}}(4) + \bar{\mathbf{x}}(6) + \bar{\mathbf{x}}(7)}{4}$$

Cluster information

cluster assignment variable.

- We introduce the concept of ***missing data or latent information***

$$\alpha_i(j) = \begin{cases} 1 & \text{if } \vec{x}(j) \in \mathcal{C}_i \\ 0 & \text{if } \vec{x}(j) \notin \mathcal{C}_i \end{cases}$$

Cluster information

- We introduce the concept of ***missing data or latent information***

$$\alpha_i(j) = \begin{cases} 1 & \bar{\mathbf{x}}(j) \in \mathcal{C}_i \\ 0 & \bar{\mathbf{x}}(j) \notin \mathcal{C}_i \end{cases}$$

Missing!

Complete data

$\bar{x}(1), \bar{x}(2), \ldots, \bar{x}(M)$:

$\alpha_i(j)$ $\quad i = 1, 2, \ldots, K$

$\qquad\quad j = 1, 2, \ldots, M$

Complete Data.

Missing Data

Latent information

Complete data

$$\underbrace{\bar{\mathbf{x}}(j), \alpha_i(j)}_{\text{Complete data}}$$

Likelihood

- The **likelihood** of the ***complete data***

Missing data.

$$\underbrace{\prod_{j=1}^{M}\prod_{i=1}^{K}\left(p_i\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\|\bar{x}(j)-\bar{\mu}_i\|^2}\right)^{\alpha_i(j)}}_{\text{joint PDF of points.}} = Q(\bar{\mu})$$

Likelihood function

Likelihood

- The <u>**likelihood**</u> of the ***complete data***

$$\prod_{j=1}^{M} \prod_{i=1}^{K} \left(p_i \times \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2} \|\bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i\|^2} \right)^{\alpha_i(j)}$$

Products become sum

for Log.

Log-Likelihood

Missing Data!

- The **log-likelihood** of the ***complete data***

$$\underbrace{\sum_{j=1}^{M}\sum_{i=1}^{K}\alpha_i(j)\left(P_i - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|\bar{x}(j)-\bar{\mu}_i\|^2\right)}_{\text{Log likelihood.}}$$

Log-Likelihood

- The **log-*likelihood*** of the ***complete data***

$$\sum_{j=1}^{M}\sum_{i=1}^{K}\alpha_i(j)\left(\ln p_i - \frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i\|^2\right)$$

Log Likelihood.

Expected value of $\alpha_i(j)$ in iteration ℓ.

$$E\{\alpha_i(j)\} = \alpha_i^{(\ell)}(j)$$

EM Algorithm

E - Step
M - Step.

- EM Algorithm ***proceeds iteratively***.
- Consider the $l - 1$th iteration with centroids

$\bar{\mu}_1^{(l-1)}, \bar{\mu}_2^{(l-1)}, \ldots, \bar{\mu}_K^{(l-1)}$: Centroids in $(l-1)^{th}$ iteration

Expectation Maximization

EM Algorithm

- EM Algorithm ***proceeds iteratively***.
- Consider the $l-1$th iteration with centroids

K = # clusters.

Centroids. cluster means.

$$\overline{\boldsymbol{\mu}}_0^{(l-1)}, \overline{\boldsymbol{\mu}}_1^{(l-1)}, \dots, \overline{\boldsymbol{\mu}}_K^{(l-1)}$$

Gaussian mixtures.

EM Algorithm

E – Step

Expectation

- The **expected value** of the log-likelihood in iteration l is

$$\sum_{j=1}^{M}\sum_{i=1}^{K}\alpha_i^{(l)}(j)\left\{\ln p_i-\frac{N}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\|x(j)-\mu_i\|^2\right\}$$

$$E\{\alpha_i(j)\}$$

EM Algorithm

- The **expected value** of the log-likelihood in iteration l is

$$\underbrace{\sum_{j=1}^{M}\sum_{i=1}^{K}\alpha_i^{(l)}(j)\left(\ln p_i - \frac{N}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\|\bar{\mathbf{x}}(j) - \bar{\boldsymbol{\mu}}_i\|^2\right)}_{Q(\bar{\boldsymbol{\mu}})}$$

$Q(\bar{u})$.

EM Algorithm

- How to calculate $\alpha_i^{(l)}(j)$?

$= 1$ if $\bar{x}(j) \in \mathcal{C}_i$

$= \Pr(\bar{x}(j) \in \mathcal{C}_i) \times 1$
$+\, 0 \times \Pr(\bar{x}(j) \notin \mathcal{C}_i)$
$= \underline{\Pr(\bar{x}(j) \in \mathcal{C}_i)}$

Probability! can take any value in [0,1]

$$\alpha_i^{(l)}(j) = \Pr(c_i|\bar{\mathbf{x}}(j))$$

$$= \frac{\Pr(\bar{x}(j)|\mathcal{C}_i)\cdot P(\mathcal{C}_i)}{\sum_k \Pr(\bar{x}(j)|\mathcal{C}_k)P(\mathcal{C}_k)}$$: Bayes rule.

$$= \frac{P_i \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\frac{1}{2\sigma^2}\|\bar{x}(j)-\bar{\mu}_i\|^2}}{\sum_k \cdot P_k \cdot \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} \cdot e^{-\frac{1}{2\sigma^2}\|\bar{x}(j)-\bar{\mu}_k\|^2}} \rightarrow E\{\alpha_i(j)\}$$

EM Algorithm

- How to calculate $\alpha_i^{(l)}(j)$?

$$\alpha_i^{(l)}(j) = \Pr\left(\mathcal{C}_i \middle| \bar{\mathbf{x}}(j)\right)$$

$$= \frac{p_i \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\left\|\bar{\mathbf{x}}(j)-\bar{\boldsymbol{\mu}}_i^{(l-1)}\right\|^2}}{\sum_{k=1}^{K} p_k \times \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{N}{2}} e^{-\frac{1}{2\sigma^2}\left\|\bar{\mathbf{x}}(j)-\bar{\boldsymbol{\mu}}_k^{(l-1)}\right\|^2}}$$

M-Step

M = Maximization.

Maximize Expected value of log likelihood.

- The M-Step is the **maximization step**
- To determine $\bar{\mathbf{\mu}}_j$, **differentiate** with respect to $\bar{\mathbf{\mu}}_j$ and set equal to zero

M-Step

To maximize take gradient and set equal to zero.

$$\nabla_{\bar{\boldsymbol{\mu}}_i} Q(\bar{\boldsymbol{\mu}}) = 0$$

$$\Rightarrow \bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(\ell)}(j)\, \bar{x}(j)}{\sum_{j=1}^{M} \alpha_i^{(\ell)}(j)}$$

M-Step

$$\nabla_{\bar{\boldsymbol{\mu}}_i} Q(\bar{\boldsymbol{\mu}}) = 0$$

$$\Rightarrow \bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}$$

centroid of
ith cluster in
iteration l.

K-Means vs EM

- Compare the expressions in K-Means and EM

EM: $\overline{\boldsymbol{\mu}}_i^{(l)} = \dfrac{\sum_j \alpha_i^{(l)}(j)\, \overline{x}(j)}{\sum_j \alpha_i^{(l)}(j)}$: EM.

K − Means: $\overline{\boldsymbol{\mu}}_i^{(l)} = \dfrac{\sum_j \alpha_i^{(l)}(j)\, \overline{x}(j)}{\sum_j \alpha_i^{(l)}(j)}$

K-Means vs EM

- Compare the expressions in K-Means and EM

$$\underbrace{\bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}}_{EM}$$

Same Expression!

$$\underbrace{\bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}}_{K-Means}$$

K-Means vs EM

- Both are **identical!**
- What then is the **difference?**

K-Means vs EM

Weighted Average

$$\underbrace{\bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}}_{EM}$$

Soft clustering.

Probabilities. $\in [0, 1]$

Average.

$$\underbrace{\bar{\boldsymbol{\mu}}_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)\bar{\mathbf{x}}(j)}{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}}_{K-Means}$$

$\in \{0, 1\}$

Either 0 or 1

Hard clustering

K-Means vs EM

Can only take 2 possible values.

- If you observe carefully, in K-Means $\alpha_i^{(l)}(j) \in \{0,1\}$
- In EM, $\underbrace{\alpha_i^{(l)}(j) \in [0,1]}$

any value in interval [0,1].
Ex: 0.95

K-Means vs EM

- Therefore, in EM, $\alpha_i^{(l)}(j)$ denote probabilities.

Prior probabilities

Sum of all aposteriori probabilities of points in cluster i

- Finally, the prior probabilities p_i can also be computed as follows

$Pr(\bar{x}(j) \in C_i)$

$$p_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}{M}$$

can also Be computed.

Prior probabilities

- Finally, the prior probabilities p_i can also be computed as follows

$$p_i^{(l)} = \frac{\sum_{j=1}^{M} \alpha_i^{(l)}(j)}{M}$$

Estimate of prior probability in iteration l.

Instructors may use this white area (14.5 cm / 25.4 cm) for the text.
Three options provided below for the font size.

Font: Avenir (Book), Size: 32, Colour: Dark Grey

Font: Avenir (Book), Size: 28, Colour: Dark Grey

Font: Avenir (Book), Size: 24, Colour: Dark Grey

Do not use the space below.