Expectation Maximization = probabilistic variant of K-means.

- Make probabilistic calculations, instead of hard assignments.
- More powerful algorithm of doing the same task.

Clustering:

Inputs: n objects (data points) $\{x_i\}_{i=1}^n$ and K clusters.

Goal: group the points into several groups.

group ID: $z_i \in \{1, \cdots, K\}$, for each x_i.

Intro to EM (Idea)

K-Means

Cluster Assignment:

$$z_i = \underset{k=1,\cdots,K}{\operatorname{argmin}} \|x_i - \mu_k\|^2$$

cluster i ← z_i

// iterate the centroids and find the one with smallest squared distance to x_i, and it to z_i.

Centroid calculation:

$$\mu_k = \frac{\sum_{i=1}^n \mathbb{I}(z_i = k)\, x_i}{\sum_{i=1}^n \mathbb{I}(z_i = k)}$$

(μ_k ← centroid k)

where $\mathbb{I}(z_i = k) = \begin{cases} 1, & z_i = k \\ 0, & z_i \neq k \end{cases}$

// calculate the average location of those belong to cluster $z_i = k$.

EM

$$p(z_i = k) = \frac{\exp\left(-\|x_i - \mu_k\|^2 / \lambda\right)}{\sum_{i=1}^K \exp\left(-\|x_i - \mu_k\|^2 / \lambda\right)}$$

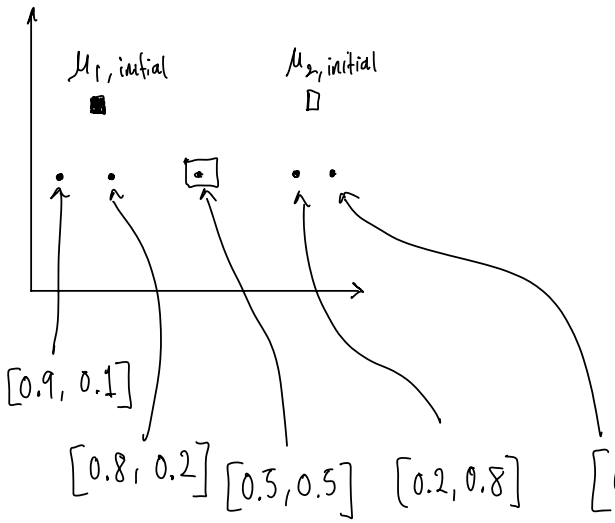
(λ ← Const)

// sigmoid function.

replaced the indicator

$$\mu_k = \frac{\sum_{i=1}^n p(z_i = k)\, x_i}{\sum_{i=1}^n p(z_i = k)}$$

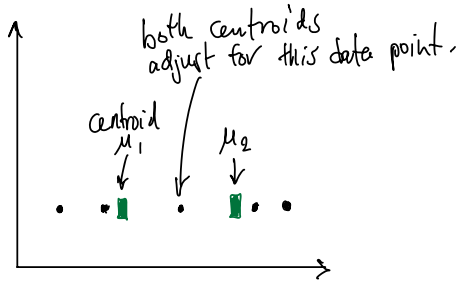
// weighted average. Instead of indicator function, it uses a probability.



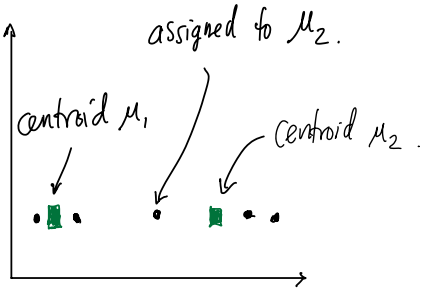
For $\mu_1 = \dfrac{0.9X_1 + 0.8X_2 + 0.5X_3 + 0.2X_4 + 0.1X_5}{0.9 + 0.8 + 0.5 + 0.2 + 0.1}$ // weighted average.

contributes the most.

$$\mu_2 = \frac{0.1X_1 + 0.2X_2 + 0.5X_3 + 0.8X_4 + 0.9X_5}{0.1 + 0.2 + 0.5 + 0.8 + 0.9}$$



Expectation Maximization
End Result.



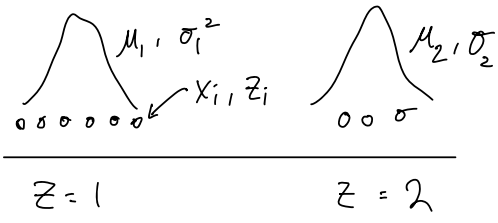
K-means
End Result.

- More generally, EM algorithm is a special optimization algorithm for maximum likelihood estimation of Gaussian mixtures (or latent variable models).

Probabilistic Modeling of Clustering: General

Probabilistic approach:

- Assume a joint distribution $p(x, z \mid \theta)$ that generates data and labels (x, z). (x = data, z = label)
  - data points, label, parameters μ, σ, π
- Estimate parameter $\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$
- Infer the labels from the posterior distribution $p(z \mid x, \theta)$.



$p(x, z) = p(z)\,p(x \mid z)$ // multiplication rule.

joint distribution of data and label

① $p(z = k) = \pi_k$ — probability of $z = k$.

$$\begin{cases} \pi_k \geq 0 \\ \sum_k \pi_k = 1 \end{cases}$$ given the axioms of probability

② $p(x \mid z = k) = N(x \mid \mu_k, \sigma_k^2)$ — data; given μ_k, σ_k^2 (from $z = k$). Assume Gaussian.

③ $p(x, z = k) = p(z = k)\,p(x \mid z = k)$ // law of conditional probability

$= \pi_k \, N(x \mid \mu_k, \sigma_k^2)$

joint: $x \cap z = k$; $p(z = k)$; $p(x \mid z = k)$

unknown parameters $\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$

$p(x, z \mid \theta)$ — we don't observe θ.

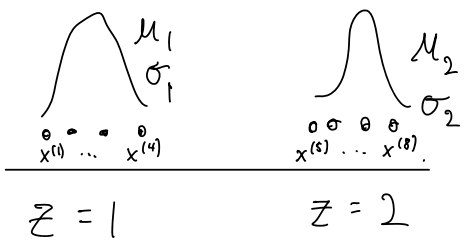
MLE with Complete Information

If we observe both data and labels $\{x_i, z_i\}_{i=1}^n$ (complete information), we can estimate the parameter θ by maximizing the joint probability:

$$\max_\theta \sum_{i=1}^n \log p(x_i, z_i | \theta)$$

joint distribution

Example:



$$\pi_1 = p(z = 1) = \frac{\#(z = 1)}{n} = \frac{4}{8} = \frac{1}{2}.$$

$$\mu_1 = \frac{\sum_{i=1}^n \mathbb{I}(z_i = 1)\, x_i}{\sum_{i=1}^n \mathbb{I}(z_i = 1)} = \frac{\sum \text{values of } x^{(1)} \text{ to } x^{(4)}}{4}$$

$$\sigma_1^2 = \text{var}(\{x_i | z_i = 1\})$$ // variance of the data points with $z_i = 1$.

$$\Rightarrow p(x_i, z_i = k | \theta) = \pi_k \cdot N(x | \mu_k, \sigma_k^2)$$

Joint Log Likelihood Function:

$$\underset{\theta}{\text{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} \underbrace{\mathbb{I}(z_i = k)}_{\text{only fires when } z_i = k} \log P(x_i, z_i = k \mid \theta)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(z_i = k) \log \left(\pi_k N(x_i \mid \mu_k, \sigma_k^2) \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{I}(z_i = k) \left(\log \pi_k + \log N(x_i \mid \mu_k, \sigma_k^2) \right)$$

If we max, then (without further derivation):

$$\pi_k = \frac{\sum_{i=1}^{n} \mathbb{I}(z_i = k)}{n}$$

μ_k = empirical mean of S_k: $\mu_k = \dfrac{\sum_{i=1}^{n} \mathbb{I}(z_i = k) x_i}{\sum_{i=1}^{n} \mathbb{I}(z_i = k)}$

σ_k^2 = empirical variance of S_k: $\text{var}(\{x_i \mid z_i = k\})$

If there is complete information, estimating θ is easy. Use the empirical mean and variance.

Gaussian Mixture Models: Incomplete Information

Clustering can be formulated as parameter estimation in GMM. If we only observe data $\{x_i\}_{i=1}^n$, but missing $\{z_i\}_{i=1}^n$, (incomplete information), we shall estimate θ by maximizing the marginal probability of x_i. (instead of joint $\{x_i, z_i\}$).

Marginal Log-Likelihood Function:

$$\sum_{i=1}^{n} \log p(x_i|\theta) = \sum_{i=1}^{n} \log \sum_{z_i} p(x_i, z_i | \theta) \quad \text{//LOTP}$$

↖ sum over z_i (from 1 to K)

we only observe x_i. z_i = missing information.

$\theta := \{\pi_k, \mu_k, \sigma_k\}_{k=1}^K$, where:

μ_k and σ_k^2 = mean, variance of the k^{th} cluster,

π_k = its percent in the data.

Need numerical method, as there is no closed form solution.

In other words:

known, unknown, unknown

$$p(X|\theta) = \sum_k p(x, z=k \mid \theta)$$

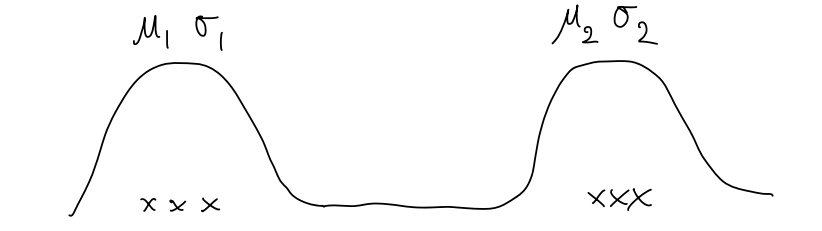
$$= \sum_k \pi_k N(x \mid \mu_k, \sigma_k^2)$$

} Mixture of Gaussian distributions

$$\Rightarrow \underbrace{\sum_{i=1}^{n} \log p(x_i|\theta)}_{\text{want to find } \theta \text{ s.t. this function is maximized.}} = \underbrace{\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \sigma_k^2) \right)}_{\text{objective function that we want to maximize.}}$$

$$\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$$

Example:



$$p(X|\theta) = \frac{1}{2} \exp\left(\frac{-(x-\mu_1)^2}{2\sigma_1^2}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right) \quad \text{// first mixture}$$

$$+ \frac{1}{2} \exp\left(\frac{-(x-\mu_2)^2}{2\sigma_2^2}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_2}\right) \quad \text{// 2nd mixture}$$

No closed form solution. Use EM Algorithm!

SUMMARY: Gaussian Mixture Models:

- Probabilistic model for clustering, in which each cluster is represented by a Gaussian distribution.
- Density function of a GMM is a weighted linear combination of several Gaussian density functions:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \, N(x; \mu_k, \sigma_k^2), \text{ where :}$$

1) $\theta = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$

2) $\{\pi_k\}_{k=1}^{K}$ = set of mixture weights that satisfy the axioms of probability: $\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0 \;\; \forall k$.

3) $N(x; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)$.

each of these is called "component" of $p(x|\theta)$, which corresponds to a cluster

Draw latent label z: $p(z=k|\theta) = \pi_k$ // draw $z=k$ with probability ω_k.

Draw observation X: $p(x=x \mid z=k, \theta) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(\frac{-1}{2\sigma_k^2}(x-\mu_k)^2\right)$

// If $z=k$, X is a Gaussian with mean μ_k, variance σ_k^2.

Joint Probability of X and Z: What is the probability of observation x belonging to cluster k? used for complete problems with known X and Z (data and labels).

$$P(X=x, Z=k \mid \theta) = P(Z=k \mid \theta) \cdot P(X=x \mid Z=k, \theta)$$
$$= \pi_k \times N(x; \mu_k, \sigma_k^2).$$

Marginal Distribution of X: can calculate by summing over Z. used for problems without known Z (labels).

$$P(X=x \mid \theta) = \underbrace{\sum_{k=1}^{K}}_{k=1} P(X=x, Z=k \mid \theta) = \sum_{k=1}^{k} w_k \times N(x; \mu_k, \sigma_2^k)$$

summing over all values of k in $Z=k$.

$\therefore$ This is the density function of the GMM.

How to infer cluster ID of an observation X, based on the posterior distribution:

$$P(Z=k \mid X=x, \theta) = \frac{P(X=x, Z=k \mid \theta)}{P(X=x \mid \theta)}$$
$$= \frac{w_k \times N(x; \mu_k, \sigma_k^2)}{\sum_{l=1}^{K} w_l \times N(x; \mu_l, \sigma_l^2)}$$

EM Algorithm: iterative method for maximizing the marginal likelihood (since we don't observe the latent variable z):

$$\max_{\theta} \sum_{i=1}^{n} \log P(x_i \mid \theta)$$

1) Initialize $\theta_0 = \{w_{k,0}, \mu_{k,0}, \sigma^2_{k,0}\}_{k=1}^{K}$ (random guess)

2) Denote $\gamma_{ik} = P(z^{(i)} = k \mid X = x^{(i)}; \theta)$ = posterior probability for the i^{th} data point belonging to cluster k.

3) Perform until convergence:

i) E-step: Fixing θ, update $\{\gamma_{ik}\}$

Given θ_t, "impute" the missing labels by drawn samples from the posterior distribution,

$$z_i \sim \underbrace{P(z \mid x_i; \theta_t)}_{\text{posterior distribution}}$$

$$\gamma_{ik;\,t+1} = P(z = k \mid X = x^{(i)}; \theta_t)$$

(i^{th} data point, k^{th} cluster)

$$= \frac{w_{k,t} \cdot N(x^{(i)}; \mu_{k,t}, \sigma^2_{k,t})}{\sum_{l=1}^{K} w_{l,t}\, N(x^{(i)}; \mu_{l,t}, \sigma^2_{l,t})}$$

// k^{th} cluster for one data point $x^{(i)}$.

// sum over all the clusters, for one data point $x^{(i)}$.

where

$\theta_t = \{ w_{k,t}, \mu_{k,t}, \sigma_{k,t} \}_{k=1}^{K}$ = values at t^{th} iteration

$\gamma_{ik,t}$ = posterior probability @ t^{th} iteration.

This step calculates the posterior probability of the cluster IDs of each point, assuming θ_t is the true parameter. For each data point $x^{(i)}$, the sum of γ_{ik} over all clusters = 1.

ii) M-Step: Fixing $\{\gamma_{ik}\}$, update θ:

update θ by maximizing the expected joint likelihood.

$$l(\theta) \triangleq \theta^{t+1} = \arg\max_{\theta} \sum_{i=1}^{n} \underset{z_i \sim p(\cdot | x_i, \theta_t)}{E} \left[\log p(x_i, z_i | \theta) \right]$$

$$\Rightarrow l(\theta) = \sum_{i=1}^{n} \log p(x_i | \theta) = \underbrace{\sum_{i=1}^{n} \sum_{k=1}^{k} \gamma_{ik}^{t} \log p(x_i, z_i = k | \theta)}_{\text{max this.}}$$

parameter updates (provided without derivation):

updating mean (for cluster k): $\mu_{k,t+1} = \dfrac{\sum_{i=1}^{n} \gamma_{ik,t+1} x^{(i)}}{\sum_{i=1}^{n} \gamma_{ik,t+1}}$

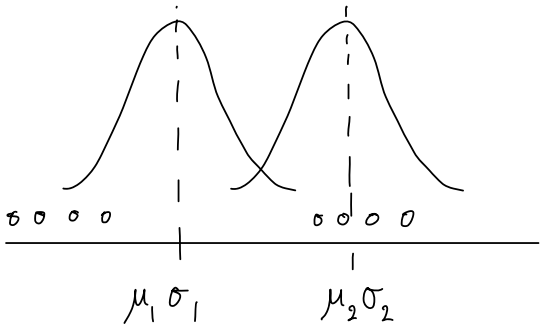
updating variance: $\sigma^2_{k,t+1}$ (for cluster k) $= \dfrac{\sum_{i=1}^{n} \gamma_{ik,t+1}\left(x^{(i)} - \mu_{k,t+1}\right)^2}{\sum_{i=1}^{n} \gamma_{ik,t+1}}$

updating weights: $w_{k,t+1}$ (for cluster k) $= \dfrac{\sum_{i=1}^{n} \gamma_{ik,t+1}}{\sum_{i=1}^{n}\sum_{k=1}^{K} \gamma_{ik,t+1}}$

where mean μ_k and variance σ_k^2 of the k-th cluster equals the empirical mean and variance of the data, when each data point is "weighted" by posterior probability of belonging to the k^{th} cluster, based on estimation $\gamma_{ik,t+1}$ from the current iteration.

converge to local optimum.

Implementation using previous example:



$$\theta_t = \left[\pi_k^t , \mu_k^t , \sigma_k^t \right]_{k=1}^{K}$$

After t iterations, these are the values.

$$p\left(\overbrace{z^{(i)} = k}^{A} \mid \overbrace{x^{(i)}, \theta_t}^{B} \right) = \frac{p\left(\overbrace{x^{(i)}, z^{(i)} = k}^{A \cap B} \mid \theta_t \right)}{\sum_l p\left(x^{(i)}, z^{(i)} = l \mid \theta_t \right)}$$

// law of conditional probability: $p(A \mid B) = \frac{p(A \cap B)}{p(B)}$

$$= \frac{\pi_k \, N\left(x^{(i)} \mid \mu_k^t \; \sigma_k^t \right)}{\sum_l \pi_l \, N\left(x^{(i)} \mid \mu_l^t \; \sigma_l^t \right)}$$

So, for example:

$$p\left(z^{(i)} = 1 \mid x^{(i)}, \theta_t \right) = \frac{\pi_1 \exp\left(- \frac{(x^{(i)} - \mu_1)^2}{2\sigma_1^2} \right) \frac{1}{\sqrt{2\pi}\sigma_1}}{\sum_l \left\{ \pi_l \exp\left(- \frac{(x^{(i)} - \mu_l)^2}{2\sigma_l^2} \right) \frac{1}{\sqrt{2\pi}\sigma_l} \right\}}$$

Derivation of EM algorithm:

- Consider general "imputation distributions" $\rho(z|x)$ ← arbitary distribution?
- we can construct a **tight lower bound** $LB(\theta, \rho)$ of the **marginal log likelihood function**:

$$l(\theta) \geq LB(\theta, \rho), \quad \forall \rho \quad \text{and} \quad l(\theta) = \max_{\rho} LB(\theta, \rho)$$

- **Maximizing $l(\theta)$ is then equivalent to maximizing $LB(\theta, \rho)$**:

$$\max_{\theta} l(\theta) = \max_{\theta, \rho} LB(\theta, \rho)$$

- Optimizing θ and ρ alternatively (**coordinate descent**) yields EM algorithm.

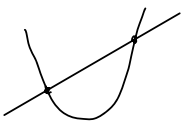
Dive In:

$$l(\theta) \triangleq \max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta)$$

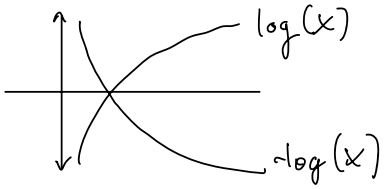
$$= \max_{\theta} \sum_{i=1}^{n} \log \sum_{z_i} P(x_i, z_i|\theta)$$ // marginal distribution = sum of joint distribution

want to move the summation outside of log ⇒ use Jensen's Inequality:

for convex function f: $f(E[x]) \leq E[f(x)]$.



log is concave function → -log is a convex function.



$$\Rightarrow \sum_{i=1}^{n} \log \sum_{z_i} \frac{p(x_i, z_i \mid \theta)}{p(z_i \mid x_i)} p(z_i \mid x_i)$$

↑ multiply both num and dem (equivalent) by arbitrary distribution p.

$$= \sum_{i=1}^{n} \log \underbrace{E_{z_i \sim p(\cdot \mid x_i)} \left[\frac{p(x_i, z_i \mid \theta)}{p(z_i \mid x_i)} \right]}_{f(E[X])}$$

// rewrite summation as expectation over p.

$$\geq \sum_{i=1}^{n} \underbrace{E_{z_i \sim p(\cdot \mid x_i)} \log \left(\frac{p(x_i, z_i \mid \theta)}{p(z_i \mid x_i)} \right)}_{E[f(X)]} \triangleq LB(\theta, p).$$

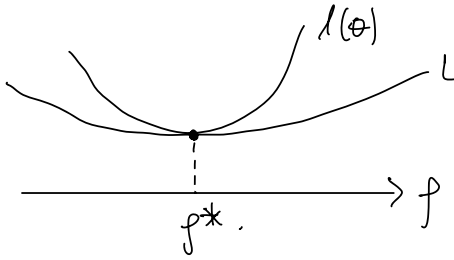
If $p(z \mid x, \theta) = p(z \mid x, \theta)$ then

$$\frac{p(x_i, z_i \mid \theta)}{p(z_i \mid x_i)} = \underbrace{\frac{p(x_i, z_i \mid \theta)}{p(z_i \mid x_i, \theta)}}_{\text{posterior}} = \underbrace{p(x_i \mid \theta)}_{\text{marginal}}$$

$$\Rightarrow LB(\theta, p^*) = \sum_{i=1}^{n} \underbrace{E_{z \sim p(\cdot \mid x_i)}}_{\text{the expectation does not depend on } z_i \text{ anymore, thus can be removed.}} \log p(x_i \mid \theta).$$

$$= \sum_{i=1}^{n} \log p(x_i \mid \theta) = L(\theta).$$

By Jensen Inequality, we showed that LB is always a Lower Bound of $\ell(\theta)$ for any ρ. If we choose ρ to be to be the posterior distribution, then the lower bound matches the marginal likelihood.



$$\left.\begin{array}{l} \ell(\theta) \geq LB(\theta) \quad \forall\, \theta, \rho. \end{array}\right\} \Rightarrow \boxed{\ell(\theta) = \max_{\rho} LB(\theta, \rho)}$$

$\Rightarrow \max_{\theta, \rho} LB(\theta, \rho)$: Use coordinate descent:

1) Initialize θ_0.
2) Iterate:

Fix θ_t : $\rho_{t+1} = \underbrace{\arg\max_{\rho} LB(\theta_t, \rho)}_{p(z \mid x, \theta_t)}$ // expectation step

Fix ρ_{t+1} : $\theta_{t+1} = \arg\max_{\theta} LB(\theta, \rho_{t+1})$ // maximization step.

To see that:

$$LB(\theta, \rho_{t+1}) = \sum_{i=1}^{n} E_{z_i \sim \rho(\cdot \mid x_i)} \Big[\log p(x_i, z_i \mid \theta) - \underbrace{\log \rho(z_i \mid x_i)}_{\text{constant.}} \Big]$$

$$= \underbrace{\sum_{i=1}^{n} E_{z \sim \rho(\cdot \mid x_i)} \big[\log (x_i, z_i \mid \theta) \big]}_{\text{Expected joint likelihood,}} + \text{constant,}$$

Properties: this procedure:

① Monotonically decrease $LB(\theta, \rho)$, and $\ell(\theta)$.

② converge to local optima of $\ell(\theta)$.