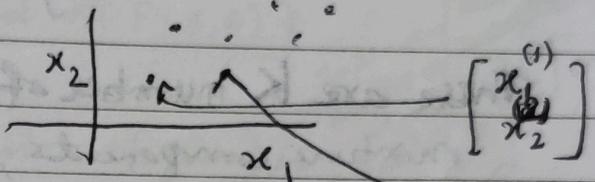


We studied K-means clustering to group the given data $\mathbf{x} = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, where $x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$ (earlier we may have used $\underline{x} \approx \underline{x}^{(i)}$) → Notation changed. Sorry is the i^{th} data vector. Note that the or feature vector entire data can be considered as \overrightarrow{N} dimensional \mathbf{X} s.

	(1)	(2)	(m)
$x_1 :$	x_1	x_1	\dots
	(1)	(2)	(m)
$x_2 :$	x_2	x_2	\dots
	(1)	(2)	(m)
\vdots	(1)	(2)	(m)
$x_n =$	x_n	x_n	\dots

If we consider 2D then each data point can be represented in a 2D plot-



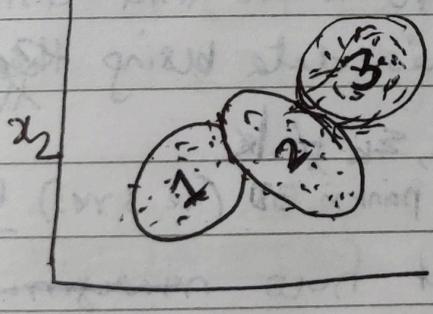
Can we have a probabilistic interpretation for clustering? i.e. how can we model

$$p(x) = p_{x_1, \dots, x_n}(x_1, x_2, \dots, x_n)$$

to reflect our intuition that the points (that we assign) remain closer to their cluster centres (centroids)

\downarrow
joint
density
 f_{X_1, \dots, X_n}

~~Consider~~ Consider the plot below
We



We can see that the given points seem to form 3 clusters. We cannot ~~fit~~ model such a data by a simple distribution such as Gaussian (with some mean & some variance)

One can think of modeling each region with a distinct distribution. We may think of mixtures of Gaussians (GMMs) to model such a data.

① Continue

How can we do it? We have to do it using unlabelled data \mathbf{x} (Unsupervised)

Since \mathbf{x} is considered as mixture of Gaussians GMM has the following pdf for \mathbf{x}

$$p(\mathbf{x}) = P_{\underbrace{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n}_{\sum}} (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) = \sum_{k=1}^{dk} w_k N(\mathbf{x} / \mu_k, \Sigma_k)$$

$$\text{If } K=3 \quad p(\mathbf{x}) = w_1 N(\mathbf{x} / \mu_1, \Sigma_1) + w_2 N(\mathbf{x} / \mu_2, \Sigma_2) \\ + w_3 N(\mathbf{x} / \mu_3, \Sigma_3)$$

There are K number of Gaussian distributions called mixture components, μ_k and Σ_k represent mean and covariance matrix of k th component. w_k represent mixture weight's telling us about how much each component

contributes to final GMM represented by $p(\mathbf{x})$. One has to have $w_k \geq 0$, $\forall k$ and $\sum_{k=1}^K w_k = 1$ to ensure that $p(\mathbf{x})$ is a valid pdf.

Now given \mathbf{x} our objective is to find GMM parameters to represent this data using ~~these~~ ^{the} parameters

Constituent parameters are $w_k, \mu_k, \Sigma_k \forall k$

Example if $K=3$ & each data point is 3D (i.e. 3 rows) then we have to find w_1, w_2, w_3 , μ_1, μ_2, μ_3 and $\Sigma_1, \Sigma_2, \Sigma_3$ all are 3×3 matrices.

To do this let us look at how marginal GMM can be expressed as the marginal (of a joint Pdf)

We know that $p(\mathbf{x}, z) = \underbrace{p(z)}_{z} p(\mathbf{x}|z)$ \mathbf{x} is a random vector. \mathbf{x} has multivariate Pdf

where z is a discrete rv taking values $k=1$ to K .

Denote $w_k = P_z(z=k)$

since we are dealing with GMM, assume conditional distributions as Gaussians

$$\text{i.e. } p(x|z=k) = N(x|\mu_k, \Sigma_k)$$

We can then write

$$P(x) = \sum_{k=1}^K p(z) p(x|z) = \sum_{k=1}^K w_k p(x|\mu_k, \Sigma_k)$$

which is GMM

[See that $P(x)$ is marginal of $p(x, z)$]

data point

As an example (from the previous figure)

$$p(x|z=1) = N(x|\mu_1, \Sigma_1)$$

$$p(x|z=2) = N(x|\mu_2, \Sigma_2)$$

$$p(x|z=3) = N(x|\mu_3, \Sigma_3)$$

$$\text{Then } p(x) = P(\text{region 1}) N(x|\mu_1, \Sigma_1)$$

$$+ P(\text{region 2}) N(x|\mu_2, \Sigma_2)$$

$$+ P(\text{region 3}) N(x|\mu_3, \Sigma_3)$$

Estimation of parameters of GMM $\Theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^K$ with complete data.

Let us consider unrealistic case where we have labels z

$$\text{Define } \mathcal{D} = \{x_n^{(n)}\}_{n=1}^m \text{ or } \mathcal{D}' = \{x_n^{(n)}, z_n^{(n)}\}_{n=1}^m$$

incomplete data → Complete data

$x_n^{(1)}$ represents first point

$x_n^{(2)}$ 2nd point

$x_n^{(m)}$ mth data point

$z_n^{(1)} = 1$ means

first data pt of C

region (cluster)

If D' is ~~given~~^{ie.} & complete data is given ML
 estimation of θ can be obtained by maximizing the
~~obtained~~ considering all data points

$$P(x^{(n)}, z^{(n)})$$

i.e. maximize $\prod_{n=1}^m P(x^{(n)}, z^{(n)})$ or θ can be estimated

$$\text{by decrease } \max_{\theta} \log \left[\prod_{n=1}^m P(x^{(n)}, z^{(n)}) \right]$$

$$= \max_{\theta} \sum_{n=1}^m \log P(x^{(n)}, z^{(n)})$$

$$= \max_{\theta} \sum_{n=1}^m \log p(z^{(n)}) p(x^{(n)}|z^{(n)})$$

$$= \max_{\theta} \sum_{k=1}^K \sum_{n: z^{(n)}=k} \log P(z^{(n)}) p(x^{(n)}|z^{(n)}) \quad \text{--- (2)}$$

i.e. group the data by its $z^{(n)}$ values (Ex: If there are 30 data points and 10 each $\in E$ to each cluster with number of clusters (regions) = $K=3$, then $\log p(z^{(n)}) p(x^{(n)}|z^{(n)})$ appears 10 times (for 10 data points each). There are total of 30 terms.

Let $r_{nk} \in \{0, 1\}$ be a binary variable, that indicates column $z^{(n)} = k$. i.e., if $\cancel{x^{(n)}}$ is n th data point $\cancel{\in}$ cluster k , ~~then~~ $r_{nk} = 1$ else 0. Note that $r_{nk} = P(z^{(n)}=k)$

We can write

$$\sum_n \log P(x^{(n)}, z^{(n)}) = \sum_k \sum_n r_{nk} \log \{ P(z^{(n)}=k) \} + p(x^{(n)}|z^{(n)}=k) \}$$

$$= \sum_k \sum_n \gamma^{nk} [\log w_k + \log P(x^{(n)} / \mu_k, \Sigma_k)]$$

(2)

$$= \sum_{k=1}^K \sum_{n=1}^m \gamma^{nk} \log w_k + \sum_{k=1}^K \left\{ \frac{\sum_{n=1}^m \gamma^{nk} \log P(x^{(n)} / \mu_k, \Sigma_k)}{\sum_{n=1}^m \gamma^{nk}} \right\}$$

One can now write the pdf for $P(x^{(n)} / \mu_k, \Sigma_k)$ and

after differentiating and equating to 0, we get

(w.r.t parameters) [for maximizing $\prod_{n=1}^m P(x^{(n)}, z^{(n)})$ or

for maximizing $\sum \log P(x^{(n)}, z^{(n)})$]

$$w_k = \frac{\sum_n \gamma^{nk}}{\sum_{k=1}^K \sum_n \gamma^{nk}}$$

\downarrow a vector (n dimensional)

$$\mu_k = \frac{1}{\sum_n \gamma^{nk}} \sum_n \gamma^{nk} x^{(n)}$$

this is known.

$$\Sigma_k = \frac{1}{\sum_n \gamma^{nk}} \sum_n \gamma^{nk} (x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T$$

$n \times n \quad n \times 1 \quad 1 \times n \quad \rightarrow n \times n$

matrix

~~we see~~ $w_k = P(z^{(n)} = k)$ represents fraction of total

~~a part~~ number of data points which are assigned to cluster $k \rightarrow$ i.e. probability of each cluster ($w_1 + w_2 + \dots + w_K = 1$)

Note that

$$\sum_k \sum_{n=1}^m \gamma^{nk} = m$$

$$w_1 = \frac{m_1}{m}, w_2 = \frac{m_2}{m}, \dots$$

$$w_3 = \frac{m_3}{m} \dots$$

$$\text{So } w_1 + w_2 + \dots + w_K = 1$$

because $m_1 + m_2 + m_3 + \dots = m$

p_k
is given by the
feature $x^{(n)}$. In this case γ^{nk} is
assumed to be
known i.e. both its
values in this

μ_k Vector represents mean of all data points which are assigned to cluster k (i.e. whose $z^{(n)} = k$)

Ex Suppose the $K=3$ & each data point is of dimension 2 ($i.e., n=2$)

$$x_1 = \begin{bmatrix} x_1^{(1)} \\ x_1^{(2)} \end{bmatrix}$$

$$x_2 = \begin{bmatrix} x_2^{(1)} \\ x_2^{(2)} \end{bmatrix}$$

If $m=30$ and if first 10 points

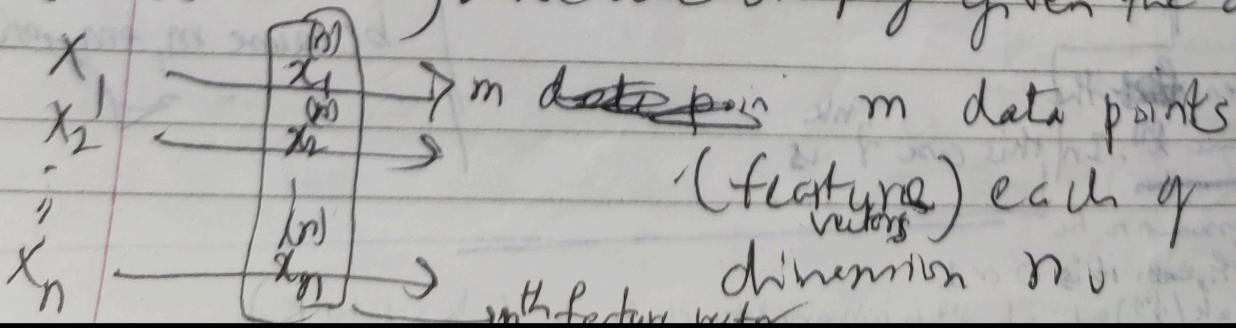
are of each rv (x_1 or x_2) \in Cluster 1

i.e. $w_1 = 10/30 = \frac{1}{3}$, mean $M_k = \begin{bmatrix} m_{x_1} \\ m_{x_2} \end{bmatrix} = \begin{bmatrix} \text{average of 10} \\ \dots \\ \dots \end{bmatrix}$

Covariance matrix ~~Σ~~ $\Sigma_1 = \Sigma_k$ will be size 2×2 computed by using the first 10 points of x_1 & x_2

Important Point: The above clustering method is unrealistic since we have assumed which data points \in to which cluster. For example for the above example we know that first 10 points (10 features) \in cluster 1. In practice which data point \in to which cluster is not known at all. So, the problem of automatic clustering becomes quite difficult.

Let us now consider estimating θ when we do not know $z^{(n)}$ (i.e. which data pts \in to which cluster). We are simply given the data



(4)

~~This~~ A method called EM (expectation maximization) is often used when we do not know $Z^{(n)}$ i.e. when we are given incomplete data (D) & not D^I

Parameter estimation for GMMs: Incomplete data (Use of EM algorithm)

When $Z^{(n)}$ is not given, one can guess γ^{nk}

by using its posterior probability i.e.

$$P(Z_{(n)}^{(n)} = k / x^{(n)}) = \frac{P(x^{(n)} / Z^{(n)} = k)}{P(x^{(n)})} P(Z^{(n)} = k) = \gamma^{nk}$$

→ given n th datapoint $P(x^{(n)})$

i.e. if we do not know that ~~which~~ data point belongs to (k) cluster k , one can guess it by ~~any~~ ^{this} ~~coexisting~~

Writing the above probability expression i.e. the

chance of assigning n th data point to k th cluster given ~~that~~
that point is equal to what is written ~~written~~ on the

RHS of the above expression. In the previous case

We assumed that this ^{was} known. That is if $\gamma^{nk} = 1$ then

$P(Z^{(n)} = k / x^{(n)}) = 1$ else it is zero. Note that this is

not practical since we do not know which data point belongs to
which cluster. It ~~can~~ ~~be~~ may ∞ to any ^{one} of the K clusters.

Hence ~~Gaussian mixture~~ obtaining the GMM parameters

i.e. $w_k, k=1 \dots K$ & μ_k & $\Sigma_k, k=1 \dots K$ is not easy.

If it is unrealistic case where $Z^{(n)}$ is known the problem can
be solved ~~easily~~ easily (since we have ^{the} complete data). ~~The~~ The

Solution is already obtained (see the expressions for w_k, μ_k, Σ_k)

Let us now obtain the same solution with the incomplete data ($z^{(n)}$ not known) by ~~making~~ guessing & arriving at solution using EM.

We ~~can write~~ know that

$$P(z^{(n)} = k/x^{(n)}) = \frac{P(x^{(n)} / z^{(n)} = k) P(z^{(n)} = k)}{\sum_{k'=1}^K P(x^{(n)} / z^{(n)} = k') P(z^{(n)} = k')} = \gamma^{nk}$$

Note that we cannot write $x^{(n)} = k$ since $x^{(n)}$ is a vector.

This is
~~P(x)~~ $P(x^{(n)})$

We have to use another random variable say Z to indicate that a particular data point x belongs to a cluster k

for this,

Now, how do you compute the ~~the~~ posterior? We need to know the parameters which are not known. Let us

start with some parameters ~~(X, C, S)~~ so that we can guess the posterior.

We have

~~Let us write~~ $\gamma^{nk} = P(z^{(n)} = k/x^{(n)})$. We know that

$$\gamma^{n,1} + \gamma^{n,2} + \gamma^{n,3} = P(z^{(n)} = 1/x^{(n)}) + P(z^{(n)} = 2/x^{(n)}) + P(z^{(n)} = 3/x^{(n)}) = 1 \quad (\text{Assuming } K=3) \quad \text{for all } n=1 \text{ to } m$$

We will assume γ^{nk} for each data point (if there are say 30 data points, the total number of γ^{nk} values are $(30 \times 3) = 90$)

With these values expectation of assigning n^{th} data point to each cluster can be written as

$$P(z^{(n)} = 1/x^{(n)}) \log P(x^{(n)} / \mu_1, \Sigma_1) + P(z^{(n)} = 2/x^{(n)}) \log P(x^{(n)} / \mu_2, \Sigma_2) + P(z^{(n)} = 3/x^{(n)}) \log P(x^{(n)} / \mu_3, \Sigma_3)$$

~~log is used here for convenience~~

i.e. it represents ~~on~~ on an average

what is the chance that it belongs to ~~a~~ cluster, which itself depends on $\mu_1 \Sigma_1 \dots \mu_3 \Sigma_3$.

Note that (E) is same as the 2nd term of (1) hence θ can be obtained using (1)

If we do this for all data points, we can write the total probability as

$$Q(\theta_{old}, \theta) = \sum_{k=1}^m \sum_{n=1}^N P(z^{(n)}=k/x^{(n)}, \theta_{old})$$

Page 5

$$\log P(x^{(n)}/z^{(n)}, \theta) - (E)$$

Here
↓
E

writing above is called expectation step

where θ i.e. $\mu, \Sigma, \dots, \pi_k$ are unknown

We can then maximize Q by choosing appropriate θ which is done iteratively.

Hence the name EM. (i.e., we arrive at the final θ by using expectation step and maximization step. That is we obtain the parameters

$$E(Y) = \sum y_i p(y_i)$$

↓
posterior

likelihood

by maximizing (over θ) the

expectation expectation expectation under Posterior of $P(z/x)$

taken w.r.t likelihood $\log(x/z)$

↑ after convergence

Note that finally we arrive at the solution as (1)

i.e. we succeed in making $\gamma^{nk} \in [0, 1]$

depending or by making γ^{nk} as 1 for those data points belonging to clusters 1, 2, ..., k iteratively.

(i.e. If ~~first~~ in 30 data points first 10 belong to cluster 1 then $\gamma^{n,1} = 1$ for $n = 1, 2, \dots, 10$ and

$\gamma^{n,1} = 0$ for $n = 11 \text{ to } 30$ (likewise for other clusters)

How we perform EM? Once we assume γ^{nk} for all n, k we already know the expressions for getting w_k, μ_k, Σ_k for all k (see (1)). Use (1) to get updated

~~wk, μk, Σk~~. Repeat until we get convergence. Then use these to get updated γ^{nk} & form E step to maximize θ

EM to get $\theta = \omega_k, \mu_k = \mu_k$.

EM to get $\omega_k, \Sigma_k, \mu_k$ iteratively.

E Step (Expectation step)

Write E step as E given in previous page.

Here $y^{nk} = P(z^{(n)}=k/x^{(n)}, \Theta)$ are assumed. Θ corresponds to are unknown.

M step (Maximization step)

Maximize E for all θ . This leads to expressions given in ① on page ③. Using these expressions we get updated parameters $\theta = \omega_k, \mu_k, \Sigma_k$

Go back to E step after computing updated y^{nk} using expression given in page ④. Use

$$P(x^{(i)} / z^{(n)}=k) = P_{x_1, x_2, \dots, x_n}(x_1^{(i)}, x_2^{(i)}, \dots, x_n^{(i)} | z^{(n)}=k)$$

Here $\mu_k = \begin{pmatrix} \mu_k^{(1)} \\ \mu_k^{(2)} \\ \vdots \\ \mu_k^{(n)} \end{pmatrix}$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \left| \sum_k \right|^{\frac{1}{2}} \exp \left(-\frac{1}{2} (x^{(i)} - \mu_k)^T \sum_k^{-1} (x^{(i)} - \mu_k) \right)$$

If $n=2$, (two r.v.s x_1, x_2 i.e., 2D feature vectors)

$$\rightarrow P(x^{(i)} / z^{(n)}=k) = P_{x_1, x_2}(x_1^{(i)}, x_2^{(i)} / z^{(n)}=k) = \frac{1}{(2\pi)^{\frac{1}{2}} |\Sigma_k|^{\frac{1}{2}}} \times \exp \left(-\frac{1}{2} (x_1^{(i)} - \mu_k) [x_2^{(i)} - \mu_k]^T \Sigma_k^{-1} (x_1^{(i)} - \mu_k) \right)$$