7/24/2023

K- MEANS

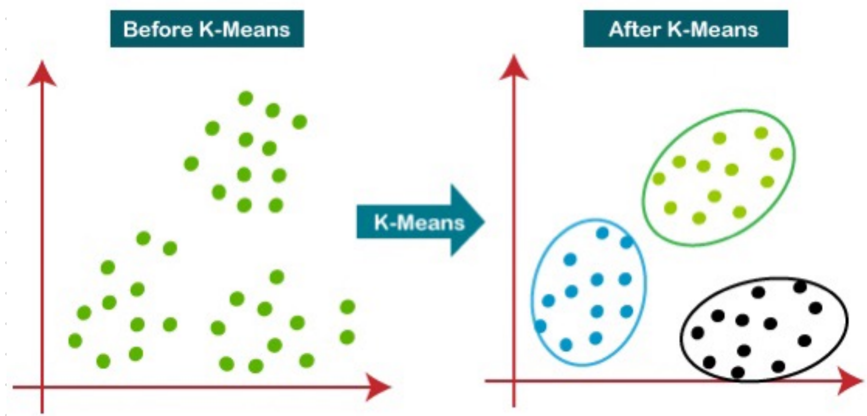
INGREDIENTS

TRAINING SET $\{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$, $x^{(i)} \in \mathbb{R}^d$

NUMBER OF CLUSTERS : K CLUSTERS

$\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^d$: K CENTROIDS

$\{C^{(i)}\}^K$, $C^{(i)} \in \{1, 2, \dots, K\}$ REPRESENTS WHICH CLUSTER DOES $x^{(i)}$ BELONG TO



K-MEANS ALGORITHM (CLASS.)

INITIALIZATION

1. INITIALIZE CENTROIDS $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^d$ RANDOMLY

ITERATION

2. REPEAT UNTIL CONVERGENCE:

FOR EACH i SET

$$c^{(i)} := \arg\min_j \| x^{(i)} - \mu_j \|^2$$

HARD ASSIGNMENT
(PICK CENTROID W/ SMALLEST DISTANCE)

FOR EACH j, SET $\{1, \dots, k\}$

$$\mu_j := \frac{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{c^{(i)} = j\}}$$

(RE-SET THE CENTROID)

CONVERGENCE CONDITION

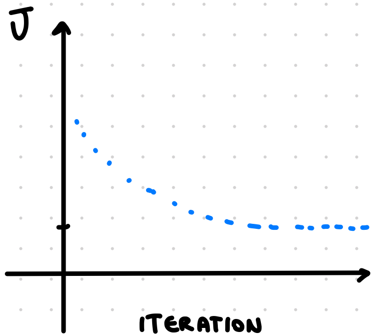
NOT GUARANTEED

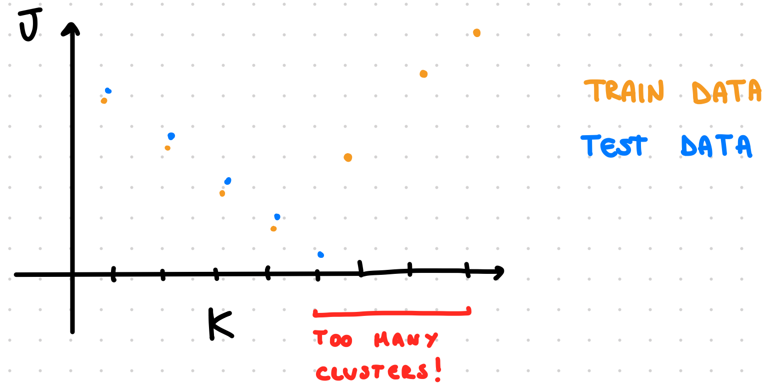
LET'S DEFINE THE DISTORTION FUNCTION J

$$\boxed{J(c,\mu) = \frac{1}{n}\sum_{i=1}^{n} \| x^{(i)} - \mu_{c^{(i)}} \|^2}$$

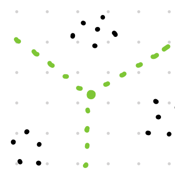
IT IS GUARANTEED TO MONOTONICALLY DECREASE AND, AT SOME POINT, <u>IT WILL CONVERGE</u>

<u>EX</u>





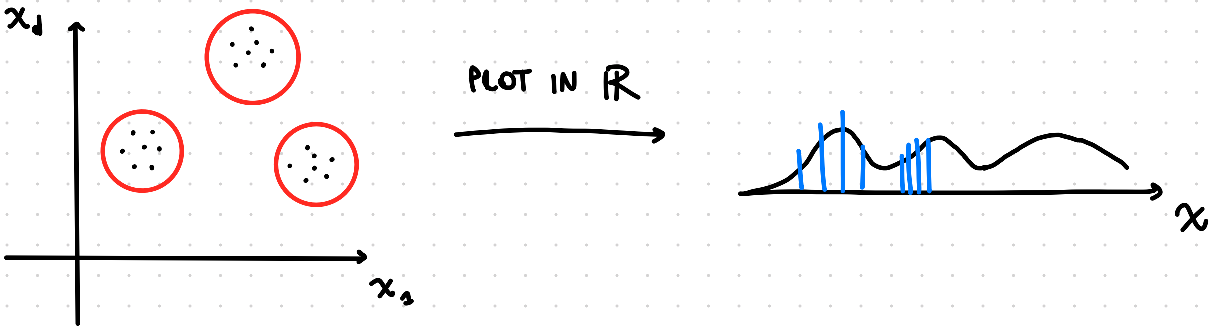
HOW DOES THE CLASSIFIER LOOK LIKE?



STILL LINEAR!
(L_2 DISTANCE)

GAUSSIAN MIXTURE MODEL (CLASS.)

WE WANT A DENSITY ESTIMATION IN A d-DIMENSIONAL SPACE



GMM IS AN APPROACH TO DENSITY EST.
IT ASSUMES THAT DATA IS IN A CLUSTER IS DEFINED BY A PROBABILITY FUNCTION (MULTINOMIAL)

$$Z \sim \text{Multinomial}(\phi) : \quad \phi_j \geq 0, \quad \sum_{j=1}^{k} \phi_j = 1$$

PROB. OF A DATA-POINT TO BELONG TO A CLUSTER
DEF. CLUSTER IDENTITY

$$x \mid Z = j \sim N(\mu_j, \Sigma_j) : \quad \mu_j \in \mathbb{R}^d, \quad \Sigma_j \in \mathbb{R}^{d \times d}_{++}, \quad j \in \{1, \dots, k\}$$

WE TAKE PARAMS AND SAMPLE IT FROM THE OBTAINED j-TH DISTRIBUTION

WE DEFINE A LIKELIHOOD

$$\ell(\phi, \mu, \Sigma) = \sum_{i=n}^{n} \log P(x^{(i)}; \phi, \mu, \Sigma) =$$

$$= \text{MARGINALIZATION} = \sum_{i=1}^{n} \log \sum_{j=1}^{K} P(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma) =$$

$P(x) = \sum P(x, z)$

$$= \boxed{\sum_{i=1}^{n} \log \sum_{j=1}^{K} P(x^{(i)} \mid z^{(i)}; \phi, \mu, \Sigma)\, P(z^{(i)}; \phi)}$$

THERE IS NO ANALYTICAL SOLUTION. HERE THE E.M. (EXPECTATION MAXIMIZATION) COMES IN CLUTCH. (WE DO NOT HAVE A CLOSED-FORM SOL, NOR CAN APPLY GRAD. DESC)

EXPECTATION MAXIMIZATION

WE HAVE A PROBABILISTIC MODEL

$$P(\underbrace{x}_{\text{OBSERVED}}, \underbrace{Z}_{\text{NOT OBSERVED}}, \underbrace{\theta}_{\text{PARAMS (EX } \phi, \mu, \Sigma)})$$

θ : PARAMETERS

WILL NOT CHANGE GIVEN THE TRAIN SET

x : OBSERVED DATA

Z : NON OBSERVED DATA

ON NEW EXAMPLE WE GET TO EST. NEW Z

THE E.M. ALGORITHM IS A RECIPE FOR A GIVEN MODEL

E.M. ALGORITHM

REPEAT UNTIL CONVERGENCE:

E-STEP : $\forall i, j$ SET

EXAMPLES CENTROIDS

$$w_j^{(i)} = P(x^{(i)} \mid z^{(i)}; \phi, \mu, \Sigma)$$

BASICALLY A SOFT ASSIGNMENT OF EX TO A CENTROID (UPDATE BELIEFS OF EACH EX UNDER A CERTAIN CEN)

$P(z)$ - PRIOR

$P(x)$ - DATA

$P(x, z)$ - JOINT (MODEL)

$P(x \mid z)$ - LIKELIHOOD

$P(z \mid x)$ - POSTERIOR

RE-SOFT-ASSIGN EACH EXAMPLE TO A CLUSTER CENTROID

M-STEP : UPDATE PARAMETERS

(**)

$$\phi_j = \frac{1}{n}\sum_{i=1}^{n} w_j^{(i)}, \quad \mu_j = \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}}, \quad \Sigma_j = \sum_{i=1}^{n} \frac{w_j^{(i)}(x^{(i)}-\mu_j)(x^{(i)}-\mu_j)^T}{\sum_{i=1}^{n} w_j^{(i)}}$$

FOR EACH x ASSIGNED TO CENT. z WE ARE CALC. $P(x \mid z)$ BY CALC. THE PARAMS

EST. OF MEAN, IF LESS PROB. OF BELONGING TO A CERTAIN CENTROID = LESS THE WEIGHT

GENERALIZED E.M.

WE HAVE A PROBABILISTIC MODEL W/ LATENT PARAMS (UNSUPERVISED LEARNING)

IF WE HAVE A MODEL, W/ WELL DEFINED LATENT / UNLATENT PARAMS, WE CAN USE E.M.

JENSEN'S INEQUALITY (MATH TOOL)

f IS CONVEX

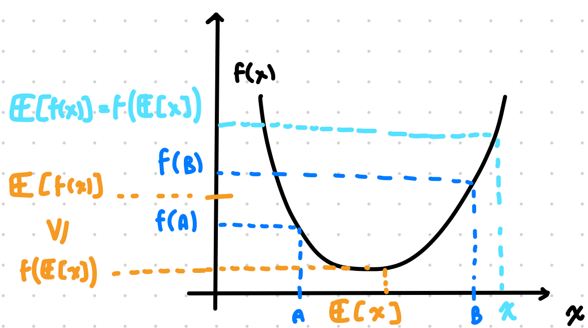
x IS A RANDOM VARIABLE

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$$

IF f IS STRICTLY CONVEX

$$\mathbb{E}[f(x)] = f(\mathbb{E}[x]) \iff x = \mathbb{E}[x] \text{ w.p. } 1 \quad (x \text{ IS A CONST})$$

ex



$f'' > 0 \quad \forall x$

GIVEN A TRAINING SET $\{x^{(1)}, \dots, x^{(n)}\}$

AND A PDF $P(x;\theta) = \sum_z P(x,z;\theta)$

WE WANT TO MAXIMIZE (JUST ONE EXAMPLE FOR EZ NOTATION)

$$\ell(\theta) = \log P(x;\theta) =$$

$$\ell(\theta) = \log \sum_z P(x,z;\theta) = (*)$$

LET $Q(z) > 0$ $\forall z$, $\sum_z Q(z) = 1$ (SOME DISTRIBUTION OVER z)

$$(*) = \log \sum_z Q(z) \frac{P(x,z;\theta)}{Q(z)}$$

$$= \log \mathbb{E}_{z\sim Q}\left[\frac{P(x,z;\theta)}{Q(z)}\right]$$

log IS CONCAVE, THEN -log IS CONVEX. FROM JENSEN'S INEQUALITY

$$\mathbb{E}[f(x)] > f(\mathbb{E}[x])$$

FLIPPED BECAUSE log IS CONCAVE!

$$\log \mathbb{E}_{z \sim Q}\left[\frac{P(x, z;\theta)}{Q(z)}\right] \geq \mathbb{E}_{z \sim Q} \log \frac{P(x, z; \theta)}{Q(z)}$$

$$\log P(x;\theta) \geq \mathbb{E}_{z \sim Q} \log \frac{P(x, z; Q)}{Q(z)}$$

EVIDENCE   LOWER BOND

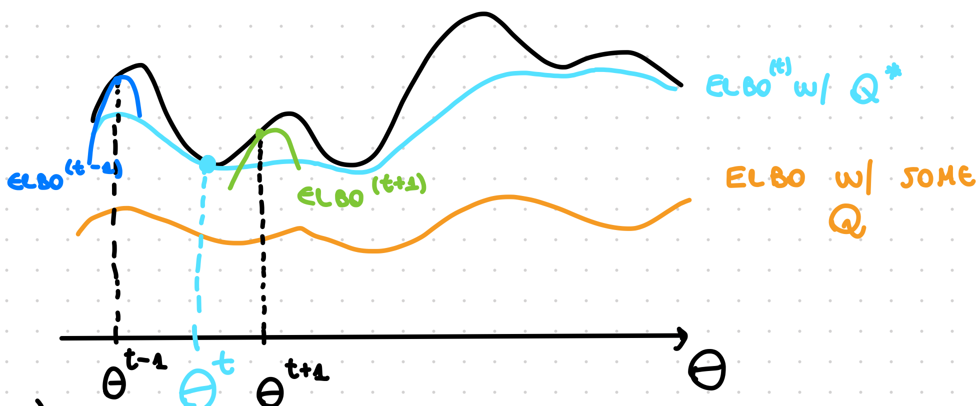
ELBO

EVIDENCE LOWER BOND

WE NEED TO TAKE AN ELBO AS CLOSE TO EVIDENCE AS POSSIBLE

ex

WE NEED A Q^* THAT MAKES ELBO AS CLOSE TO $P(x;\theta)$ AS POSSIBLE (FOR ANY GIVEN VALUE OF θ)



WE CHOOSE Q (DESIGN)

TO DO THAT, WE ASK OURSELVES, FOR WHICH VALUE OF Q ?

$$\log P(x;\theta) = \mathbb{E}_{z\sim Q} \log \frac{P(x,z;Q)}{Q(z)}$$

WE WANT $\frac{P(x,z;\theta)}{Q(z)} = K$ HAS TO BE A <u>CONSTANT</u>

WE KNOW THAT $Q(z) \propto P(x,z)$

$$Q^*(z) = \frac{\cancel{K} P(x,z)}{\sum_z \cancel{K} P(x,z)} = P(z|x)$$

NORMALIZE ←

FOR THIS SPECIFIC CHOICE OF Q^* THE RESULTING ELBO FOR THE CURRENT ESTIMATE OF θ^t WILL BE TIGHT.

WE CAN MAXIMIZE ELBO INSTEAD OF OG PDF. Q^* IS THE POSTERIOR DIST. FUN. $P(z|x)$

WE OPTIMIZE ELBO AND UPDATING ON θ^t WE HAVE MADE PROG. (ON OG OBJECTIVE)

THEN NEW ELBO, NEW THETA, NEW PROGRESS... AND SO ON

IT'S LIKE DRIVING IN THE DARK, W/ A GPS

WE CONTRUCT AN ELBO F.E. θ WHERE $ELBO(\theta) = \log P(x;\theta)$. SO

$$\ell(\theta^{(t+1)}) \overbrace{\geq}^{\text{JENSEN'S INEQUALITY}} ELBO^{(t)}(\theta^{(t+1)}) \underbrace{\geq}_{\text{DEFINITION OF ARGMAX}} ELBO^{(t)}(\theta^{(t)}) \underbrace{=}_{\substack{\text{JENSEN'S INEQUALITY}\\ \text{COROLLARY}}} \ell(\theta^{(t)})$$

$$\Rightarrow \ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$$

FROM HERE, IT IS TRIVIAL TO STATE

$$\boxed{\theta^{(t+1)} = \arg\max_{\theta} ELBO^{(t)}(\theta)}$$

CAN BE CLOSED-FORM θ OR CAN BE OBTAINED NUMERICALLY (EX G.D.)

RECAP : E.M. RECIPE

- GIVEN $\begin{cases} P(x, z; \theta) \\ x \text{ OBSERVED} \\ z \text{ LATENT} \\ \theta \text{ ESTIMATE} \end{cases}$

- INITIALIZE θ RANDOMLY

- REPEAT UNTIL CONVERGENCE

<u>E-STEP</u> :

CONSTRUCT POSTERIOR

$$Q^{(t)}(z) = P(z \mid x; \theta^{(t)})$$

<u>M-STEP</u> :

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \ \mathrm{ELBO}_i^{(t)}(\theta)$$

$$\mathrm{ELBO}^{(t)}(\theta) = \mathbb{E}_{z \sim Q^{(t)}(z)}\left[\log \frac{P(x, z; \theta)}{Q^{(t)}(z)}\right]$$

GMM WITH EM

E-STEP

$$w_j^{(i)} = Q_i(z^{(i)} = j) = P(x^{(i)} \mid z^{(i)}; \underbrace{\phi, \mu, \Sigma}_{\theta})$$

M-STEP

$$\nabla_\theta \, \mathrm{ELBO}^{(t)}(\theta) = 0$$

AND WE GET (**)