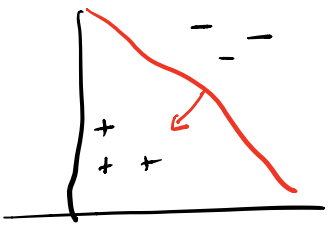
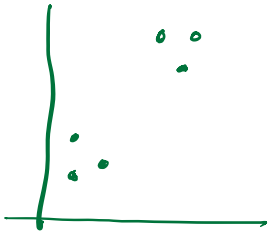
Unsupervised Learning

Today: K-means, mixture of Gaussians, EM.



Supervised Points and Labels



Unsupervised - no labels!

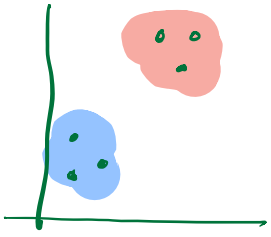
Techniques are valuable (pedagogically & practically)

K-means

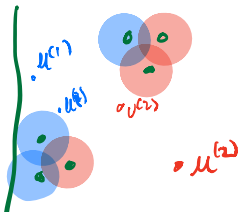
Given $\{x^{(1)}, \cdots, x^{(n)}\} \in \mathbb{R}^d$ & k the # of clusters

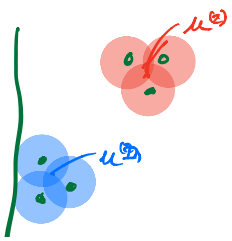
Do find an assignment of $x^{(i)}$ to one of k clusters

$c^{(i)} = j$ "Point i in cluster j"



How do we find these clusters?





1. Randomly Init $\mu^{(1)}, \mu^{(2)}$

2. Assign each point to cluster $\longleftrightarrow$ $c^{(i)} = \underset{j=1\ldots k}{\text{argmin}} \|\mu^{(j)} - x^{(i)}\|^2$

3. Compute new clusters centers $\longleftrightarrow$ $\mu^{(j)} = \frac{1}{|\Omega_j|} \sum_{i \in \Omega_j} x^{(i)}$

Repeat until no points change

where $\Omega_j = \{i : c^{(i)} = j\}$ "compute mean"

Comments

+ Does K-means terminate? Yes!

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

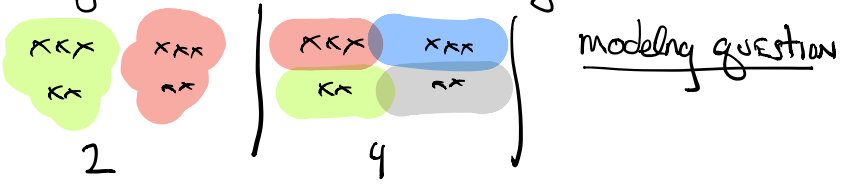
(In notes)

+ Does it find a minimum? Not necessarily (NP-hard)

Side note: K-means++ 2007 from great Stanford students

- Improved approximation bounds through clever init
- Default in sklearn

+ How do you choose K? No one right answer



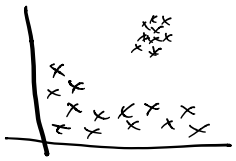
modeling question

Mixture of Gaussians

Toy Astronomy Example (based on real one example)

- Quasars & stars are sources of light

Both emit light, and we observe photons



Goal: Assign each photon to light source $P(z^{(i)} = j)$

"Probability point $z^{(i)}$ belongs to object j"

of kmeans. This is a soft assignment

Challenges

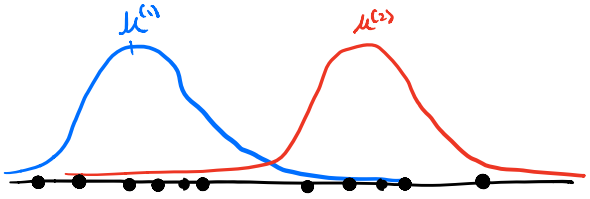
- Many sources (say we know # of sources)
- Sources have different intensity & shape.

Assume: 1. Sources are Gaussian like (μ_j, σ_j^2)

2. We do not assume equal # of points per source (mixture)

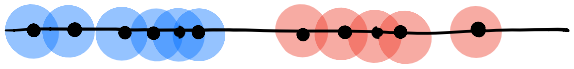
NB: In this example, physics can check how plausible it is.

Mixture of Gaussians: Model & setup



We observe the black points only

Observation 1: If we knew "cluster labels" → solve with GDA



Compute $\mu^{(1)}$ $\mu^{(2)}$ ... Done!

Challenge: we don't!

GIVEN: $x^{(1)} \dots x^{(n)} \in \mathbb{R}$ AND $k > 0$

DO: FIND Prob $z^{(i)}$ for $i = 1 \dots n$ to one of k clusters

$P(z^{(i)} = j)$ SOFT ASSIGNMENT

Gmm model

$P(x^{(i)}, z^{(i)}) = P(x^{(i)} \mid z^{(i)}) P(z^{(i)})$ Bayes Rule

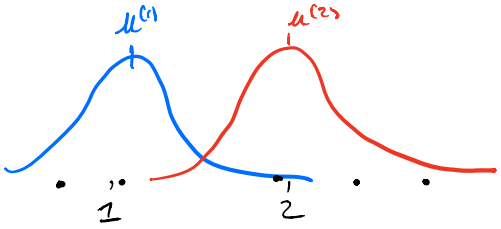
$z^{(i)} \sim \text{multinomial}(\phi)$ s.t. $\sum_{i=1}^{k} \phi_i = 1$, $\phi_j \geq 0$

$x^{(i)} \mid z^{(i)} = j \sim N(\mu_j, \sigma_j^2)$

the PARAMETERS TO BE FOUND ARE IN BLUE

WE CALL $z^{(i)}$ A hidden OR LATENT VARIABLE

$z^{(i)}$ IS NOT DIRECTLY Observed.



GENERATE DATA:

$\phi_1 = 0.7$ $\phi_2 = 0.3$

$\mu^{(1)} = 1$ $\mu^2 = 2$

$\sigma_1^2 = \sigma_2^2 = 1$

Gmm Algorithm (FAMOUS Algorithm & type)

MIRRORS K-MEANS

1. (E-STEP) "Guess" latent VALUES of $z^{(i)}$ (FOR EACH point)
2. (M-STEP) UPDATE PARAMETERS (MLE)

Why Abstract? VERY general EM Algorithm

E-STEP 1:

GIVEN: DATA & CURRENT PARAMETER

Gmm

$\times \frac{1}{k} (\phi, \mu, \sigma)$

DO: PREDICT LATENT VALUE $z^{(i)}$ for $i = 1 \ldots n$

$$w_j^{(i)} = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \sigma)$$ "How likely is data"

$$= \frac{P(z^{(i)} = j, x^{(i)}; \phi, \mu, \sigma)}{P(x^{(i)})}$$ (BAYES RULE)

$$= \frac{P(x^{(i)} \mid z^{(i)} = j; \phi, \mu) P(z^{(i)} = j \mid \phi)}{\sum_l P(x^{(i)} \mid z^{(i)} = l; \phi, \mu) P(z^{(i)} = l \mid \phi)}$$

- $\propto \exp\{-(x^{(i)} - \mu)^2 \sigma^{-2}\}$ "How likely from this GAUSSIAN"
- $= \phi_j$ or ϕ_l "How likely this point came from Cluster"

∴ WE CAN COMPUTE ALL TERMS! RETURN $w_j^{(i)}$

M-STEP

GIVEN $w_j^{(i)} = P(z^{(i)} = j)$ ESTIMATE for all $i \in [n]$ j clusters

DO: ESTIMATE the other OBSERVED PARAMETERS USING MLE

e.g. $\phi_j = \frac{1}{n} \sum_{i=1}^{n} w_j^{(i)} \approx$ "fraction of elements in Cluster j"

$$\mu_j = \frac{\frac{1}{n} \sum_{i=1}^{n} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{n} w_j^{(i)}} \approx$$ "SOFT CLUSTER CENTER"

$\vdots$

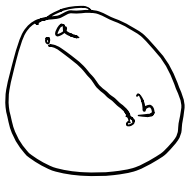
σ_1^2 .

MLE more generally. Let's make rigorous!

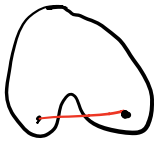
DETOUR: Convexity & JENSEN.

THIS IS A REG RESULT, SO WANT TO GO SLOWLY! ASK!

A SET Ω IS CONVEX if for ANY $a, b \in \Omega$
THE LINE joining a, b IS IN Ω



CONVEX



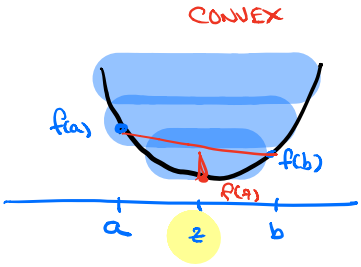
NOT CONVEX

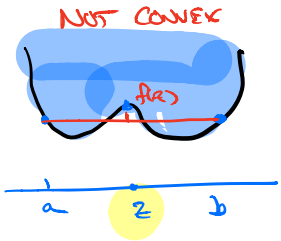
IN symbols,
$\forall \lambda \in [0,1]$, $a, b \in \Omega$
$\lambda a + (1-\lambda) b \in \Omega$

(NEED TO CHECK $\forall a, b \in \Omega$)

GIVEN A function f the graph of f is G_f defined
$G_f = \{(x, y) : y \geq f(x)\}$

A function IS CONVEX if its graph IS CONVEX (AS A SET)





CONVEX MEANS
$\lambda(a, f(a)) + (1-\lambda)(b, f(b)) \in \Omega$

ex Let $z = \lambda a + (1-\lambda) b$
$\lambda f(a) + (1-\lambda) f(b) \geq f(z)$

"EVERY CHORD IS ABOVE function"

If f is twice differentiable, $\forall\, f''(x) \geq 0 \Rightarrow f$ is CONVEX

Pf: $f(a) = f(z) + f'(z)(a-z) + f''(z_a)(a-z)^2$ $\quad z_a \in [a, z]$

$$f(b) = f(z) + f'(z)(b-z) + f''(\xi_b)(a-z)^2 \qquad \xi_b \in [z,b]$$

$$\lambda f(a) + (1-\lambda) f(b) = f(z) + f'(z)(\lambda a + (1-\lambda)b - z) + C \qquad C \geq 0$$

$$\leq \quad \square$$

WE say f is strictly convex if $\forall x \in \mathrm{Dom}(f)$ $f''(x) > 0$

Ex: $f(x) = x^2 \Rightarrow f''(x) = 2 \Rightarrow$ strongly convex

$f(x) = x^2(x-1)^2$ is the graph above that is not convex.

JENSEN'S INEQUALITY $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$ for convex f.

Ex: X takes value a with prob. λ

takes value b with prob. $1-\lambda$

$\mathbb{E}[f(x)] = \lambda f(a) + (1-\lambda) f(b)$

$f(\mathbb{E}[x]) = f(z)$ for $z = \lambda a + (1-\lambda) b$

for convex f, our definition above implies Jensen's thm.

NB: for finitely supported distributions, PROVE Jensen's by induction

Stronger: if f is strongly convex and $\mathbb{E}[f(x)] = f(\mathbb{E}[x])$

then X is a constant (experts: almost surely)

WE NEED Concave functions g is concave if $-g$ is convex

$$\mathbb{E}[g(x)] \leq g(\mathbb{E}[x])$$

Ex: $g = \log(x) \Rightarrow g''(x) = -x^{-2}$ on $(0, \infty)$ NEGATIVE

curve below function.

What about $h(x) = ax + b$? $h''(x) = 0 \Rightarrow$ convex and concave!

$$\mathbb{E}[h(x)] \geq h(\mathbb{E}[x]) \quad \& \quad \mathbb{E}[h(x)] \leq h(\mathbb{E}[x])$$

$$\Rightarrow \mathbb{E}[h(x)] = h(\mathbb{E}[x]) \quad \text{linearity of expectation}$$

End Detour

Em Algorithms as max likelihood

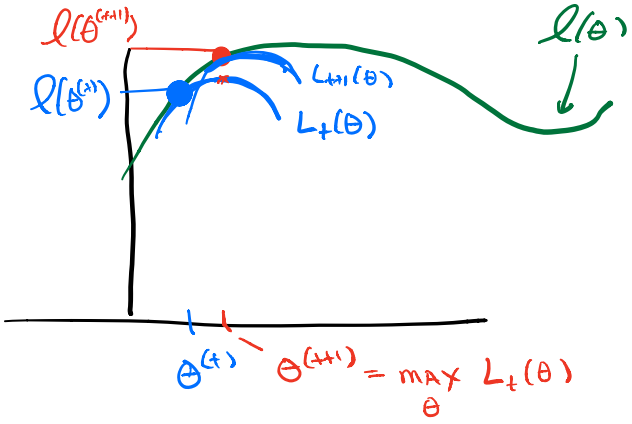
$$\ell(\theta) = \sum_{i=1}^{n} \log P(x^{(i)}; \theta)$$

(θ: parameters; $x^{(i)}$: data)

We assume $P(x;\theta) = \sum_z P(x, z; \theta)$ cf. GMM

(z: latent variable)

Picture of Algorithm



$L_t(\theta) \leq \ell(\theta)$ lowerbound

$L_t(\theta^{(t)}) = \ell(\theta^{(t)})$ tight

Hope: $L_t(\theta)$ is easier to optimize than $\ell(\theta)$

Rough Algorithm

(E-step) 1. Find $L_t(\theta)$ given $\theta^{(t)}$

(M-step) 2. $\theta^{(t+1)} = \arg\max_\theta L_t(\theta)$

How do we construct $L_t(\theta)$

Let's examine a single point

$$\log \sum_z P(x,z;\theta) = \log \sum_z \frac{Q(z)\,P(x,z;\theta)}{Q(z)} \quad \text{(for any } Q(z)\text{)}$$

Pick $Q(z)$ s.t. $\sum_z Q(z) = 1$, $Q(z) \geq 0$ (*)

$$= \log \mathop{\mathbb{E}}_{Q(z)}\left[\frac{P(x,z;\theta)}{Q(z)}\right] \quad \text{(just symbol pushing)}$$

$$\geq \mathbb{E}\left[\log \frac{P(x,z;\theta)}{Q(z)}\right] \quad \text{(JENSEN, log CONCAVE)}$$

$$= \sum_z Q(z) \log \frac{P(x,z;\theta)}{Q(z)} \quad \text{(symbol pushing, def of } \mathbb{E}\text{)}$$

① this holds for ANY Q satisfying (*)

② this gives a family of lower bounds (PROPERTY, ABOVE)

Pick $Q(z)$ TO SATISFY PROPERTY 2. that is

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

i.e. WHEN IS JENSEN'S tight?

if WE PICK $\frac{P(x,z;\theta)}{Q(z)} = c$ i.e. $Q(z) = P(z|x;\theta)$

NB: $Q(z)$ depends on θ & x — different $Q^{(i)}(z)$ for every point.

WE DEFINE Evidence-Based lower bound (ELBO), SUM OVER z

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

WE'VE SHOWN $\ell(\theta) \geq \sum_{i=1}^{n} ELBO(x^{(i)}, Q^{(i)}; \theta)$ for any $Q^{(i)}$

$\ell(\theta^{(t)}) = \sum_{i=1}^{n} ELBO(x^{(i)}, Q^{(i)}; \theta^{(t)})$ for choice of $Q^{(i)}$ ABOVE

<u>WRAP UP</u>:

1. (E-STEP) $Q^{(i)}(z) = P(z^{(i)} | x^{(i)}; \theta)$
2. (M-STEP) $\theta^{(t+1)} = \underset{\theta}{\text{ARGMAX}}\ L_t(\theta)$

in which $L_t(\theta) = \sum_{i=1}^{n} ELBO(x^{(i)}, Q^{(i)}, \theta)$

WHY DOES THIS TERMINATE $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$

IS IT GLOBALLY OPTIMAL? NOPE! (SEE PICTURE)

IN this lecture, we say hard & soft clustering methods

WE DERIVED General Algorithm (EM) IN terms of MLE.