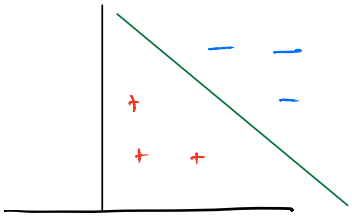
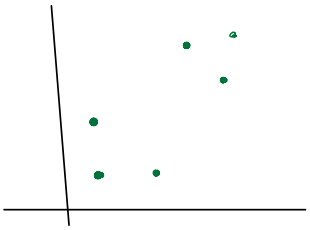
Unsupervised Learning

TODAY: K-means, mixture of Gaussians, EM



Supervised Setting



Unsupervised, no labels!

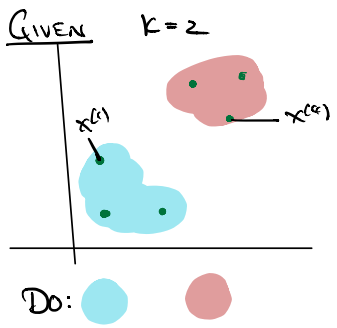
Unsupervised is harder than Supervised

allow Stronger assumptions

accept weaker guarantees

TECHNIQUES & IDEAS ARE VALUABLE

K-MEANS



GIVEN $x^{(1)} \cdots x^{(n)} \in \mathbb{R}^d$ & INTEGER K, # of clusters

DO find assignment of $x^{(i)}$ to ONE of K clusters

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

eg. $c^{(2)} = 2$ while $c^{(1)} = 1$

How do we find these clusters? ITERATIVE Approach

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

2. Assign each point to closest cluster $\longleftrightarrow$ for each $i = 1 \dots n$ $c^{(i)} = \underset{j=1\dots k}{\text{Argmin}} \|\mu^{(j)} - x^{(i)}\|^2$

3. Compute New cluster centers for $j = 1 \dots k$ $\mu^{(j)} = \frac{1}{|\Omega_j|} \sum_{i \in \Omega_j} x^{(i)}$ s.t. $\Omega_j = \{i : c^{(i)} = j\}$

Repeat until no points change

Comments

Does k-means terminate? Yes!

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

(see notes)

Does it find a global minimum? not necessarily... NP-hard

Side note: k-means++ from great Stanford students

+ Improved Apx Ratio through Clever Init
+ Default in Sklearn

How do you choose k? No one Right Answer.

2 clusters

4 clusters

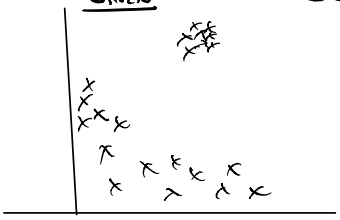
Modeling Question!

Mixture of Gaussians

Toy Astronomy Example (based on a paper from UW)

- Quasars & stars are sources of light
- Both emit light, and we observe Photons

Given



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

"Probability Point $z^{(i)}$ belongs to Object j"

cf K means. This is a soft Assignment

Challenges + Many Sources (say we know K, # of sources)

+ Sources have different intensity & shapes

Assume 1. Sources are well modeled by Gaussian (μ_j, σ_j^2)

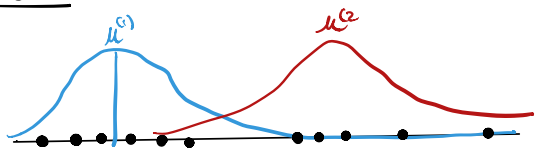
2. We do not assume equal # of points per source

→ Unknown mixture

NB: Physics folks can check if recovered values make sense.

Mixture of Gaussians (model & setup) – 1d for simplicity

Model:



We observe points w/o sources:

$x^{(i)} \in \mathbb{R}$

Observation 1 if we knew "cluster labels" → solve ol GDA.

Compute $\mu^{(1)}, \mu^{(2)}$ and be done. **Challenge** we don't

GIVEN $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}$ AND POSITIVE INTEGER k

DO find P s.t. for $i=1\dots n$ & $j=1\dots k$ clusters

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

According to the "GMM MODEL"

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

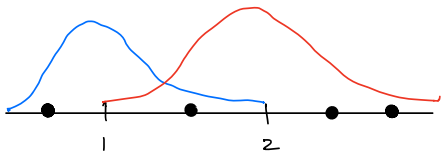
$$z^{(i)} \sim \text{Multinomial}(\phi) \quad \phi_j \geq 0 \quad \sum_{j=1}^{k} \phi_j = 1 \quad \text{"which source"}$$

$$x^{(i)} \mid z^{(i)}=j \sim N(\mu_j, \sigma_j^2) \quad \text{GAUSSIAN IN EACH SOURCE}$$

THE PARAMETERS TO BE FOUND ARE highlighted

WE CALL $z^{(i)}$ A hidden OR LATENT VARIABLE. $z^{(i)}$ IS NOT DIRECTLY OBSERVED

helpful to think in terms of Sampling.



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

$\mu_1 = 1 \quad \mu_2 = 2 \quad \sigma_1^2 = \sigma_2^2 = \frac{1}{3}$ (roughly)

Gmm Algorithm (Famous Algo & Class)

Mirrors K-means

1. (E-step) "Guess latent values" of $z^{(i)}$ for each point
2. (M-step) Update Parameters

Abstractly our first example of EM-Algorithm (Expectation Maximization)

(E-step) Given Data & current guess at parameters $(\phi, \mu, \sigma^2 \ldots)$

Do Predict latent variable $z^{(i)}$ for $i = 1 \ldots n$

$$w_j^{(i)} = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \sigma)$$ our goal

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

$$= \frac{P(x^{(i)} \mid z^{(i)} = j; \mu, \sigma) \, P(z^{(i)} = j; \phi)}{\sum_{l=1}^{k} P(x^{(i)} \mid z^{(i)} = l; \mu, \sigma) \, P(z^{(i)} = l; \phi)}$$

ϕ_j in the numerator; ϕ_l in the denominator

- $\propto \exp\left\{-\frac{(x^{(i)} - \mu_j)^2}{\sigma_j^2}\right\}$ "How likely is $x^{(i)}$ according to Gaussian (μ_j, σ_j^2)"
- "How likely point from cluster"

Key Point We can compute all terms! Return $w_j^{(i)}$

M-STEP

GIVEN $w_j^{(i)}$ OUR CURRENT ESTIMATE of $P(z^{(i)} = j)$ for $i = 1 \ldots n$
$j = 1 \ldots k$ clusters

DO ESTIMATE OBSERVED PARAMETERS (USING MLE)

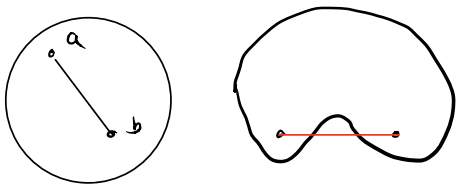
e.g. $\phi_j = \frac{1}{n} \sum_{i=1}^{n} w_j^{(i)} \approx$ fraction of elements in cluster j

$$\mu_j = \frac{\sum w_j^{(i)} x^{(i)}}{\sum w_j^{(i)}} \quad \ldots \text{ etc} \ldots$$

MLE. Let's MAKE REASONS!

Detour Convexity & Jensen (this is a key result, we'll go slowly)

A set Ω is convex if for any $a, b \in \Omega$ the line joining a, b is in Ω as well.



Convex

NOT CONVEX!

In symbols,

$\forall \lambda \in [0,1], \quad a, b \in \Omega$

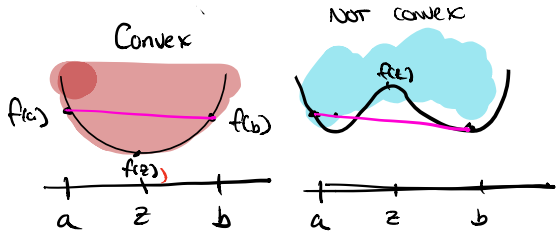
$\lambda a + (1-\lambda) b \in \Omega$

(need to check $\forall a, b \in \Omega$)

Given a function f, the graph of f G_f is defined as

$$G_f = \{ (x, y) : y \geq f(x) \}$$

A function is convex if its graph is convex (as a set)



"Every chord is above function"

In symbols, $\forall \lambda \in [0,1]$

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

or let $z = \lambda a + (1-\lambda) b$

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

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

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

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

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

$\lambda a + (1-\lambda) b - z = 0$ by def of z.

i.e. $\lambda f(a) + (1-\lambda) f(b) \leq f(z)$ □

We say f is strongly convex if $\forall x \in \text{Dom}(f)$ $f''(x) > 0$.

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

$f(x) = x^2(x-1)^2$: graph above (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 $1-\lambda$

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

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

NB: Can prove finitely supported distribution by induction

for convex f, definition implies thm in this case!

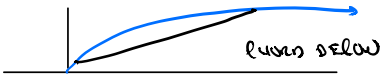
Strongly if f is strongly convex, and $\mathbb{E}[f(x)] = f(\mathbb{E}[x])$

$\Rightarrow$ X is a constant (recall: almost surely)

We need concave functions g concave iff $-g$ is convex

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

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



What about $h(x) = ax + b$ convex & concave since $h''(x) = 0$.

End Detour

EM Algorithm as max likelihood

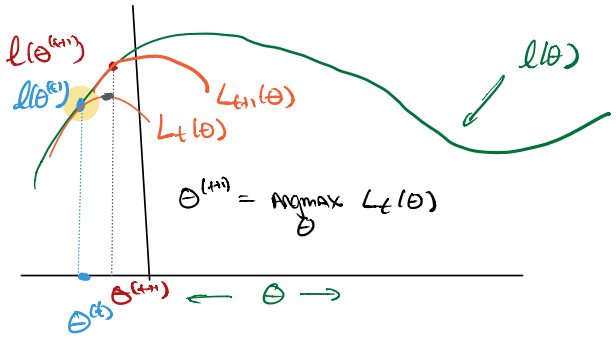
$$\ell(\theta) = \sum_{i=1}^{n} \log P(x^{(i)}; \theta)$$

PARAMETERS

DATA

WE ASSUME $P(x;\theta) = \sum_z P(x, z; \theta)$ of GMM LATENT VARIABLE

Picture of Algorithm



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

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

HOPE L_t easier to optimize than $\ell(\theta)$

Rough Algo

(E-STEP) 1. GIVEN $\theta^{(t)}$ find L_t

(M-STEP) 2. GIVEN L_t, SET $\theta^{(t+1)} = \text{argmax}_\theta L_t(\theta)$

How do we construct L_t?

(let's look at single pnt)

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

WE PICK $Q(z)$ S.T. $\sum_z Q(z) = 1$ AND $Q(z) = 0$ (*)

$$= \log \mathop{\mathbb{E}}_z \left[\frac{P(x, z; \theta)}{Q(z)} \right] \quad \text{(Symbol Pushing)}$$

$$\geq \mathop{\mathbb{E}}_{z}\left[\log \frac{P(x,z;\theta)}{Q(z)}\right] \qquad \text{JENSEN'S (log IS CONCAVE)}$$

$$= \sum_{z} Q(z) \log \frac{P(x,z;\theta)}{Q(z)} \qquad \text{(DEF of } \mathbb{E}\text{)}$$

Key step holds for ANY such Q: (*)

This gives a family of lower bounds, one for each choice of Q ($B_t \leq \ell$)

How do we make it tight? Select Q to make inequality tight

What if... $\log \frac{P(x,z;\theta)}{Q(z)} = c$ for some constant, then Jensen's is Equality!

So, $Q(z) = P(z|x;\theta)$ then $\qquad P(x,z;\theta) = P(z|x;\theta)P(x;\theta)$

$c = \log P(x;\theta)$ does not depend on z, so constant!

NB: $Q(z)$ does depend on θ & x — we will select a $Q^{(i)}(z)$ for every point independently.

We define Evidenced-based lower bound (ELBO), sum over z

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

We've shown $\ell(\theta) \geq \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, Q^{(i)}; \theta)$ for ANY $Q^{(i)}$ satisfying (*)

lower bound

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

WRAP UP

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

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

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

IS IT GLOBALLY OPTIMAL? (NOPE SEE PICTURE)

WE DERIVED HARD & SOFT CLUSTERING METHODS

EM ALGORITHM IN TERMS OF MLE.