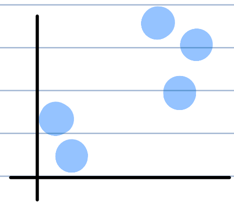
Unsupervised Learning

Supervised



Unsupervised

- K-means
- Mixture of Gaussians
- EM

K-means

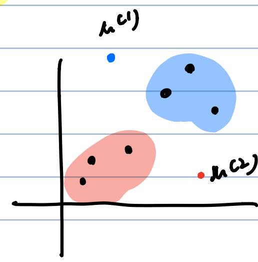
<u>Given</u> $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$ points

K is the number of clusters

<u>Do</u> Find an assignment of $x^{(i)}$ to j clusters, for $j = 1 \dots k$

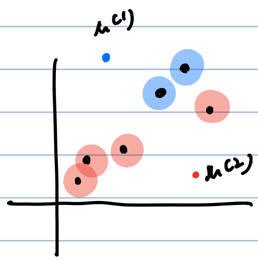
$c^{(i)} = j$ "Point i is to cluster j"

How do we find these clusters?

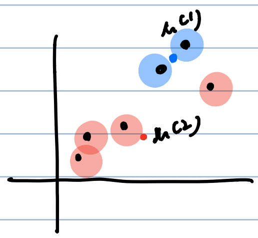


$\mu^{(1)}$: Blue cluster center

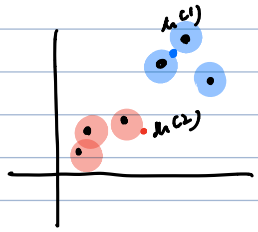
$\mu^{(2)}$: Red cluster center



Assign each point to its nearest cluster centroid.



Recompute centroids



Assign points again. Converge!

Terminates till no point changes.

Algo

Repeat until no changes

- Randomly Init $\mu^{(1)}, \mu^{(2)} \ldots \mu^{(k)}$

2. Assign each point to a cluster center

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

3. Compute new cluster centers

$$\mu^{(j)} = \frac{1}{|\Omega_j|} \sum_{i \in \Omega_j} x^{(i)} \qquad \Omega_j = \{i : c^{(i)} = j\}$$

Comments

① Does this terminates? Yes! $J(c, \mu) = \sum_{i=1}^{n} \|x^{(i)} - \mu^{c^{(i)}}\|^2$

Monotonic decreasing

② Is it globally optimal? Not necessary (NP-hard)

K-Means ++ (By Stanford Alumn 2007)

- Clever init! Check CS246- Clustering slides
- Default in sklearn

How do we choose k? No one right answer.

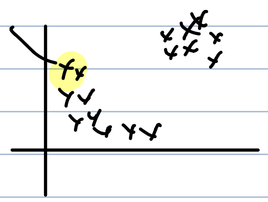
Mixture of Gaussians

Toy Astronomy Example (Based on real on example)

- Quasaas. source of light

$z^{(i)}$

Photons



Goal: Assign $P(z^{(i)} = j)$

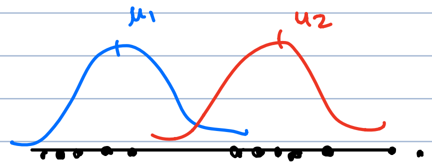
Probability of Point $z^{(i)}$ came from source j

Mixture model is K-means with a soft assignments

Assume

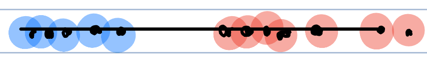
- Sources are Gaussian like (μ_j, σ_j^2)
- We do not assume equal # of points for each source (unknown mixture of points)

Mixture of Gaussians : Model setup (1d)



This is what we observe.

Observation 1. : If we knew which points came from each Gaussian



$\mapsto$ compute $\mu^{(j)}, \sigma^2_{(j)}$

Challenge: we don't see that!

<u>Given</u>: $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^d$ and $k > 0$

<u>Do</u>: $z^{(i)} = j$ " $x^{(i)}$ comes from source j "

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

<u>Model</u>

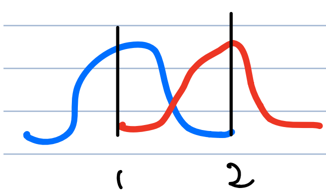
$$P(x^{(i)}, z^{(i)}) = P(x^{(i)} \mid z^{(i)})\, P(z^{(i)})$$

$$z^{(i)} \sim \text{multinomial}(\phi) \qquad \sum_k \phi_k = 1 \qquad \phi_k \geq 0$$

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

We call $z^{(i)}$ a hidden or latent variable

$z^{(i)}$ is not directly observed



$\phi_1 = 0.9 \qquad \phi_2 = 0.1$

$\mu^{(1)} = 1 \qquad \mu^{(2)} = 2 \qquad \sigma_1^2 = \sigma_2^2 = 1$

> Recover these parameters

GMM Algorithm (mirror k-means)

① (E-Step) " Guess " latent values of $z^{(i)}$ (For each point)

② (M-Step) Upate the variables of the model

$p(x^{(i)} \mid z^{(i)} = j; \cancel{\phi}, \mu, \sigma^2)\ P(z^{(i)} = j; \phi, \mu, \sigma^2)$

$= P(x^{(i)} | z^{(i)} = j, \mu, \sigma^2) P(z^{(i)} = j; \phi) P(\cancel{\mu, \sigma^2})$

E-Step

<u>Given</u>: Data & current parameters $x^{(i)}, \phi, \mu, \sigma$

<u>Do</u>: Predict the latent variable $z_j^{(i)}$

$$w_j^{(i)} = P(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \sigma^2)$$

$$= \frac{P(z^{(i)} = j, x^{(i)}; \phi, \mu, \sigma^2)}{P(x^{(i)}; \phi, \mu, \sigma^2)}$$ Bayes Rule

How likely $x^{(i)}$ in cluster j

$\exp\left(\frac{-(x^{(i)} - \mu_j)^2}{\sigma^2}\right) \propto$

$$= \frac{P(x^{(i)} | z^{(i)} = j; \sigma^2, \mu) \; P(z^{(i)} = j; \phi)}{\sum_l P(x^{(i)} | z^{(i)} = l; \sigma_l^2, \mu_l) \; P(z^{(i)} = l; \phi)}$$

$\to \phi_j$

$\to \phi_l$.

$\therefore$ We can compute easy one of these elements

M-Step

<u>Given</u> $w_j^{(i)} = P(z^{(i)} = j)$ for all $i = 1 \sim n$, $j = 1 \sim k$

<u>Do</u> Estimate all parameters using <u>MLE</u>

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

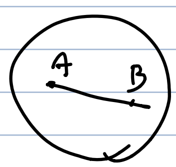
$$\mu_j = \frac{\sum_{i=1} w_j^{(i)} x^{(i)}}{\sum w_j^{(i)}}$$ Soft cluster center

MLE!

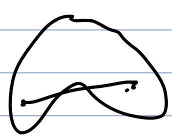
$$\sigma_j^2 = \frac{\sum_{i=1} w_j^{(i)} (x^{(i)} - \mu_j)^2}{\sum_{i=1} w_j^{(i)}}$$ (Not sure)

Detour: Convexity

A set Ω is convex of for any $a, b \in \Omega$ the line joining a, b is entirely in Ω



Convex



Not Convex

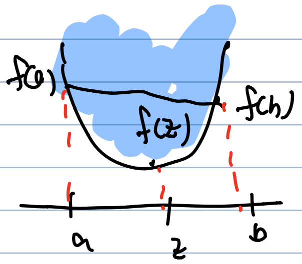
Symbols
$\forall \lambda \in [0, 1]$, $a, b \in \Omega$

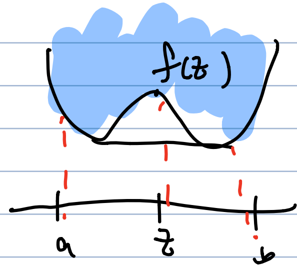
$\lambda a + (1 - \lambda) b \in \Omega$

Given a function f, the graph of the function G_f
$G_f = \{(x, y) : y \geq f(x)\}$

A function is convex if it's graph is convex (as a set)

x^2





$(x-1)^2(x-2)^2$

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

or

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

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

- If f is twice differentiable, $f''(x) \geq 0 \Rightarrow$ Convex
- We say f is strictly convex if $f''(x) > 0$

ex: $f(x) = x^2$ $f''(x) = 2 \Rightarrow$ strictly convex

Jensen's Inequality

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

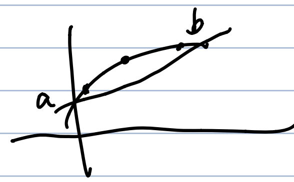
ex: x takes a w/ prob λ
b w/ prob $(1-\lambda)$

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

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

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

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



Recap:

1. K-means (Iterative Solution)