Unsupervised Learning

Today: K-Means, Mixture of Gaussians, EM.

Supervised Points and Labels vs. Unsupervised - No Labels!

Techniques are valuable (pedagogically & practically)

K-Means

Given $\{x^{(1)}, \ldots, x^{(n)}\} \in \mathbb{R}^d$ & $k$ 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?

Given $\mu^{(1)}, \ldots, \mu^{(k)}$
1. Randomly Init $\mu^{(0)}_d, \mu^{(x)}_d$

2. Assign each point to cluster $c^{(x)} = \arg\min_d \| x^{(x)} - \mu^{(x)}_d \|$

3. Compute new clusters centers

Repeat until no points change $\mu^{(x+1)}_d = \frac{1}{\omega_d} \sum_{x^{(x)} \in \Omega_d} x^{(x)}$ where $\omega_d = \sum_i \omega^{(x)} = \| \Omega_d \|$

"Compute mean"

Comments

+ Does K-Means terminate? *Yes!*

$J(c, \mu) = \sum_{x^{(x)}} \| x^{(x)} - \mu^{(x)} \|^2$ 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

  2 4

  Mixture of Gaussians

**Toy Astronomy Example** *(based on real on example)*

  * Quasars & stars are sources of light
  * Both emit light, and we observe clusters
**Goal**: Assign each photon to light source $P(Z^{(s)} = j)$

"Probability Point $Z^{(s)}$ belongs to Object $j"$

Of K-means, this is a soft assignment.

**Challenges**

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

**Assume**:

1. Sources are Gaussian like $(c_j, \sigma_j^2)$
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

![Graph showing two Gaussian distributions and points]

We observe the black points only

**Observation 1**: If we knew "Cluster lasers" $\rightarrow$ Solve with GMM

Compute $\mu^{(s)}, \sigma^{(s)}$ $\rightarrow$ Done!

**Challenge**: we don't!
**Given:** \( x^{(i)} \ldots x^{(n)} \in \mathbb{R} \) and \( k > 0 \)

**Do:** Find Prob \( z^{(i)} \) for \( i = 1 \ldots n \) to one of \( k \) clusters

\[
P(z^{(i)} = j) \text{ soft assignment}
\]

**Gmm model**

\[
P(x^{(i)}, z^{(i)}) = P(x^{(i)} | z^{(i)}) P(z^{(i)}) \text{ Bayes Rule}
\]

\[
z^{(i)} \sim \text{multinomial}(\Phi)
\]

\[
x^{(i)} | z^{(i)} = j \sim N(\mu_{j}^{(i)}, \sigma_{j}^{2})
\]

*The parameters to be found are in blue.*

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

\( z^{(i)} \) is not directly observed.

**Gmm Algorithm** (Famous Algorithm for topic)

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:**

**Given:** Data \& Current Parameters

\[
x^{(i)} (\phi, \mu, \sigma)
\]
Do: Predict latent value $z_i^{(n)}$ for $i = 1, ..., n$

$w_j^{(n)} = P(z_i^{(n)} = j | x_i^{(n)}, \phi, \mu, \sigma)$

"How likely is class $j$ for $x_i^{(n)}"$

$= \frac{P(x_i^{(n)} | z_i^{(n)} = j, \phi, \mu, \sigma) P(z_i^{(n)} = j | \phi)}{P(x_i^{(n)})}$

"Bayes Rule"

$\sum_j \frac{P(x_i^{(n)} | z_i^{(n)} = j, \phi, \mu, \sigma) P(z_i^{(n)} = j | \phi)}{\sum_j P(x_i^{(n)} | z_i^{(n)} = j, \phi, \mu, \sigma) P(z_i^{(n)} = j | \phi)}$

$\propto \exp \left[-\frac{(x_i^{(n)} - \mu)^2}{2 \sigma^2} \right]" \text{How likely from the Gaussian}"$

$= \phi_j \text{ or } \phi_2 " \text{How likely this point came from Cluster}"$

; WE CAN COMPUTE ALL TERMS! RETURN $w_j^{(n)}$

M-STEP

\textbf{Given} $w_j^{(n)} = P(z_i^{(n)} = j)$ Estimate for all $i \in \{1, ..., n\}$ use EM

\textbf{Do:} Estimate the other observed parameters use MLE

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

$\mu_j = \frac{1}{\sum_{i=1}^{n} w_j^{(n)}} \sum_{i=1}^{n} w_j^{(n)} x_i^{(n)} \in \text{"soft center center"}$

$\sigma^2$
more generally, let's make a road!

**Detour: Convexity & Jensen.**

*This is a key result, so want to go slowly! Ask!*

A set $\Omega$ is **convex** if for any $a, b \in \Omega$

the line joining $a, b$ is in $\Omega$

In symbols,

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

**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*

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

"Every chord is above function"

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

Proof: $f(a) = f(z) + f'(z)(a - z) + \frac{f''(z)}{2}(a - z)^2$ $\forall \in (a, b)$
\[ f(b) = f(z) + f'(z)(b-z) + \frac{f''(z_0)}{2!}(b-z)^2 \quad z \in [a,b] \]
\[ \lambda f(a) + (1-\lambda)f(b) = f(z) + f'(z)(\lambda a + (1-\lambda)b - z) + \epsilon \]

We say \( f \) is strictly convex if \( \forall x \in \text{dom}(f) \) \( f''(x) > 0 \)

\[ f(x) = x^2 \Rightarrow f''(x) = 2 \Rightarrow \text{strongly convex} \]
\[ f(x) = x^3 - 1 \text{ is the graph above that is not convex.} \]

**Jensen's Inequality**

\[ \mathbb{E}[f(x)] \geq f(\mathbb{E}[x]) \quad \text{for convex } f. \]

**Ex:** \( x \) takes value \( a \) with prob \( \lambda \)
\( \) takes value \( b \) with prob \( 1-\lambda \)

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

For convex \( f \), our definition above implies Jensen's thm.

**Note:** 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 (exptb: almost surely)

**WE NEED CONCAVE FUNCTION**
\[ g \text{ is concave if } -g \text{ is convex} \]
\[ \mathbb{E}[g(x)] \leq g(\mathbb{E}[x]) \]

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

\[ \text{concave below function.} \]
What about $f(x) = ax + b$? $f'(x) = a \Rightarrow \text{convex and concave}$

$\mathbb{E}[f(x)] \geq \mathbb{E}[f(0)] \land \mathbb{E}[f(x)] \leq \mathbb{E}[f(E(x))]$

$\Rightarrow \mathbb{E}[f(x)] = \mathbb{E}[f(0)] \quad \text{linear and concave}$

End Detour

Em Algorithms as max likelihood

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

We assume $P(x; \theta) = \sum_{z} P(x, z; \theta)$ of GMM

Picture of Algorithm

$\theta(1), \theta(\infty) = \max_{\theta} L_t(\theta)$

$\mathcal{L}_t(\theta) \leq l(\theta) \quad \text{looser and}$

$\mathcal{L}_t(\theta(\infty)) = l(\theta(\infty)) \quad \text{tight}$

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

Rough Algorithm

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

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

How do we construct $L_t(\theta)$
Let's examine a single point

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

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

\[ \log \sum_z \frac{Q(z) P(x, z; \theta)}{Q(z)} = \log \mathbb{E}_{Q(z)} \left[ \frac{P(x, z; \theta)}{Q(z)} \right] \quad \text{(Just symbolic pushing)} \]

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

\[ = \sum_z Q(z) \log \frac{P(x, z; \theta)}{Q(z)} \quad \text{(symbol pushing, def 2)} \]

1. This holds for any Q satisfying \((*)\)
2. This gives a family of lower bounds (PROPERTY 1 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 tight? if we pick \( \frac{P(x, z; \theta)}{Q(z)} = 1 \) i.e., \( Q(z) = P(z | x; \theta) \)

NB: Q(z) depends on \( \theta + x - \) different \( Q(z) \) for any \( x \).

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 $L(\theta) \geq \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, q^{(i)}; \theta)$ for any $q^{(i)}$

\[ l(q^{(i)}) = \sum_{i=1}^{n} \text{ELBO}(x^{(i)}, q^{(i)}, \theta) \text{ for choice of } \theta^{(i)} \text{ above} \]

Wrap up:

1. (E-STEP) $Q^{(i)}(z) = P(z^{(i)} | x^{(i)}, \theta)$
2. (M-STEP) $\theta^{(i+1)} = \text{argmax} L_{\theta}(\theta)$

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

Why does this terminate $L(\theta^{(i+1)}) \geq L(\theta^{(i)})$?

Is it globally optimal? No! (See picture)

In this lecture, we say hard & soft clustering methods

We derived general algorithm (GEM) in terms of MLE.