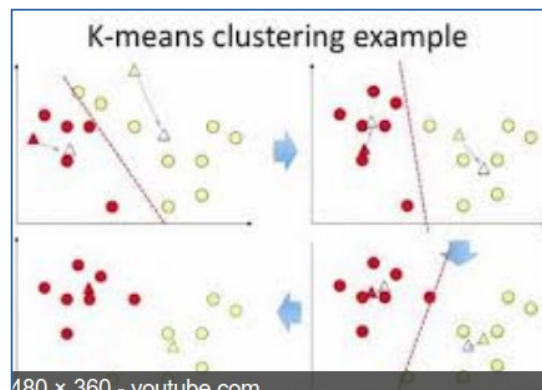# K-mean Algorithm (1)

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \left\| \mathbf{x} - \boldsymbol{\mu}_i \right\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$
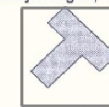
Example:

Form feature vectors

Cluster Seekingg

Example On Simple Pattern Recognition

Given two binary images, derived from two objects, T and O, design a technique to identify them

Example: Computation of
(1) Area (size);
(2) X-bar;
(3) Y-bar;
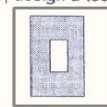(4) Orientation, theta angle
(5) Perimeter of an object

Fig1(a),(b)

Fig2(a),(b)

Good continuation or noise? What to do with this noise?

Feature Vector		Size	X-bar	Y-bar	Orientation	Perimeter	
V_1(v1,..., v5)	T	v11	v12	v13	v14	v15	From Fig1(b)
V_2(v1,..., v5)	L	v21	v22	v23	v24	v25	From Fig2(b)

Harry Li, PhD. SJSU

K-means clustering example

480 x 360 - youtube.com

K-mean Algorithm (2)

Assignment step: Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean.[7] (Mathematically, this means partitioning the observations according to the Voronoi diagram generated by the means).

$$S_i^{(t)} = \left\{ x_p : \left\| x_p - m_i^{(t)} \right\|^2 \leq \left\| x_p - m_j^{(t)} \right\|^2 \; \forall j, 1 \leq j \leq k \right\},$$

Update step: Calculate the new means to be the centroids

$$m_i^{(t+1)} = \frac{1}{\left| S_i^{(t)} \right|} \sum_{x_j \in S_i^{(t)}} x_j$$

Algorithm:

1) Assum Number of Classes = k.
   Pick $\vec{m_i}(t)\big|_{t=1}$, $i=1,2,\ldots,k$.
   Arbitrary.
   e.g. $\vec{m_i}(t) = \vec{x_i}$

2) $\left\| \vec{x_\ell} - \vec{m_i}(t) \right\| \leq \left\| \vec{x_\ell} - \vec{m_j}(t) \right\|$ for $j \neq i$
   $1 \leq j \leq k$
   Regroup $\vec{x_\ell}$ such that $\vec{x_\ell} \in \omega_i(S_i(t))$

3) Update $\vec{m_i}(t+1)$ for $i=1,2,\ldots k$, Check
   Termination $\vec{m_i}(t+1) = \vec{m_i}(t)$ yes, stop
   No, Goto 2

K-mean Example

CMPE297 Video Analytics March 15, 2018 HL 1/.

Non-Supervised Learning vs. Supervised Learning.

Today's Topics: K-mean for Cluster Analysis.

Background: Computer Vision → Feature Extraction Techniques.
Pre-processing Tasks.
2D Convolutions.
Kernel Design. LoG

$B(x,y)$ Binary Image Analysis.
Area (Size), Perimeters,
$\bar{x}, \bar{y}$, moments. m_{ij}
Feature Vectors. $i, j = 0, 1, 2$.
Images from CAT I

Better/Best "Cluster" of the Data,
Define Objective function for Optimization.

Error — minimized — Takes Care of Possible "–" Cancellations

$$\sum_{i=1}^{K} \sum_{\vec{v} \in S_i} \| \vec{v} - \vec{u} \|^2 \quad \cdots (1)$$

Sample Data Vector — Ideal Center "Cluster"

S_i : CAT. I , CAT. II
Total Number of Samples Per CAT.
For Total Number of CAT's.

Data Collection
$\{ \vec{V}_{I,i} | i = 1, 2, \cdots N \}$
"N" Samples (I)
CAT.I   CAT.II
$\{ \vec{V}_{II,j} | j = 1, 2, \cdots M \}$
"M" Samples for CAT. II

Images from CAT II
$\vec{v}_2$ ← $I_1(x,y)$ from CAT.I.
$I_2(x,y)$ from CAT.II
Decision Function.
K-mean Technique.

Machine Learning
for $\vec{V}_{I,i}$ $(\vec{V}_{I1}, \vec{V}_{I2}, \cdots \vec{V}_{Ik})$ K-Dimensional Space →
$\vec{V}_{I1} : A$ (size), $\vec{V}_{I2} :$ Perimeter, $\vec{V}_{I3} : P/A$,
$\vec{V}_{I4} : \bar{x}, \vec{V}_{I5} : \bar{y}$, ... (8 Additional moments) Here k=13

Algorithm: Step 1 Assignment Step.

$$S_i^{(t)} = \{ \vec{x}_p : \| \vec{x}_p - \vec{m}_i^{(t)} \|^2 \leq \| \vec{x}_p - \vec{m}_j^{(t)} \|^2 \; \forall j, 1 \leq j \leq k \} \quad \cdots (2)$$

Step 2   where $\vec{x} = \vec{v}$
Update :

$$m_i = \sum_{x_j \in \{ S_i^{(t)} \}} \vec{x}_j \frac{1}{|S_i^{(t)}|} \quad \cdots (3)$$

Harry Li, Ph.D.

Hand Calculation

CMPE 2.9 — March 15, 2018 HL.21.

Similarly

Example: Given 2 CATs (classes), find the means, m_j, for $j=1,2$ Based on K-mean Algorithm.

(See Handout [P1.]) In Class.

$\overline{X} : \overline{V}$

Notation

CAT$_2$: $\underline{W}_2$

CAT$_1$: $\underline{W}_1$ Perimeter $x_1(v_1)$ Area

$\|\overrightarrow{X_1} - \overrightarrow{m_2}\| = (x_{11} - m_{21})^2 +$

$(x_{12} - m_{22})^2 = (0-1)^2 + (0-0)^2$

$= 1 \rightarrow \overrightarrow{X_1} \in W_1 \ (S_1(t-1))$

And $\overrightarrow{X_2}$, we have

$\|\overrightarrow{X_2} - \overrightarrow{m_2}\| \leq \|\overrightarrow{X_2} - \overrightarrow{m_1}\|$

$\overrightarrow{X_2} \in W_2 \ (S_2(t-1))$

for $\overrightarrow{X_3}$,

$\|\overrightarrow{X_3} - \overrightarrow{m_1}\| \leq \|\overrightarrow{X_3} - \overrightarrow{m_2}\|$

$\overset{=1}{\overrightarrow{X_3}} \in W_1 \ (S_1(1)) \overset{=2}{}$

Step 1. Arbitrary 2 pts as 2 classes

$m_i(t) \begin{cases} m_1(1) = \overline{X_1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ m_2(1) = \overline{X_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = (1,0)^t \end{cases}$

9 means   K=2

mean.

Step 2. Compare Each Pt with the mean from Each Class, to make sure Eqn(2) Satisfied.

$\|\overrightarrow{X_1} - \overrightarrow{m_1}\| \leq \|\overrightarrow{X_1} - \overrightarrow{m_2}\|$ holds good

Since,

$\left\|\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} - \begin{pmatrix} m_{11} \\ m_{12} \end{pmatrix}\right\| = (x_{11} - m_{11})^2 + (x_{12} - m_{12})^2 = 0 + 0 = 0$

Step 3. Update means.

$\overrightarrow{m}(2) = \frac{1}{|S_1(1)|} \sum_{x_i \in S_1(1)} \overrightarrow{X_i} = \frac{1}{2}(\overrightarrow{X_1} + \overrightarrow{X_3})$

$= \frac{1}{2}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$
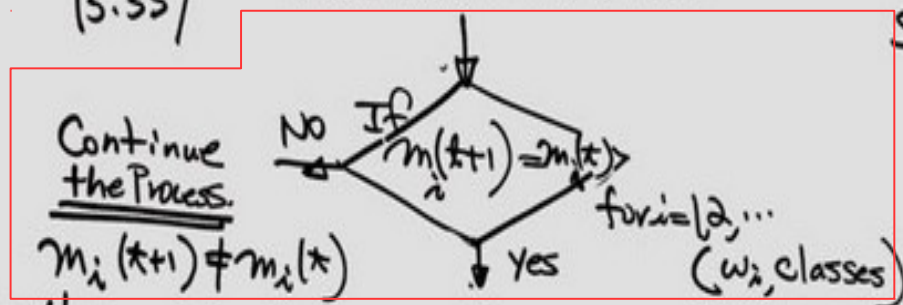
Harry Li, Ph.D.

K-mean

$$\overrightarrow{m_2}(3) = \frac{1}{|S_2(1)|} \sum_{j \in \{S_2(1)\}} \overrightarrow{X_j} = \frac{1}{18}\left(\overrightarrow{X_2} + \overrightarrow{X_4} + \cdots + \overrightarrow{X_{20}}\right)$$

$$= \begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}$$

CMPE 297 Maven, 2018, Harry Li 3.

Terminate the Process?

Continue the Process.
No If
$m_i(t+1) = m_i(t)$ for $i = 1, 2, \ldots$
$m_i(t+1) \neq m_i(t)$
Yes
$(\omega_2, \text{classes})$
Stop

Step 4
$$\left\| \overrightarrow{X_\ell} - \overrightarrow{m_1}(2) \right\| \leq \left\| \overrightarrow{X_\ell} - \overrightarrow{m_2}(2) \right\|, \text{ for } \ell = 1, 2, \ldots, 8$$

and
$$\left\| \overrightarrow{X_p} - \overrightarrow{m_2}(2) \right\| \leq \left\| \overrightarrow{X_p} - \overrightarrow{m_1}(2) \right\|, \text{ for } p = 9, 10, \ldots, 20$$

Hence, $S_1(2) = \{\overrightarrow{X_1}, \overrightarrow{X_2}, \ldots, \overrightarrow{X_8}\}$, $S_2(2) = \{\overrightarrow{X_9}, \overrightarrow{X_{10}}, \ldots, \overrightarrow{X_{20}}\}$

update mean,
$$\overrightarrow{m_1}(3) = \frac{1}{10}\left(\overrightarrow{X_1} + \overrightarrow{X_2} + \cdots + \overrightarrow{X_8}\right) = \begin{pmatrix} 1.25 \\ 1.33 \end{pmatrix}$$

$$\overrightarrow{m_2}(3) = \frac{1}{10}\left(\overrightarrow{X_9} + \overrightarrow{X_{10}} + \cdots + \overrightarrow{X_{20}}\right) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}$$

Check if $\overrightarrow{m_i}(t+1) \overset{?}{=} \overrightarrow{m_i}(t)$, $i = 1, 2$.

Since, not, continue,

Step 5. Check
$$\left\| \overrightarrow{X_\ell} - \overrightarrow{m_1}(3) \right\| \overset{?}{\leq} \left\| \overrightarrow{X_\ell} - \overrightarrow{m_2}(3) \right\|$$
for $\ell = 1, 2, \ldots, 8$.

Similarly, $\left\| \overrightarrow{X_p} - \overrightarrow{m_2}(3) \right\| \overset{?}{\leq} \left\| \overrightarrow{X_p} - \overrightarrow{m_1}(3) \right\|$

Both is Satisfied, No Regrouping, therefore,
$$\overrightarrow{m_i}(t+1) \overset{?}{=} \overrightarrow{m_i}(t), \text{ for } i = 1, 2.$$

YES! Converged, Stop.

Harry Li, Ph.D.

Use kmean.py To Compute Example

Example: w1: B1(x,y), …, B6(x,y), w2: B'1(x,y), …, B'10(x,y), V = (v1, v2, v3), where v1 = m11, v2=m10, v3=m01, find m1 and m2 by using kmean.py

Midterm Review

Midterm Review. Questions. Basic Concepts.

Colour Image Plane, Coordinate System Setup, Convolution (Shift + mult + Addition)

LoG

$\nabla^2 G(x,y)$ ← Kernels. (math Formulation) to a mask

Computer Vision Filter Ops.

$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$

$G(x,y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_x)^2 + (y-\mu_y)^2}{2\sigma^2}}$ Non-Skewed

Forward Difference
Backward ~
Central Difference

$(\sigma_x = \sigma_y)$

↓

Binarization $B(x,y)$

⌐ Flood Fill Algorithm → Feature Extraction → Cluster Analysis.

Implementation   Area/Perimeter   K-mean Algorithm.

Computation.   Moments.

$\vec{V_I} = (V_{I1}, V_{I2}, \ldots V_{IN})$

Programming Language.
Python

Question: On Kernel Design, → Cluster Analysis.

2D Convolution

Feature Extraction.

Qa. 14~20

Harry Li, Ph.D.