Clustering

Unsupervised Learning

In Supervised learning we are given a label associated with each data point, which helps in Classification of data.

In unsupervised we don't have any labels asso with it and tasked to find some structure

One such approach could be to divide the given into various distinct cluster based on some pr

Application

→ Use of to find pattern in stars

→ Dividing Customer base into segment on some

Extra

Can we labeled data for clustering?

The thing is that we can use whatever we wa to use, what matters is results. & results f labelled data is not as good as that of su Algorithm for labelled data such as regression They perform better.

Why?
Because Clustering tries to find relation b/w data points where as in labelled data we want relation b/w data point and its label so they perform poorly as they fail to do so.

for eg. if some person visits some ~~webs~~ e-comm

label when ~~the~~ buy something

Classifier tries to predict whether he will buy this time or not (or what he will buy)
whereas Clustering may try to cluster it type of user with its parameter based ~~on time~~ data points

K-Means Clustering

This is an iterative algorithm which perform algo ~~which~~ two steps in each.

first: randomly takes k - data point as center
~~Second~~ Then
for every iteration, colour the points based on its nearest centroid
Then move centroid to center of all colored points
then repeat until not a single print centroid change its color.

Can k-means only segment well separated data?

Even when it is not used to segment data, k-means clustering separate our data points. This use heavily on market segmentation

k-means clustering always give non overlapping k-segments

In Linear Regression we have cost function to reduce. In Clustering we take similar approach to minimize the avg. distance of all the clusters centroids with the datapoints in that cluster

Formally let

$c^{(i)}$ = index of cluster to which data point $x^{(i)}$ is assigned

μ_k = cluster centroid

$\mu_{c^{(i)}}$ = centroid of cluster to which $x^{(i)}$ is assigned

$$J = \frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - \mu_{c^{(i)}} \|^2$$

And our objective is to minimize this cost function

This cost function is also called distortion.

$$c_i = \underset{1 \le k \le K}{\operatorname{argmin}} (x^i - \mu_k)^2$$
$$1 \le i \le m$$

$$J(c, \mu) = \frac{1}{m} \sum_{i=1}^{m} (x^i - \mu_{c^i})^2$$

$\delta(i,k) = 1 \quad c^i = k$ Delta function
$= 0$ otherwise

Delta assignment help us to write similar functions

$$N_k = N_k + \delta(i,k)$$

N_k will contain no. of data points in cluster

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{m} x^i \delta(i,k)$$

$$\hat{\mu} = \{\hat{\mu}_1, \hat{\mu}_2, \ldots\}$$

$$\mu = \hat{\mu}$$

Go back to cluster assignment until convergence

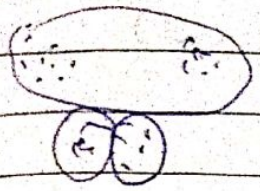
$$(J_k(c,\mu) - J_{k-1}(c,\mu)) \le \text{eps}$$

$\Downarrow$ convergence

k means of the given data point

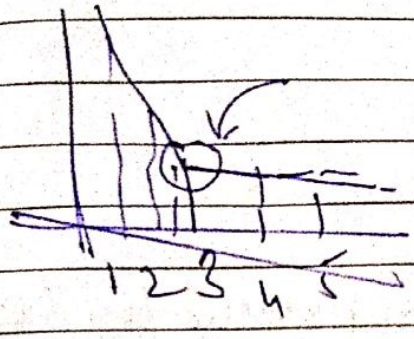
Page No.
Date

Local optima problem



Practical Solution: run it n-times with random initialization.
one with least error is choosen then.

Elbow method



K=3 is more accurate
After
After some "k" performance decreases.

Fuzzy C-mean Clustering

Fuzzy set

Normal set A, x
Either $x \in A$, or $x \notin A$

Fuzzy set
μ: membership function
$0 \leq \mu_A(x) \leq 1$
Not clear diversion in set.

Fuzzy logic
A statement is true with membership value

Neural Nets are considered to be part of soft computing. As it tell about being something with some membership value

fuzzy c-partition

Usual c-partition ($c \geq 2$, $c \in I$) of a set $S = \{x_1, \ldots x_n\}$ represented

$P(A_1, A_2 \ldots A_c)$

Defination

$A_i \neq \phi$

$A_i \cap A_j = \phi$

$\bigcup_{i=1}^{c} A_j = S$.

fuzzy c-partition

Give membership value, represend by (U, S) where $U_{n \times c}$ matrix

$U = ((U_{ij}))_{n \times c}$ where U_{ij} : memebership of ith point to jth fuzzy set

$1 \leq i \leq n$ $1 \leq j \leq c$

Satisfying the property stated here.

i) $0 \leq u_{ij} \leq 1$

ii) $\sum_{i=1}^{c} u_{ij} = 1 \quad \forall i = 1 \ldots n$

iii) $0 < \sum_{i=1}^{n} u_{ij} < n \quad \forall i = 1 \ldots c$

Let $r > 1$ let $v_i = \left(\sum_{i=1}^{n} x_i u_{ij}^{r}\right) / \left(\sum_{i=1}^{n} u_{ij}\right)^{r}$ mean for ith cluster

$J_r(U, S, A) = \sum_{i=1}^{n} \sum_{i=1}^{c} u_{ij}^{r} \| x_i - v_j \|_A (x_i - v_j)$

Minimize

A is positive definite matrix

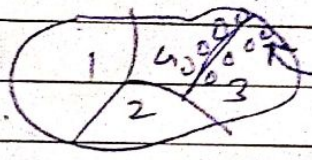
Given ~~E~~, S, R, A need to find optimal point.
So J is minimum.

FCM Algorithm

1 → we are given s, r, A and c
2 → start with fuzzy c-partition U_0
3 → Compute fuzzy c-partition V_j $j=1 \dots c$
4 → $$V_{ij} = \left(\sum_{K=1}^{C} \left(\frac{d_{ij}^2}{d_{iK}^2} \right)^{\frac{1}{r-1}} \right)^{-1} \quad 1 \le i \le n \quad 1 \le j \le C \quad r > 1$$

$$d_{ij}^2 = (x_i - v_j)' A (x_i - v_j)$$

5 → ~~Goto~~ if (convergence) break
else goto (3)



← high membership func for K=4
← lesser mem func for k=1, 2
highest for 3, lesser for 4

Data point near to a cluster mean will have higher membership function for that cluster.

Advantage
worst case performance is same has k-means and where data is less fuzzy tends to give more stable cluster means.

⊗ FCM
→ Same initialization as of k-means.

Page No.
Date

K-means

Cluster assigment

$$c^i = \arg\min_{1 \le k \le K} (x^i - \mu_k)^2$$

$1 \le i \le M$

FCM

$$c_k^i = \frac{1}{\sum_{j=1}^{K} \frac{(x^i - \mu_k)^2}{(x^i - \mu_j)^2}}$$

gives higher membership to cluster which are closer and lower to others

$$\sum_k c_k^i = 1.0$$

~~Loss~~

$$J(\mu, c) = \frac{1}{M} \sum_{i=1}^{M} (x^i - \mu_{c^i})^2$$

$$J(c, \mu) = \frac{1}{M} \sum_{i=1}^{M} \sum_{k=1}^{K} (x^i - c_k^i \mu_{c_k^i})^2$$

New means

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{M} x^i \delta(i,k)$$

$$\hat{\mu}_k = \frac{1}{M} \sum_{i=1}^{M} c_k^i x^i$$

$1 \le k \le K$

Minimization

$\mu = \hat{\mu}$

Step ① — ④

$\mu = \hat{\mu}$

Step ① — ④

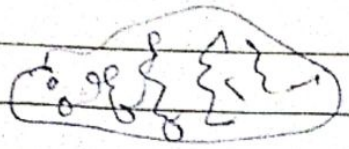
Model based Clustering

Principle :- Assume data points are from specific mathematical model

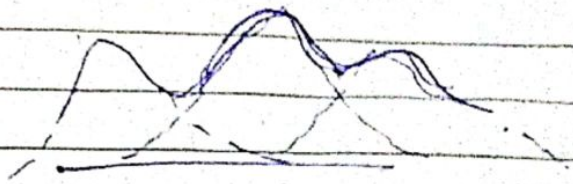
using training data point estimate model parameter

Gaussian Mixture Model

based on Gaussian method
data points are distributed as Mixture of Many Gaussians models



Top - view



Side view
can be showed with 3 Gaussian distribution.

1) Gaussian distribution
or univariate Gaussian distribution

$$g(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$x = [x_1, x_2]$ ⇒ 2D Multivariate GD

(μ_1, σ_1) (μ_2, σ_2)

$$g(x_1) = \frac{1}{\sqrt{2\pi}\,\sigma_1} e^{-\frac{1}{2}\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2}$$

$$g(x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_2} e^{-\frac{1}{2}\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2}$$

$$g(x) = g(x_1) \cdot g(x_2)$$

$$= \frac{1}{(2\pi)^{2/2}\,\sigma_1\sigma_2\,|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$x = [x_1, x_2]$$

$$\mu = [\mu_1, \mu_2]$$

probabilistic Constraints.

Should be followed by EM

N-dimension

$$g(x) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\bar{x}-\bar{u})^T \Sigma^{-1} (\bar{x}-\bar{u})}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \sigma_2^2 & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\lambda(x) = \sum_{k=1}^{K} w_k g(x, u_k, \Sigma_k)$$

$$\sum_{k=1}^{K} w_k = 1.0 \leftarrow \text{weight factor.}$$

$k = 3$ for provided eg.

Expectation Maximization Approach

- "X" Train + Initial Model 'λ'

unimportant

→ Iterative procedure.

after 1st iteration through λ

$\lambda = \lambda'$

weights generated from features.

λ' should be more representative of data patient

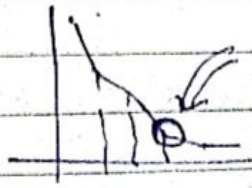
history is not carried away

Page No.
Date

Extras (Source Medium and tds.com)

2. Ways of finding k- in k-means algorithm.

1. Elbow method
It is very common method start to it tells when wr accuracy increase at much slower rate than it was before u have reached some optimal pt.



↑ - before this pt. Δ was high. b/w 2 consecutive values of k.

2) Silhouette Curve

Silhouette coefficient calculates the density of the cluster the cluster by generating a score for each sample based on the difference between the average intra cluster distance and mean nearest-cluster distance for that sample normalized by maximum value we select ~~best~~ k using best score calculated

~~16. De~~
DBSCAN: Density Based clustering

Disadvantage of k-Means / Partition based Clustering.

Unable to fit arbitrary-shape Clusters

→ Prone to outlier: outlier are assign to
→ Cluster, after that they pull mean vect
towards them which makes their cluster
prone to far point than near points.

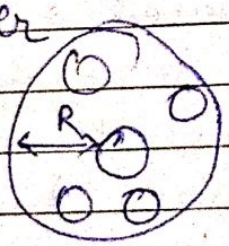
~~DBSCAN~~ Can
DBScan Density based Can find out any arbitary
shape cluster without being affected by
noise.

Density Based Spatial Clustering of Application

R: Radius of Neighbour hood
M: Min number of neighbour

Intitution / Idea.
If a point is in a cluster ~~of~~ it
should be near to lots of other points in
that cluster

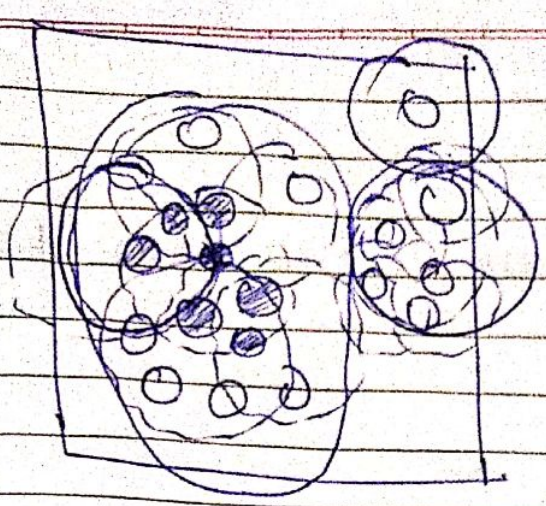
R



If a taking any point if in a given Radi
'R' it have more than M neighbour w
call it Dense ~~area or cluster~~ cluster.

3 types of Data point

→ Core ~~border~~
→ Border
→ Outlier



R = 2
M = 6.

Core : If there are M points or more in a given radius.

Border If there are points but less than M it is border

Outlier : if points in neighbourhood is zero

~~Soft Computing~~

Core points surrounded by border points generate arbitory shape

Disadvantage.

⇒ Need to select 2 parameter 'R' & 'M'

Advantage.

⇒ Arbitary shaped cluster
⇒ Robust to Classifier
⇒ Doesn't require specification of cluster.