INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

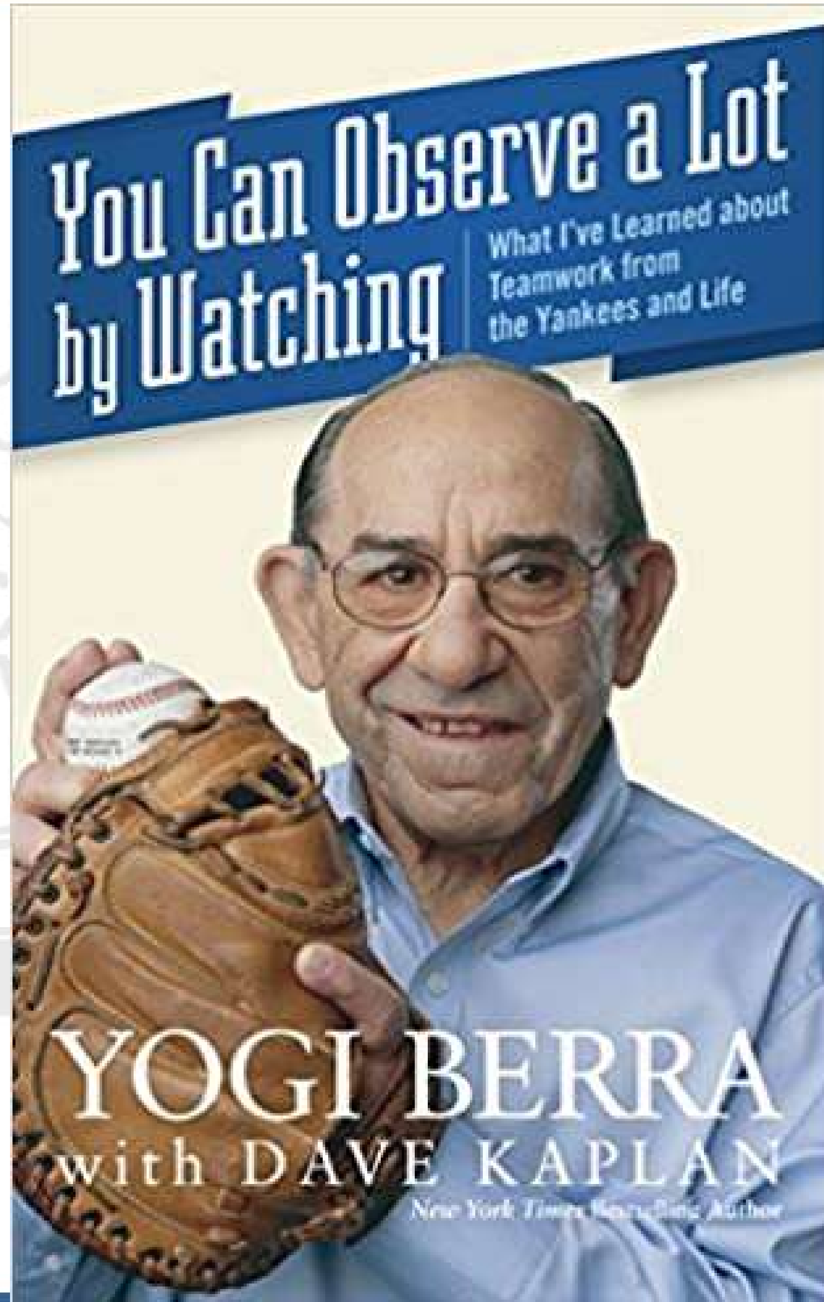
Clustering Methods and Applications

Sumit Kumar Yadav

Department of Management Studies

Tuesday 27th August, 2024

Clustering


You Can Observe a Lot by Watching
What I've Learned about Teamwork from the Yankees and Life
YOGI BERRA
with DAVE KAPLAN
New York Times Bestselling Author

Clustering

- View points as union of k disjoint clusters - $C_1, C_2, ..., C_k$
- Each point lies in exactly one

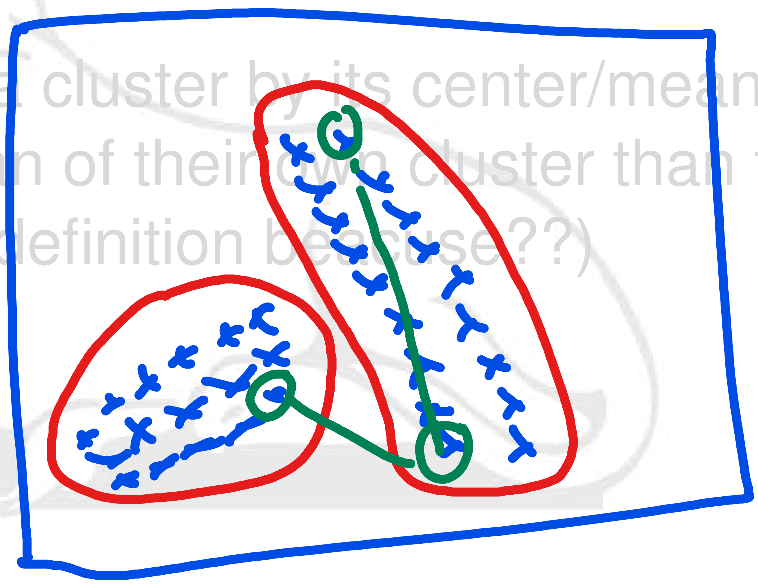
Clustering - Possible Goals

- Learn structure of data (e.g., that it consists of clusters or is low dimensional)
- Automatically organizing data
- Customer Segmentation - `https://www.kaggle.com/code/kushal1996/customer-segmentation-k-means-analysis`

Clustering

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points

Definition Attempt 2 - Represent a cluster by its center/mean. Points in a cluster are closer to center/mean of their own cluster than to the mean of other clusters. (Circular definition because??)



Clustering

Definition Attempt 1 - "subset of points that are closer to each other than to all other data points

Definition Attempt 2 - Represent a cluster by its center/mean. Points in a cluster are closer to center/mean of their own cluster than to the mean of other clusters. (Circular definition beacuse??)

k-means Clustering problem

- Let the points be $x_1, x_2, ..., x_n$
- Mean of the j^{th} cluster =

$$\boxed{c_j = \frac{1}{m_j} \sum_{i \in C_j} x_i}$$

$c_1, c_2, \ldots, c_k$

(j^{th})

m_j is the number of points in the j^{th} cluster

- Define cost of a cluster as - sum of squared distance from the points to the mean -

$$\sum_{i \in C_j} ||x_i - c_j||^2 \rightarrow (j^{th})$$

- k-means problem : Partition points into k clusters so as to minimize sum of cluster costs - $\sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - c_j||^2$

k-Means algorithm

$\geq 10^{9900}$ seconds

$\#$

many many centuries

Is "k-means Clustering problem" a difficult problem to solve? Think about 10000 points and 10 groups.

- A heuristic to solve k-means Clustering problem
- Maintain clusters $C_1, C_2, ..., C_k$
- Compute the cluster centers for these clusters
- Iteration - For each point, assign it to the c_j that it is closest to. Update $C_1, C_2, ..., C_k$ and proceed to the next iteration

Example , k-Means

$k = 3$

$(10-10)^2 + (11-12.67)^2 + (12-12.67)^2 + (15-12.67)^2 + (18-27)^2 + \ldots$

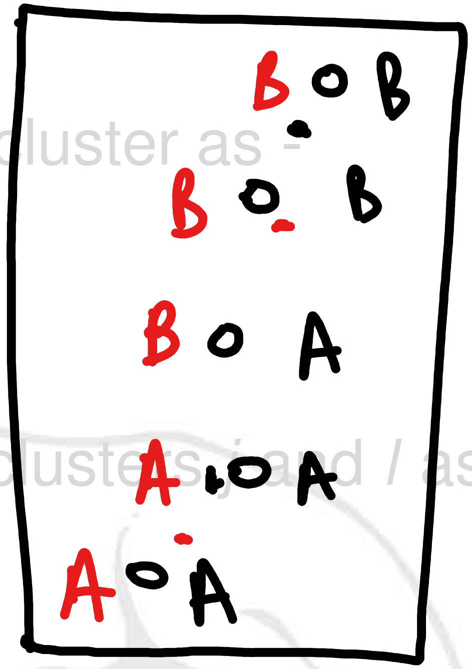
$(10-10)^2 + (11-10)^2 + (12-12.67)^2 + \ldots + (18-12.67)^2 + \ldots$

A	B	C	C	C	C	C	C
10	11	12	15	18	20	30	40
A	B	B	B	C	C	C	C
A	A	B	B	B	C	C	C
A	A	A	B	B	B	C	C
A	A	A	B	B	B	C	C

A	B	C
10	11	22.5
10	12.67	27
10.5	15	30
11	17.67	35
11	17.67	35

Finding the value of K

- Elbow Method
- DB Index
  Define cluster dispersion for the j^{th} cluster as -
  $$d_j = \sqrt{\frac{1}{m_j}\sum_{i \in C_j} ||x_i - c_j||^2}$$
- Define cluster similarity between 2 clusters j and l as -
  $$S_{jl} = \frac{d_j + d_l}{||c_j - c_l||}$$
- $$V_{DB} = \frac{1}{K}\sum_{i=1}^{K} \max_{l \neq i} S_{il}$$



Finding the value of K

- Elbow Method
- DB Index

  Define cluster dispersion for the j^{th} cluster as -

$$d_j = \sqrt{\frac{1}{m_j} \sum_{i \in C_j} ||x_i - c_j||^2}$$

- Define cluster similarity between 2 clusters j and l as -

$$S_{jl} = \frac{d_j + d_l}{||c_j - c_l||}$$

- $V_{DB} = \frac{1}{K} \sum_{i=1}^{K} \max_{l \neq i} S_{il}$

Model Based Clustering

Gaussian Mixture Models

Pre-requisites for GMM

- Normal distribution
- Multivariate normal distribution
- Probability Basics
- Maximum Likelihood

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Case - 1 : Assume these parameters to be known,

$p_1 = 0.8, p_2 = 0.2, p_{h1} = 0.9, p_{h2} = 0.75$

We do the experiment once and observe 95 heads. What is the probability it came from coin 1?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.
Case - 1 : Assume these parameters to be known,
$p_1 = 0.8, p_2 = 0.2, p_{h1} = 0.9, p_{h2} = 0.75$
We do the experiment once and observe 95 heads. What is the probability it came from coin 1?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.
Case - 1 : Assume these parameters to be known,
$p_1 = 0.8, p_2 = 0.2, p_{h1} = 0.9, p_{h2} = 0.75$
We do the experiment once and observe 95 heads. What is the probability it came from coin 1?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.

Case - 2 : The parameters are not known, all we observe is data from several trials of this experiment. Let us say that the observations are - 19,24,89,88,92,16,94,86,21,92

What are the guesses we would like to make for the parameters?

Can you group the data points into 2 and say one group came from coin 1, and other came from coin 2?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.
Case - 2 : The parameters are not known, all we observe is data from several trials of this experiment. Let us say that the observations are - 19,24,89,88,92,16,94,86,21,92
What are the guesses we would like to make for the parameters?
Can you group the data points into 2 and say one group came from coin 1, and other came from coin 2?

Analogous problem

There are 2 coins. We pick coin 1 with probability p_1. We pick the other coin with probability $p_2 = 1 - p_1$. We then toss it 100 times. The chances of heads for coin 1 and coin 2 are p_{h1} and p_{h2} respectively.
Case - 2 : The parameters are not known, all we observe is data from several trials of this experiment. Let us say that the observations are - 19,24,89,88,92,16,94,86,21,92
What are the guesses we would like to make for the parameters?
Can you group the data points into 2 and say one group came from coin 1, and other came from coin 2?

GMM

Comparison with the analogous case - In the background, we don't have coins generating the data, but we have normal distributions, and we are interested in making the best guess for the parameters of the normal distributions along with the probability that a randomly chosen data point will come from.

GMM

Comparison with the analogous case - In the background, we don't have coins generating the data, but we have normal distributions,and we are interested in making the best guess for the parameters of the normal distributions along with the probability that a randomly chosen data point will come from.

Example

Consider the 30 data points -
109, 10079, 8, 106, 9898, 7, 117, 9920, 11, 84, 10034, 11, 116, 9951, 10, 117, 9980, 13, 115, 9970, 11, 94, 9948, 11, 95, 12, 106, 12, 8, 7
Alright, it is too messy, let us organize it better maybe.

Example

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7

If we have to think of this as data coming from 3 normal distributions, what could be some sensible parameters of the data generation process.

Example - How about this??

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7
mean	**10**	**100**	**1**
sigma	**5**	**50**	**0.5**
probability	**1/3**	**1/3**	**1/3**

Example - How about this one??

S.No	Set-1	Set-2	Set-3
1	109	10079	8
2	106	9898	7
3	117	9920	11
4	84	10034	11
5	116	9951	10
6	117	9980	13
7	115	9970	11
8	94	9948	11
9	95		12
10	106		12
11			8
12			7
mean	**100**	**10000**	**10**
sigma	**10**	**100**	**2**
probability	**10/30**	**8/30**	**12/30**

Example

We will not get into how these parameters are estimated.
We will just keep in mind that it is done with an approach that is similar to what is done in Logistic Regression or SoftMax. Maximum Likelihood approach
This is usually done using an Iterative algorithm called Expectation Maximization algorithm

GMM

In the example. the data was 1 dimensional. It will not always be the case.

Welcome Multi-variate normal distribution

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

The above equation is density of a D-dimensional normal distribution, Σ is the variance-covariance matrix

GMM

In the example. the data was 1 dimensional. It will not always be the case.

Welcome Multi-variate normal distribution

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

The above equation is density of a D-dimensional normal distribution,
Σ is the variance-covariance matrix

GMM

In the example. the data was 1 dimensional. It will not always be the case.

Welcome Multi-variate normal distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\}$$

The above equation is density of a D-dimensional normal distribution, Σ is the variance-covariance matrix

GMM

So, if we want to make 3 clusters from the data,
we would think of the data as a simulation of a data generation process going on in the background. The data generation process will from 3 normal distributions with their respective parameters. Each normal distribution will be picked with some probability.
So, the parameters will be -
p_1, μ_1, Σ_1
p_2, μ_2, Σ_2
p_3, μ_3, Σ_3
with the condition that $p_1 + p_2 + p_3 = 1$

Deciding the value of k

Two ways - AIC and BIC,
pick the one for which this is minimum.
Again, we will not be getting into details of these.