Machine Learning



Amin Golzari Oskouei

a.golzari@azaruniv.ac.ir a.golzari@tabrizu.ac.ir https://github.com/Amin-Golzari-Oskouei

Azarbaijan Shahid Madani University 2023

Unsupervised Learning : Clustering

Table of Contents

2

- □ Unsupervised Learning
- □ Clustering
- □ K-Means Algorithm
- □ Improved clustering algorithm
- □ Two-part generator algorithm
- □ Hierarchical clustering

Reminder: Supervised learning

□ Supervised Learning. For each example, the correct answer is given.

$$\{ (x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)}) \}$$

Training set

□ Types Of Supervised Learning.

3

- Regression: Estimation of a continuous quantity.
- □ Classification: Estimation of a discrete quantity.



Unsupervised Learning

4

□ Unsupervised Learning. Not knowing the correct answers.

□ Goal. Structure detection in input data



Applications of clustering: grouping related news

5

					THE BP Kills M	facondo, But Its ×	- your prairies - after		
Google News		and the second second		- - ×	← → C	O blogs.wsj.com/source,	/2010/09/20/bp-kills-macondo-but-its-legacy-live	s-on/	
← → C ③ news.ç	google.com			12 A	THE WALL Dig	tal Network WSJ.com M	arketWatch BARRONS AIThingsDigital FINS SmartM	105y More▼	SEARCH
Web Images Videos M	taps News Shopping Gmail more 🔻	andrew	yantakng@gmail.com <u>Web History</u> <u>Settings</u> ▼	Sign out	THE	WALL STREET JOURNAL		Log In • Register For Free • Sul	scribe Now, Get 2 Weeks Free
Google nev	VS	Search News Search th	he Web				TUE	UDCE	
		Advanced	news search U.S. edition V Add a	section a	≪ PR	EVIOUS	I HE SO	URCE	NEXT >>
Top Stories Deepwater Horizon	Top Stories	☆	Recent Recession officially anded in June 2009	A10	Fin	ancial Transport	Leisura Insurance Oil & Gas	Sport Caudit on the	Retting Technology
Fed meeting Foreign exchange	White House official denies Tea Party-foc CNN International - Ed Henry - 1 hour ago	used ad campaign	CNNMoney - Chris Isidore - 39 minutes ago	A 8	Se	INCES		Web	county recentrely
Lindsay Lohan	Democratic sources say the White House is not cons tying Republicans to the Tea Party. Washington (CNN official sharply denied a report that claims President C	Idering an ad campaign I) A top White House Obama's political	USA Today - Gerry Broome - 5 minutes ago 'Explain what you want from us.' reads front-	A.B.	BP	Kills Macondo	o, But Its Legacy Lives On	Search The Source	SEARCH
Tom Brady Toronto	advisers are weighing a national Tea Party is misplacing the blame, former President E New York Daily News	Bill Clinton claims CNN Interna.	page editorial msnbc.com - Olivia Torres - 10 minutes ago			Article Comments (2	1		THE SOURCE HOME PAGE
International Film Festival	GOP tea party backer defends Christine O'Donnell Th Atlanta Journal Constitution - Politics Daily - MyFox V	e Associated Press Vashington DC - Salon	Crisis response: Pakistan floods					About The Source	
Paris Hilton Iran	US Stocks Climb After Recession Called	Over,	San Francisco Bay Area - Edit			Email 🔛 Print 💽 Perm	anox Line Z L3 W - More - Yeod	The Payment is MP Learn France's	
Starred 🕸	Homebuilders Gain MarketWatch - Kristina Peterson - 16 minutes ago NEW YOPK (MarketWatch) - US stocks climbad Mo	RECESSIO	Clorox » Bay Biz Buzz: Clorox close to selling STP.	公 回	By Ja	mes Herron		fire analysis of the day's big busin stories. It is edited by Lauren Mills	based in
San Francisco Bay Area	after a key nonprofit organization officially called the re investors a boost of confidence in the gradual econom	ecession over, giving tic recovery.	Armor All San Jose Mercury News - 48 minutes ago		that t	he Macondo well that		London.	3.4
U.S.	Los Angeles Times Downturn Was Longest in Decades, Panel Confirms N	New York Times	Google's official beekses a keeps the compa	nny vir 💌	barre	is of oil into the Gulf of		Most Recent	
Sci/Tech More Top Stories	Wall Street Journal - AFP - CNN - USA Today all 276 news articles »		San In Americary News - Bruce Newman -		perm	anently sealed, but the		Articles Comments	
Spotlight Health	Deepwater Horizon » BP Oil Well, Site of National Catastrophe, Venite For July Melanar, 22 minutes and	Dies at One	Jon Sylvia » Martinez man still unconscious as police	12 P	and t	he wider oil industry for		1. Who Needs Plaza II Any	rway
Sports Entertainment	The BP oil well, site of the Deepwater Horizon explosi oil spill in US history, died today at one year old	on that led to the worst	San Jose Mercury News - Robert Salonga - 48 minutes ago - all 6 articles »		The	nost immediate worry	Contraction de	2. Will Banks Be Forced to Arms?	Split Retail And Banking
ill news	Weiss Doubts BP Would End Operations in Gulf dis CNN International - Wall Street Journal (bloc) - The Gu	The Associated Press exico: Video Bloomberg vardian -	Spotlight		for Bi	P and its shareholders w the authorities will	41	3. Timing of Ratings Awar	d Intriguing
leadlines mages	New York Times all 2,292 news articles »	K	Sarkozy rages at EU 'humiliation' Financial Times - Peggy Hollinger -	sta 💌	appo BD's	ntion blame for the spill.	Associated I	4 BP Kills Macondo, But I	ts Legacy Lives On
			Sep 16. 2010	<u> </u>	STVA3	d responsibility armss	re boat response crews befied the blazing remnants of the off shore Parameter Henroepen April 21, 2010.	5. We Already Need a Sen	sel to Rasel III
	Allen: Well is dead, but m ×				BP oil spill cost hits nearl × +	A STREET			
	← → C ③ edition.cnn.co	om/20.0/US/09/20/gulf.oil.disaster	r/		← → C ③ www.guardian.co.uk/en	vironment/2010/sep/20/l	op-oil-spill-dee		
	EDITION: INTERNATIONAL	U.S. MÉXICO ARABIC	GANIA		Mobile site Sign in Register	A Text larger · smaller	About us 🔻		
	Set earlion preterence								
	Home Video World	ILS Africa Asia Europe Latir	n America Middle Fast Business We		guardian.co.uk				
					News Sport Comment Cul	ture Business Money	Life & style		
	All	to the other building of	L O . K O		Business BP				
	Allen: well	is dead, but mud	ch Gulf Coast		BP oil spill cost hit	s nearly \$10br	1		
	By the CNN Wire Staff	ins			BP has set up a \$20bn compe Horizon disaster, which has so	nsation fund after the De far paid out 19,000 clai	eepwater ms totalling		
	September 20, 2010 Updated	1317 GMT (2117 HKT)			more than \$240m				
			and the second		Julia Kollewe				
			and the second		guardian.co.uk, Monday 20 September Article history	2010 08.33 BST			
		Click to play							
		Chick to play			The second second				
		I see 1	at the						
						Tr. The			
						The second			
	What next for Gulf oil spill?					a total			
	STORY HIGHLIGHTS	(CNN) The ruptured Macondo	well, a mile under the Gulf of		BP's costs for the Deepwater Horizon of Ho/Reuters	disaster have hit \$10bn. Photo	graph:		

K-means Clustering Algorithm

6

□ Its an iterative clustering algorithm

□ It choosing K points randomly as the center of the clusters.

□ Repeat the following steps:

- Assign each data to a cluster with the closest center.
- Update the centroid of each cluster by averaging the data assigned to that cluster

□ Stop time : When no data changes its cluster in an iteration.



K-means Clustering Algorithm

□ Its an iterative clustering algorithm

□ It choosing K points randomly as the center of the clusters.

□ Repeat the following steps:

- Assign each data to a cluster with the closest center.
- Update the centroid of each cluster by averaging the data assigned to that cluster

□ Stop time : When no data changes its cluster in an iteration.



K-means Clustering Algorithm

□ Its an iterative clustering algorithm

8

□ It choosing K points randomly as the center of the clusters.

□ Repeat the following steps:

- Assign each data to a cluster with the closest center.
- Update the centroid of each cluster by averaging the data assigned to that cluster

□ Stop time : When no data changes its cluster in an iteration.





Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

9

10



Assign each data to the nearest cluster 0 0

Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

12



13



14







16



Assign each data to the nearest cluster 0 0

18





Convergence : None of the data clusters were changed .

19

Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

20

□ Inputs.

□ Number Of Clusters : K

☐ Training set

$$\{x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(m)}\}$$

□ Note: In the training set, no label is assigned to the data.

□ Note: There is no need to add the feature $x_0 = 1$ in the clustering.



21

randomly initialize *K* cluster centroids $\mu_1, \mu_2, ..., \mu_k \in \mathbb{R}^n$

repeat

for i = 1 to m $c^{(i)} = \arg\min_{k} ||x^{(i)} - \mu_k||$

for k = 1 to KUpdating the cluster center μ_k = average of points assigned to cluster k

Assigning data to clusters

22

```
centroids = np.random.random((K, n))
while True:
    for i in range(m):
         c[i] = np.argmin(np.linalg.norm(X[i] - centroids, axis=1))
    for k in range(K):
        centroids[k] = np.mean(X[c == k], axis=0)
```

Clustering: Objective function

Objective function

24

□Symbols.

- \square μ_k : center of the cluster k
- \Box c⁽ⁱ⁾ : Cluster number assigned to data x⁽ⁱ⁾
- $\mu_{c}(i)$: Cluster center assigned to data $x^{(i)}$



□ Objective function

$$J(c^{(1)}, c^{(2)}, \dots, c^{(m)}, \mu_1, \mu_2, \dots, \mu_k) = \frac{1}{m} \sum_{i=1}^m \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

randomly initialize <i>K</i> cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$								
repeat {								
for $i = 1$ to m $c^{(i)} = \arg\min_{k} x^{(i)} - \mu_k $	Minimization of the objective with respect to $c^{(i)}$ parameters							
for $k = 1$ to K μ_k = average of points assigned to cluster k	Minimization of the objective with respect to μ parameters							
}								

Machine I

Setting initial value for cluster centers

27



Setting initial value for cluster centers

\Box Setting initial value. [$K \leq m$]

28

Selection of K training samples randomly
 Assigning center of each cluster to K selected samples

A center may be selected in such a way that no data is assigned to it.



Setting initial value for cluster centers : Better way

\Box Setting initial value. [$K \leq m$]

- Choosing k amount of training samples randomly
- Assigning center of each cluster to K selected samples

C = np.random.permutation(X)[:K]



Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

نیشام یریگدای -یدنبهشوخ -یوضر رصان دیس - ۱۳۹۷

Global optimization and Local optimization

30

Local optimization



Global optimization



Avoid local optima

```
for t = 1 to MAX
ł
   randomly initialize cluster centroids \mu_1, \mu_2, \dots, \mu_k
   run K-means to get c^{(1)}, c^{(2)}, ..., c^{(m)}, \mu_1, \mu_2, ..., \mu_k
   compute cost function J(c^{(1)}, c^{(2)}, ..., c^{(m)}, \mu_1, \mu_2, ..., \mu_k)
pick clustering with minimum cost
```

Determine the number of clusters

What is the right value for K?

33



Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

نیشام یریگدای -یدنبهشوخ -یوضر رصان دیس - ۱۳۹۷

What is the right value for K?

34



What is the right value for K?

35



Determining the appropriate number of clusters: Elbow Method



Improving clustering

Improving clustering with post-processing of clusters



Division.

Splitting a cluster with the highest error into

two clusters.

By running the K-means on the data of this cluster with a value of K = 2



Merge the two closest clusters
Merging two clusters with minimal

increase in total error.





Algorithm of bipartite generator K-Means

□ Bipartite algorithm.

□ Start with a cluster containing all the data.

Choose one cluster at a time :

- Divide the selected cluster into two clusters using K-means algorithm.
- Calculate the total clustering error.
- Choose the clustering with the least error.

□ Repeat the process above until you reach the desired number of clusters.

Algorithm of bipartite generator K-Means

Start with all the points in one cluster while the number of clusters is less than K measure the total error for every cluster perform K-means clustering with k=2 on the given cluster measure the total error after splitting choose the cluster split that gives the lowest error

Hierarchical clustering

41

□ Hierarchical clustering.

☐ First, merge very similar data.

□ Gradually create larger clusters by merging smaller clusters.

□ Algorithm.

At first, each data represents a cluster.
Repeat the following steps:

Choose the two closest clusters each time.

- Merge those two clusters into a new cluster.
- Stop: when only one cluster remains.

□ Create a tree diagram containing a wide range of clusters.



Hierarchical clustering

42

□ Hierarchical clustering.

☐ First, merge very similar data.

□ Gradually create larger clusters by merging smaller clusters.

□ Algorithm.

At first, each data represents a cluster.
 Demost the following stores.

□ Repeat the following steps:

- Choose the two closest clusters each time.
- Merge those two clusters into a new cluster.
- Stop: when only one cluster is left.

□ Create a tree diagram containing a wide range of clusters.



Hierarchical clustering

43

 \Box How to define the closest two clusters.

□ Criteria for determining clusters similarity.

- □ Nearest pair (one-link clustering)
- □ Farthest pair (all-link clustering)
- □ Average distance of all pairs
- □ «Enter» method (least variance, like k-means)

□ Different criteria create different clusters.



Clustering of digits

44



Image compression

45



Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

نیشام یریگدای -یدنبهشوخ -یوضر رصان دیس - ۱۳۹۷

Image compression

46





Machine Learning : Unsupervised Learning-Amin Golzari Oskouei - 2023

نیشام یریگدای -یدنبهشوخ -یوضر رصان دیس - ۱۳۹۷



47

□ Unsupervised learning, Finding structure in data

□ Clustering Grouping similar data

- K-mean Algorithm
 - Easy implementation
 - Slow for very large data sets
 - Possibility of getting stuck in the local optimum
- □ Post-processing of clusters : Splitting and merging clusters
- Algorithm of bipartite generator K-Means
 - Better clustering than the K-Means algorithm
- □ Hierarchical clustering algorithms