

Clustering

Stephen Boyd

EE103

Stanford University

September 28, 2017

Outline

Clustering

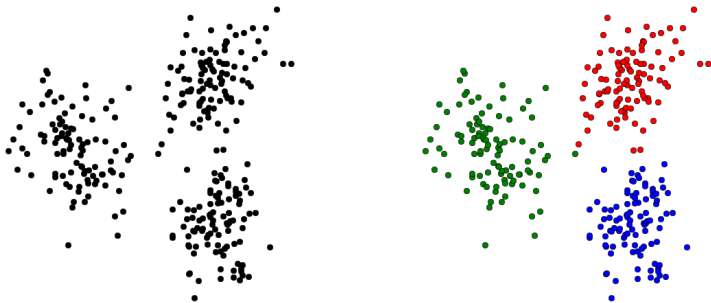
Algorithm

Examples

Applications

Clustering

- ▶ given N n -vectors x_1, \dots, x_N
- ▶ goal: partition (divide, cluster) into k groups
- ▶ want vectors in the same group to be close to one another



Example settings

- ▶ topic discovery and document classification
 - x_i is word count histogram for document i
- ▶ patient clusering
 - x_i are patient attributes, test results, symptoms
- ▶ customer market segmentation
 - x_i is purchase history and other attributes of customer i
- ▶ color compression of images
 - x_i are RGB pixel values
- ▶ financial sectors
 - x_i are n -vectors of financial attributes of company i

Clustering objective

- ▶ $G_j \subset \{1, \dots, N\}$ is group j , $j = 1, \dots, k$
- ▶ c_i is group that x_i is in: $x_i \in G_{c_i}$
- ▶ group *representatives*: n -vectors z_1, \dots, z_k
- ▶ clustering objective is

$$J = \frac{1}{N} \sum_{i=1}^N \|x_i - z_{c_i}\|^2$$

mean square distance from vectors to associated representative

- ▶ J small means good clustering
- ▶ goal: choose clustering c_i and representatives z_j to minimize J

Outline

Clustering

Algorithm

Examples

Applications

Partitioning the vectors given the representatives

- ▶ suppose representatives z_1, \dots, z_k are given
- ▶ how do we assign the vectors to groups, *i.e.*, choose c_1, \dots, c_N ?
- ▶ c_i only appears in term $\|x_i - z_{c_i}\|^2$ in J
- ▶ to minimize over c_i , choose c_i so $\|x_i - z_{c_i}\|^2 = \min_j \|x_i - z_j\|^2$
- ▶ *i.e., assign each vector to its nearest representative*

Choosing representatives given the partition

- ▶ given the partition G_1, \dots, G_k , how do we choose representatives z_1, \dots, z_k to minimize J ?
- ▶ J splits into a sum of k sums, one for each z_j :

$$J = J_1 + \dots + J_k, \quad J_j = (1/N) \sum_{i \in G_j} \|x_i - z_j\|^2$$

- ▶ so we choose z_j to minimize mean square distance to the points in its partition
- ▶ this is the mean (or average or centroid) of the points in the partition:

$$z_j = (1/|G_j|) \sum_{i \in G_j} x_i$$

k -means algorithm

- ▶ alternate between updating the partition, then the representatives
- ▶ a famous algorithm called *k-means*
- ▶ objective J decreases in each step

given $x_1, \dots, x_N \in \mathbf{R}^n$ and $z_1, \dots, z_k \in \mathbf{R}^n$.

repeat

Update partition. assign i to G_j , $j = \operatorname{argmin}_{j'} \|x_i - z_{j'}\|^2$

Update centroids. $z_j = \frac{1}{|P_j|} \sum_{i \in P_j} x_i$

until z_1, \dots, z_k stop changing

Convergence of k -means algorithm

- ▶ J goes down in each step, until the z_j 's stop changing
- ▶ but (in general) the k -means algorithm *does not find the partition that minimizes J*
- ▶ k -means is a *heuristic*: it is not guaranteed to find the smallest possible value of J
- ▶ the final partition (and its value of J) can depend on the initial representatives
- ▶ common approach:
 - run k -means 10 times, with different (often random) initial representatives
 - take as final partition the one with the smallest value of J

Outline

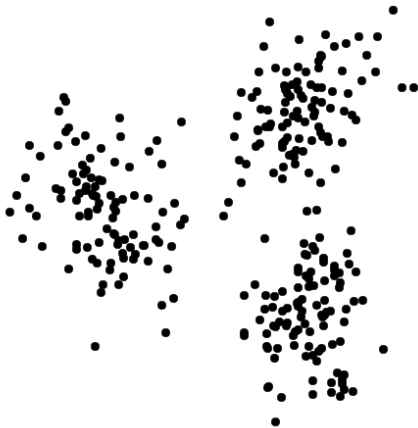
Clustering

Algorithm

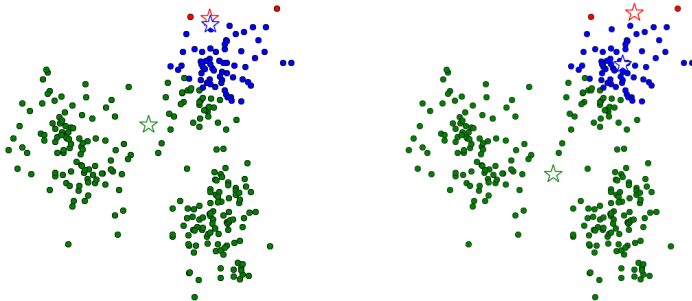
Examples

Applications

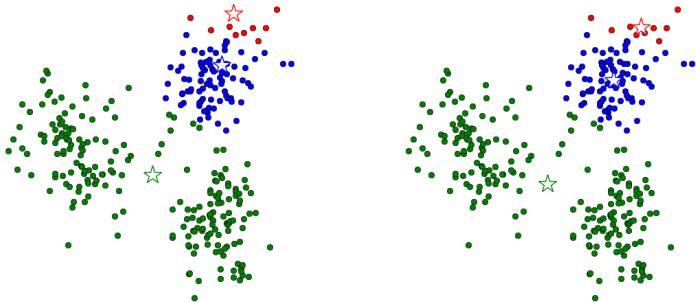
Data



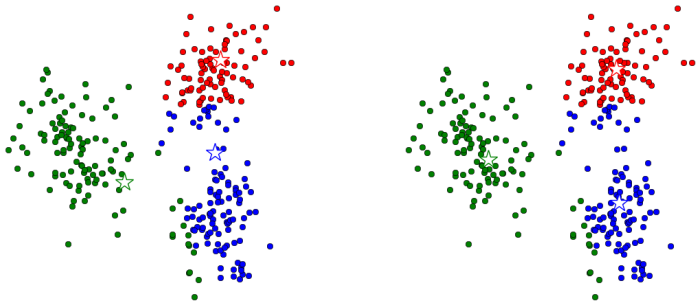
Iteration 1



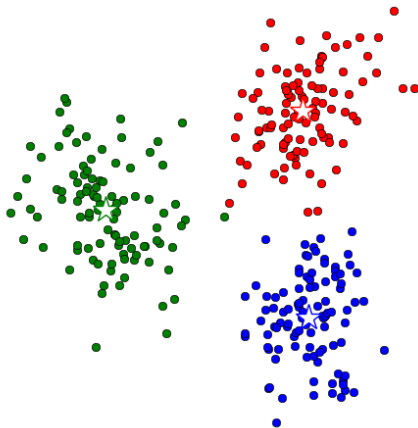
Iteration 2



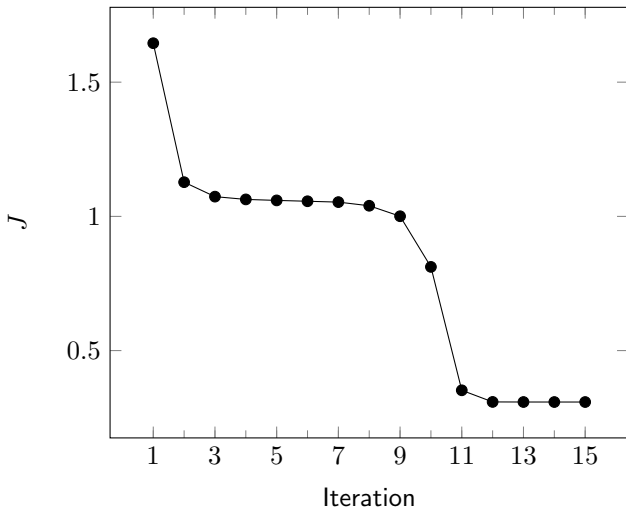
Iteration 10



Final clustering



Convergence



Outline

Clustering

Algorithm

Examples

Applications

Color compression

- ▶ 3-vectors x_1, \dots, x_N represent RGB values of each pixel in an image
- ▶ in 24-bit color representation $(x_i)_m \in \{1, 2, \dots, 256\}$
- ▶ total of $256^3 \approx 1.7 \times 10^7$ possible colors
- ▶ compress color vectors to k colors using k -means

Compressed image for various values of k

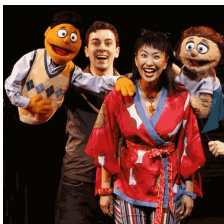
$k=4$



$k=8$



$k=64$

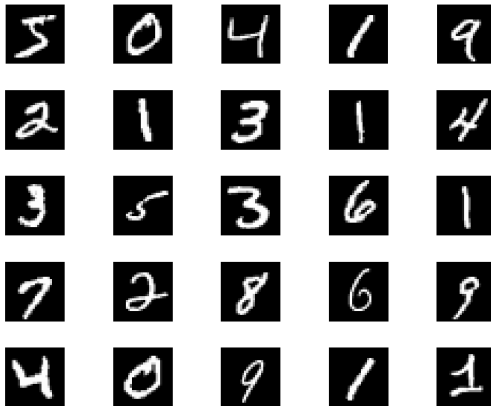


Original image



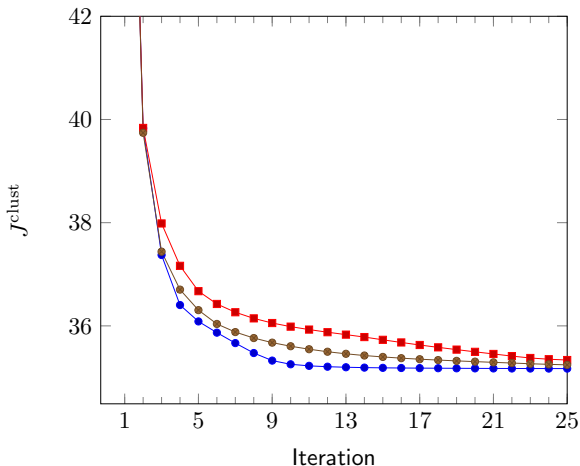
Handwritten digit image set

- ▶ MNIST images of handwritten digits (via Yann Lecun)
- ▶ $N = 60,000$ 28×28 images, represented as 784-vectors x_i
- ▶ 25 examples shown below

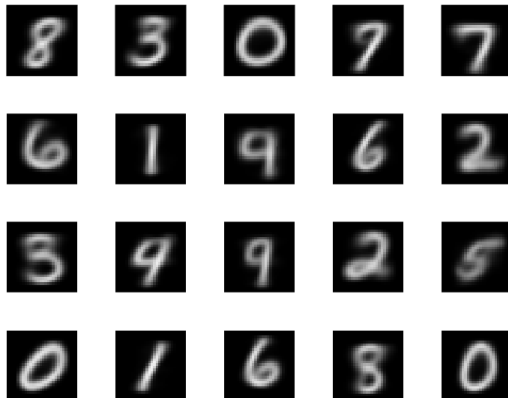


k -means image clustering

- ▶ $k = 20$, run 20 times with different initial assignments
- ▶ convergence shown below (including best and worst)

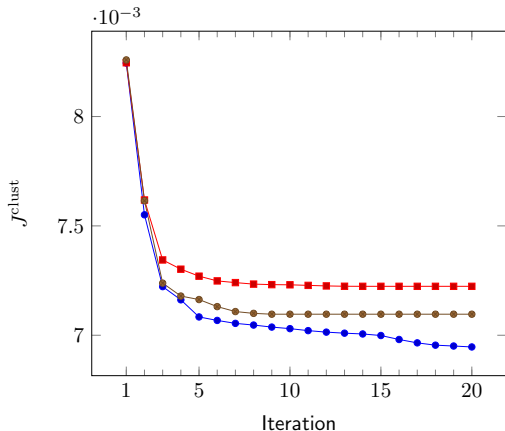


Group representatives, best clustering



Topic discovery

- ▶ $N = 500$ Wikipedia articles, word count histograms with $n = 4423$
- ▶ $k = 9$, run 20 times with different initial assignments
- ▶ convergence shown below (including best and worst)



Topics discovered

- for clusters 1–3, words with largest representative coefficients

Cluster 1		Cluster 2		Cluster 3	
Word	Coef.	Word	Coef.	Word	Coef.
fight	0.038	holiday	0.012	united	0.004
win	0.022	celebrate	0.009	family	0.003
event	0.019	festival	0.007	party	0.003
champion	0.015	celebration	0.007	president	0.003
fighter	0.015	calendar	0.006	government	0.003
bout	0.015	church	0.006	school	0.003
title	0.013	united	0.005	american	0.003
ali	0.012	date	0.005	university	0.003
championship	0.011	moon	0.005	city	0.003
bonus	0.010	event	0.005	attack	0.003

Topics discovered

- ▶ for clusters 1–3, articles closest to representative
- ▶ *Cluster 1 (21 articles)*: Floyd Mayweather, Jr; Kimbo Slice; Ronda Rousey; José Aldo; Joe Frazier; Wladimir Klitschko; Saul Álvarez; Gennady Golovkin; Nate Diaz; Conor McGregor.
- ▶ *Cluster 2 (43 articles)*: Halloween; Guy Fawkes Night; Diwali; Hanukkah; Groundhog Day; Rosh Hashanah; Yom Kippur; Seventh-day Adventist Church; Remembrance Day; Mother's Day.
- ▶ *Cluster 3 (189 articles)*: Mahatma Gandhi; Sigmund Freud; Carly Fiorina; Frederick Douglass; Marco Rubio; Christopher Columbus; Fidel Castro; Jim Webb; Genie (feral child); Pablo Escobar.