3. Norm and distance

- norm, distance, angle
- application examples
- standard deviation, correlation
- complexity
- clustering

Euclidean norm

Euclidean norm of vector $a \in \mathbb{R}^n$:

$$\|a\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_n^2} = \sqrt{a^T a}$$

- reduces to absolute value $|a|$ when $n = 1$
- measures the magnitude of a
- examples

$$\left\| \begin{bmatrix} 2 \\ -1 \\ 2 \end{bmatrix} \right\| = \sqrt{9} = 3, \quad \left\| \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\| = 1$$

Properties

Positive definiteness

$$\|a\| \geq 0 \quad \text{for all } a, \quad \|a\| = 0 \quad \text{only if } a = 0$$

Homogeneity

$$\|\beta a\| = |\beta| \|a\| \quad \text{for all vectors } a \text{ and scalars } \beta$$

Triangle inequality

$$\|a + b\| \leq \|a\| + \|b\| \text{ for all vectors } a \text{ and } b \text{ of equal length}$$

- any real function that satisfies these properties is called a (general) *norm*
- Euclidean norm is often written as $\|a\|_2$ to distinguish from other norms
- examples are the one-norm and infinity-norm

$$\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n|$$
$$\|a\|_\infty = \max\{|a_1|, |a_2|, \ldots, |a_n|\}$$

Norm of block vector and norm of sum

Norm of block vector: for vectors a, b, c,

$$\left\| \begin{bmatrix} a \\ b \\ c \end{bmatrix} \right\| = \sqrt{\|a\|^2 + \|b\|^2 + \|c\|^2} = \|(\|a\|, \|b\|, \|c\|)\|$$

Norm of sum: for vectors a, b,

$$\|a + b\| = \sqrt{\|a\|^2 + 2a^T b + \|b\|^2}$$

Cauchy-Schwarz inequality

$$|a^T b| \le \|a\|\|b\| \quad \text{for all } a, b \in \mathbb{R}^n$$

moreover, equality $|a^T b| = \|a\|\|b\|$ holds if:

- $a = 0$ or $b = 0$; in this case $a^T b = 0 = \|a\|\|b\|$
- $b = \gamma a$ for some $\gamma > 0$; in this case

$$0 < a^T b = \gamma \|a\|^2 = \|a\|\|b\|$$

- $b = -\gamma a$ for some $\gamma > 0$; in this case

$$0 > a^T b = -\gamma \|a\|^2 = -\|a\|\|b\|$$

Proof of Cauchy-Schwarz inequality

1. trivial if $a = 0$ or $b = 0$

2. assume $\|a\| = \|b\| = 1$; we show that $-1 \leq a^T b \leq 1$

$$\begin{aligned} 0 &\leq \|a - b\|^2 \\ &= (a-b)^T(a-b) \\ &= \|a\|^2 - 2a^Tb + \|b\|^2 \\ &= 2(1 - a^Tb) \end{aligned}$$

with equality only if $a = b$

$$\begin{aligned} 0 &\leq \|a + b\|^2 \\ &= (a+b)^T(a+b) \\ &= \|a\|^2 + 2a^Tb + \|b\|^2 \\ &= 2(1 + a^Tb) \end{aligned}$$

with equality only if $a = -b$

3. for general nonzero a, b, apply case 2 to the unit-norm vectors

$$\frac{1}{\|a\|}a, \quad \frac{1}{\|b\|}b$$

Triangle inequality from Cauchy-Schwarz inequality

for vectors a, b of equal size

$$
\begin{aligned}
\|a+b\|^2 &= (a+b)^T(a+b) \\
&= a^Ta + b^Ta + a^Tb + b^Tb \\
&= \|a\|^2 + 2a^Tb + \|b\|^2 \\
&\leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \\
&= (\|a\| + \|b\|)^2
\end{aligned}
$$

- taking square roots gives the triangle inequality
- triangle inequality is an equality if and only if $a^Tb = \|a\|\|b\|$
- also note from line 3 that $\|a+b\|^2 = \|a\|^2 + \|b\|^2$ if $a^Tb = 0$

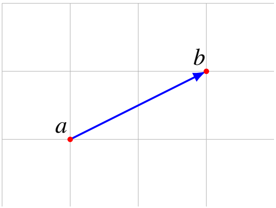
Euclidean distance

Euclidean distance between two vectors a and b,

$$\mathrm{dist}(a, b) = \|a - b\|$$

- agrees with ordinary distance for $n = 1, 2, 3$

2-D illustration



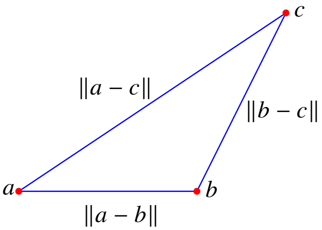
- when the distance between two vectors is small, we say they are 'close' or 'nearby', and when the distance is large, we say they are 'far'

Triangle inequality

- triangle with vertices at positions a, b, c
- edge lengths are $\|a-b\|, \|b-c\|, \|a-c\|$
- by triangle inequality

$$\|a-c\| = \|(a-b)+(b-c)\| \leq \|a-b\| + \|b-c\|$$

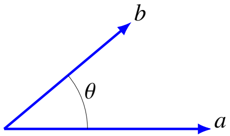
i.e., third edge length is no longer than sum of other two



Angle between vectors

the *angle* between nonzero real vectors a, b is defined as

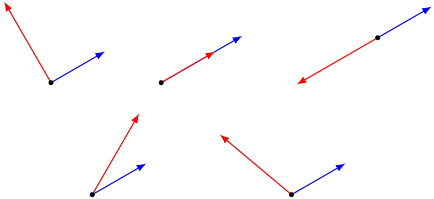
$$\theta = \angle(a, b) = \arccos\left(\frac{a^T b}{\|a\|\|b\|}\right)$$



- this is the unique value of $\theta \in [0, \pi]$ that satisfies $a^T b = \|a\|\|b\| \cos\theta$
- coincides with ordinary angle between vectors in 2-D and 3-D
- symmetric: $\angle(a, b) = \angle(b, a)$
- unaffected by scaling: $\angle(\alpha a, \beta b) = \angle(a, b)$ for positive α, β

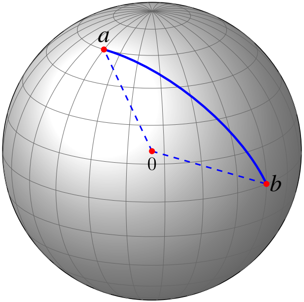
Classification of angles

$\theta = 0$	$a^T b = \|a\|\|b\|$	vectors are aligned or parallel
$0 \leq \theta < \pi/2$	$a^T b > 0$	vectors make an acute angle
$\theta = \pi/2$	$a^T b = 0$	vectors are orthogonal ($a \perp b$)
$\pi/2 < \theta \leq \pi$	$a^T b < 0$	vectors make an obtuse angle
$\theta = \pi$	$a^T b = -\|a\|\|b\|$	vectors are anti-aligned or opposed



Example: Spherical distance

if a, b are on sphere of radius R, distance along the sphere is $R\angle(a, b)$



Norm of sum via angles

for vectors a and b we have

$$\begin{aligned}\|a+b\|^2 &= \|a\|^2 + 2a^Tb + \|b\|^2 \\ &= \|a\|^2 + 2\|a\|\|b\|\cos\theta + \|b\|^2\end{aligned}$$

- if a and b are aligned $(\theta = 0)$, then $\|a+b\| = \|a\| + \|b\|$
- if a and b are orthogonal $(\theta = 90^\circ)$, then

$$\|a+b\|^2 = \|a\|^2 + \|b\|^2$$

and $\|a+b\| = \sqrt{\|a\|^2 + \|b\|^2}$ (called the Pythagorean theorem)

Outline

- norm, distance, angle
- **application examples**
- standard deviation, correlation
- complexity
- clustering

Feature distance and nearest neighbors

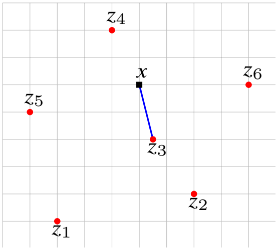
Feature distance

- let x and y be feature vectors for two entities
- $\|x - y\|$ is the *feature distance*; gives a measure of how different the objects are
  - example: features associated with patients in a hospital (weight, age, results of tests)
  - feature vector distance gives similarity between one patient case and another one

Nearest neighbor

- $z_1, \ldots, z_m$ is a list of vectors
- z_j is the nearest neighbor of x if

$$\|x - z_j\| \leq \|x - z_i\|, \quad i = 1, \ldots, m$$



Units for heterogeneous vector entries

$$\|a - b\|^2 = (a_1 - b_1)^2 + \cdots + (a_n - b_n)^2$$

- suppose entries of vectors a_i, b_i represent different types of quantities
- choice of units for each entry affects the distance/angle between a and b
- general rule: choose units so typical vector entries have similar ranges of values

Document dissimilarity

- if x_i represent histogram of word occurrence in document i
- $\|x_i - x_j\|$ measures the dissimilarity between documents

Example

- 5 Wikipedia articles: 'Veterans Day', 'Memorial Day', 'Academy Awards', 'Golden Globe Awards', 'Super Bowl'
- word count histograms, dictionary of 4423 words
- pairwise distances shown below

	Veterans Day	Memorial Day	Academy Awards	Golden Globe Awards	Super Bowl
Veterans Day	0	0.095	0.130	0.153	0.170
Memorial Day	0.095	0	0.122	0.147	0.164
Academy A.	0.130	0.122	0	0.108	0.164
Golden Globe A.	0.153	0.147	0.108	0	0.181
Super Bowl	0.170	0.164	0.164	0.181	0

Document dissimilarity by angles

- if n-vectors x_i are word counts for documents, their angle $\angle(x_i, x_j)$ can be used as a measure of document dissimilarity
- example: pairwise angles (in degrees) for 5 Wikipedia pages shown below

	Veterans Day	Memorial Day	Academy Awards	Golden globe Awards	Super Bowl
Veterans Day	0	60.6	85.7	87.0	87.7
Memorial Day	60.6	0	85.6	87.5	87.5
Academy A.	85.7	85.6	0	58.7	86.1
Golden Globe A.	87.0	87.5	58.7	0	86.0
Super Bowl	87.7	87.5	86.1	86.0	0

Outline

- norm, distance, angle
- application examples
- **standard deviation, correlation**
- complexity
- clustering

RMS value

the *root-mean-square* value of $a \in \mathbb{R}^n$ is the root of the average squared entry

$$\mathrm{rms}(x) = \sqrt{\frac{a_1^2 + \cdots + a_n^2}{n}} = \frac{\|a\|}{\sqrt{n}}$$

- it is root of *mean-square value*: $\mathrm{ms} = (a_1^2 + \cdots + a_n^2)/n$
- RMS value useful for comparing sizes of vectors of different lengths
- $\mathrm{rms}(a)$ gives 'typical' value of $|a_i|$
- *e.g.*, $\mathrm{rms}(\alpha \mathbf{1}) = |\alpha|$ (independent of n)
- $\mathrm{rms}(a - b)$ is called the RMS *deviation* between a and b

Standard deviation

the *standard deviation* of $a \in \mathbb{R}^n$ is

$$\mathrm{std}(a) = \mathrm{rms}(a - \mathrm{avg}(a)\mathbf{1}) = \left\| a - ((\mathbf{1}^T a)/n)\mathbf{1} \right\| / \sqrt{n}$$

- std is RMS deviation from the average
- std “tells” us the typical amount a vector entries deviate from their average
- $\tilde{a} = a - \mathrm{avg}(a)\mathbf{1}$ is called *de-meaned* vector (since $\mathrm{avg}(\tilde{a}) = 0$)
- other notation: μ and σ are often used for mean and standard deviation

Standard deviation formula

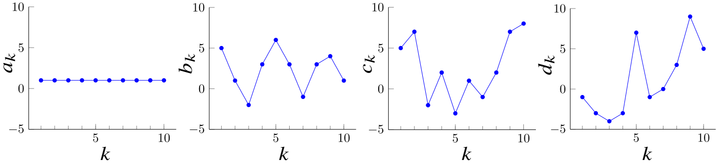
$$\mathrm{rms}(a)^2 = \mathrm{avg}(a)^2 + \mathrm{std}(a)^2$$

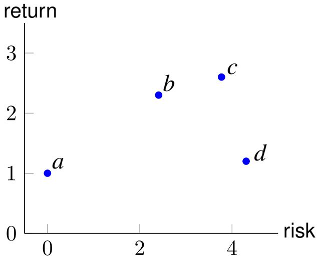
Proof

$$\begin{aligned}
\mathrm{std}(a)^2 &= \frac{\|a - \mathrm{avg}(a)\mathbf{1}\|^2}{n} \\
&= \frac{1}{n}\left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right)^T \left(a - \frac{\mathbf{1}^T a}{n}\mathbf{1}\right) \\
&= \frac{1}{n}\left(a^T a - \frac{\left(\mathbf{1}^T a\right)^2}{n} - \frac{\left(\mathbf{1}^T a\right)^2}{n} + \left(\frac{\mathbf{1}^T a}{n}\right)^2 n\right) \\
&= \frac{1}{n}\left(a^T a - \frac{\left(\mathbf{1}^T a\right)^2}{n}\right) \\
&= \mathrm{rms}(a)^2 - \mathrm{avg}(a)^2
\end{aligned}$$

Mean return and risk of investment

- vectors represent time series of returns on an investment (as a percentage)
- average value is (mean) return of the investment
- standard deviation measures variation around the mean, *i.e.*, risk





Chebyshev inequality

- assume that k of the numbers $|x_1|, \ldots, |x_n|$ are $\geq \alpha$
- then k of the numbers $x_1^2, \ldots, x_n^2$ are $\geq \alpha^2$
- so $\|x\|^2 = x_1^2 + \cdots + x_n^2 \geq k\alpha^2$

Chebyshev inequality

$$k \leq \|x\|^2/\alpha^2$$

- number of x_i with $|x_i| \geq \alpha$ is no more than $\|x\|^2/\alpha^2$
- in terms of RMS value:

$$(\text{fraction of entries with } |x_i| \geq \alpha) = \frac{k}{n} \leq \left(\frac{\mathrm{rms}(x)}{\alpha}\right)^2$$

- for $\alpha = 5\mathrm{rms}(x)$, no more than $\frac{1}{25} = 4\%$ of entries of x satisfy $|x_i| \geq 5\mathrm{rms}(x)$
- RMS value indicate that not too many of the entries of a vector can be much bigger (in absolute value) than its RMS value

Chebyshev inequality for standard deviation

if k is the number of entries of x that satisfy $|x_i - \mathrm{avg}(x)| \geq \alpha$, then

$$\frac{k}{n} \leq \left(\frac{\mathrm{std}(x)}{\alpha}\right)^2$$

- rough idea: most entries of x are not too far from the mean
- example: for return time series with mean 8% and standard deviation (risk) 3%, loss $(x_i \leq 0)$ can occur in no more than $(3/8)^2 = 14.1$% of periods
- by Chebyshev inequality, fraction of entries of x with

$$|x_i - \mathrm{avg}(x)| \geq \beta \mathrm{std}(x)$$

  is no more than $1/\beta^2$ (for $\beta > 1$)

  – fraction of entries of x within β standard deviations of $\mathrm{avg}(x)$ is at least $1 - 1/\beta^2$

Correlation coefficient

correlation coefficient (between a and b)

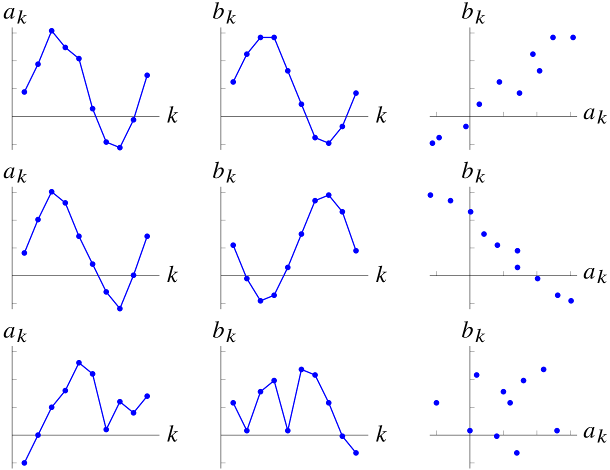
$$\rho = \frac{\tilde{a}^T \tilde{b}}{\|\tilde{a}\| \|\tilde{b}\|}$$

where vectors $\tilde{a}$ and $\tilde{b}$ are de-meaned vectors ($\tilde{a} \neq 0, \tilde{b} \neq 0$):

$$\tilde{a} = a - \mathrm{avg}(a)\mathbf{1}, \quad \tilde{b} = b - \mathrm{avg}(b)\mathbf{1}$$

- $\rho = \cos \angle(\tilde{a}, \tilde{b})$ hence $-1 \leq \rho \leq 1$
- $\rho = 0$, a and b are uncorrelated
- $\rho > 0.8$ (or so), a and b are highly correlated
- $\rho < -0.8$ (or so), a and b are highly anti-correlated
- highly correlated "means" many a_i, b_i are both above (below) their means

Example



$\rho_{ab} = 0.968$

$\rho_{ab} = -0.988$

$\rho_{ab} = 0.004$

Examples

highly correlated vectors:

- rainfall time series at nearby locations
- daily returns of similar companies in same industry
- word count vectors of closely related documents (*e.g.*, same author, topic, ...)
- sales of shoes and socks (at different locations or periods)

approximately uncorrelated vectors

- unrelated vectors
- audio signals (even different tracks in multi-track recording)

(somewhat) negatively correlated vectors

- daily temperatures in Palo Alto and Melbourne

Properties and standardization

Properties of standard deviation

- *adding a constant:* $\mathrm{std}(a + \beta\mathbf{1}) = \mathrm{std}(a)$ for vector a and number β
- *multiplying by a scalar:* $\mathrm{std}(\beta a) = |\beta|\,\mathrm{std}(a)$ for vector a and number β
- *sum:* $\mathrm{std}(a + b) = \sqrt{\mathrm{std}(a)^2 + 2\rho\,\mathrm{std}(a)\,\mathrm{std}(b) + \mathrm{std}(b)^2}$ for vectors a, b

Standardization

- de-meaned vector of a in standard units is

$$z = \frac{1}{\mathrm{std}(a)}(a - \mathrm{avg}(a)\mathbf{1})$$

- z is called *standardized* or z*-scored* version of a ($\mathrm{avg}(z) = 0$ and $\mathrm{std}(z) = 1$)
- $z_4 = 1.4$ means a_4 is 1.4 standard deviations above the mean of entries of a

Example: Hedging investments

- a and b are time series of returns for two assets with the same return (average) μ, risk (standard deviation) σ, and correlation coefficient ρ

- $c = (a + b)/2$ is time series of returns for an investment with 50% in each asset

- this blended investment has the same return as the original assets, since

$$\mathrm{avg}(c) = \mathrm{avg}((a + b)/2) = (\mathrm{avg}(a) + \mathrm{avg}(b))/2 = \mu$$

- the risk (standard deviation) of this blended investment is

$$\mathrm{std}(c) = \sqrt{2\sigma^2 + 2\rho\sigma^2}/2 = \sigma\sqrt{(1 + \rho)/2}$$

- risk of the blended investment is never more than the risk of the original assets, and is smaller when the correlation of the original asset returns is smaller

- investing in two uncorrelated or negativ. correlated assets is called *hedging*

Outline

- norm, distance, angle
- application examples
- standard deviation, correlation
- **complexity**
- clustering

Complexity of norms

for n-vectors

- $\|x\|$ requires $2n$ flops
  - n multiplications (to square each entry)
  - $n-1$ additions (to add the squares)
  - one squareroot
- RMS value costs $2n$ (ignore two flops from division of $\sqrt{n}$)
- distance between two vectors costs $3n$ flops
- angle between them costs $6n$ flops
- de-meaning an n-vector requires $2n$ flops
  - n for forming the average
  - n flops for subtracting the average from each entry
- standard deviation costs $4n$ flops
  - $2n$ for computing the de-meaned vector
  - $2n$ for computing its RMS value
- correlation coefficient costs $10n$ flops to compute

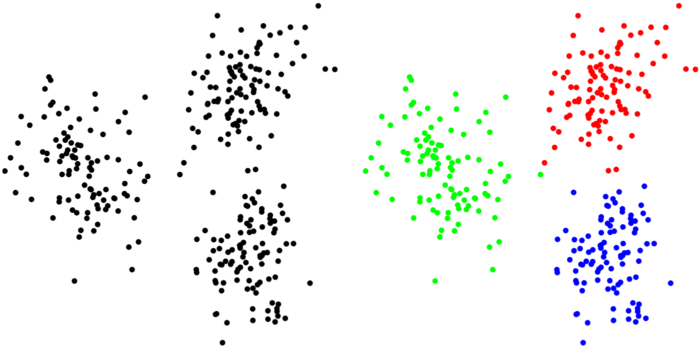
Outline

- norm, distance, angle
- application examples
- standard deviation, correlation
- complexity
- **clustering**

Clustering

- given N n-vectors $x_1, \ldots, x_N$ (features)
- goal: partition (divide, group, cluster) vectors into k groups ($k << N$)
- we want vectors in the same group to be close to each other

Example ($N = 300$, $k = 3$)



Examples

- topic discovery
  - x_i is word count histogram for document i
  - clustering algorithm groups documents with similar topics, genre, or author
- patient clustering
  - x_i are patient features (test results, symptoms, ..etc)
  - clustering algorithm groups similar patients together
- customer market segmentation
  - x_i is purchase quantities of items purchased by customer i
  - clustering algorithm groups customers with similar purchasing patterns
- financial sectors
  - x_i is financial attributes of company i (total capitalization, quarterly return, profits,...)
  - clustering algorithm groups companies into *sectors* (companies with similar attributes)
- color images
  - x_i are RGB pixel values
  - clustering algorithm groups images with similar colors

Clustering objective

Specifying clusters assignment

- c_i is group number that x_i is assigned to ($i = 1, \ldots, N$)
- G_j is set of (indices) corresponding to group $j = 1, \ldots, k$
  - example: $N = 5$ vectors and $k = 3$ groups
  - $c = (3, 1, 1, 1, 2)$ means x_1 is assigned to group 3, x_2 is assigned to group 1,...
  - $G_1 = \{2, 3, 4\}$, $G_2 = \{5\}$, $G_3 = \{1\}$

Group representatives

- n-vectors $z_1, \ldots, z_k$ are *group representatives*
- we want $\|x_i - z_{c_i}\|$ to be small (z_{c_i} is representative vector associated x_i)

Objective: mean square distance from vectors to associated representative

$$J^{\text{clust}} = \left(\|x_1 - z_{c_1}\|^2 + \cdots + \|x_N - z_{c_N}\|^2\right)/N$$

- J^{clust} small means good clustering
- goal: choose clustering c_i and representatives z_j to minimize J^{clust}

Partitioning the vectors given the representatives

- assume group representatives $z_1, \ldots, z_k$ are given (fixed)
- how to choose $c_1, \ldots, c_N$ to minimize J^{clust}? (how to assign vectors to groups)

Partitioning the vectors given z_i

- c_i only appears in term $\|x_i - z_{c_i}\|^2$ in J^{clust}
- to minimize over c_i, choose c_i to be the value of j that minimizes $\|x_i - z_j\|^2$

$$c_i = \underset{j=1,\ldots,k}{\operatorname{argmin}} \|x_i - z_j\|^2$$

  i.e., assign each vector to its nearest neighbor representative

- so the value of J^{clust} is

$$J^{\text{clust}} = \left(\min_{j=1,\ldots,k} \|x_1 - z_j\|^2 + \cdots + \min_{j=1,\ldots,k} \|x_N - z_j\|^2 \right) / N$$

  this is mean squared distance from data vectors to their closest representative

Choosing representatives given the partition

given $G_1, \ldots, G_k$, how do we choose $z_1, \ldots, z_k$ to minimize J^{clust}?

Choosing z_j given G_i

- J^{clust} splits into a sum of k sums, one for each z_j:

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

- 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:

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

(we will see later how to get this solution)

k-means algorithm

given initial representatives $z_1, \ldots, z_k$ for the k groups and repeat:

1. assign x_i to the nearest group representative z_j
2. set the representative z_j to be the mean of the vectors in group j

Math description

---

given $x_1, \ldots, x_N \in \mathbb{R}^n$ and $z_1, \ldots, z_k \in \mathbb{R}^n$
repeat

1. *partition vectors:* assign i to G_j, $j = \operatorname{argmin}_{j'} \|x_i - z_{j'}\|^2$
2. *update representatives:* $z_j = \frac{1}{|G_j|} \sum_{i \in G_j} x_i$

until $z_1, \ldots, z_k$ stop changing

(in practice, often restarted a few times, with different starting points)

Complexity

k-means cost $(3k+1)Nn$ flops per iteration (order Nkn flops)

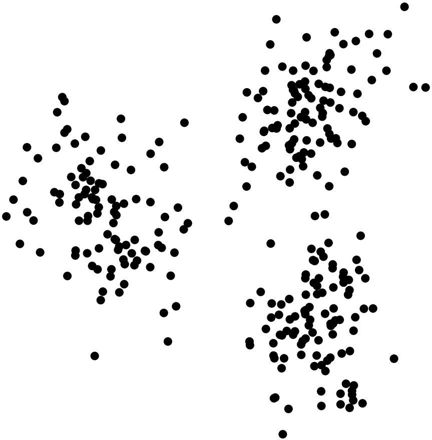
step 1:

- each distance $\|x_i - z_j\|$ costs $3n$ flops
- computing all distances $\|x_i - z_j\|$ over groups costs $3kn$
- comparisons to find the minimum costs $k-1$ flops
- repeat above N times to get approximately $3Nkn$ flops

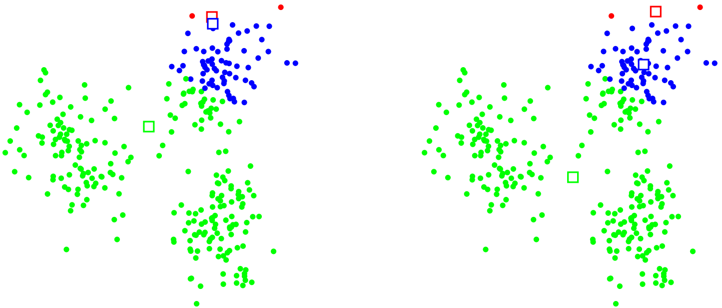
step 2: approximately Nn flops

- averaging $(1/p)\sum_{i=1}^{p} x_i$ clusters requires a total of np flops
- averaging all clusters requires a total of Nn flop

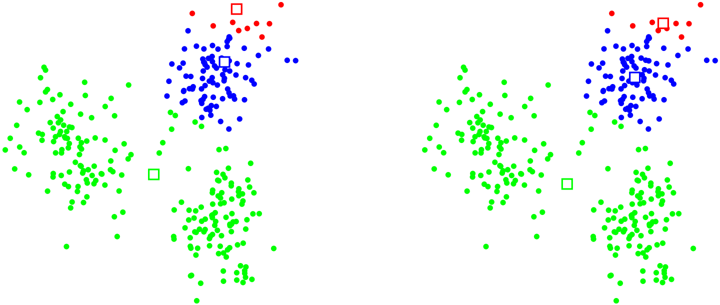
Data



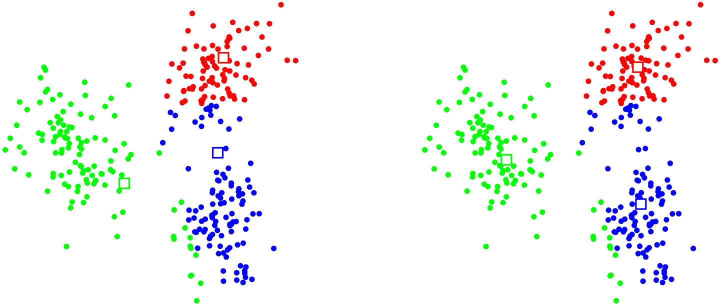
Iteration 1

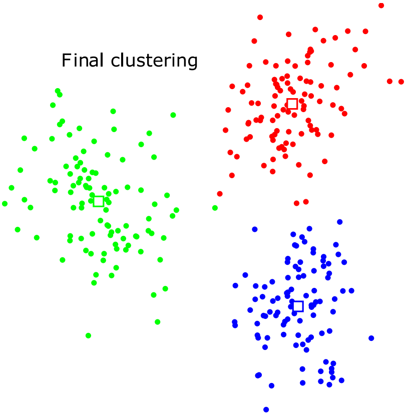


Iteration 2



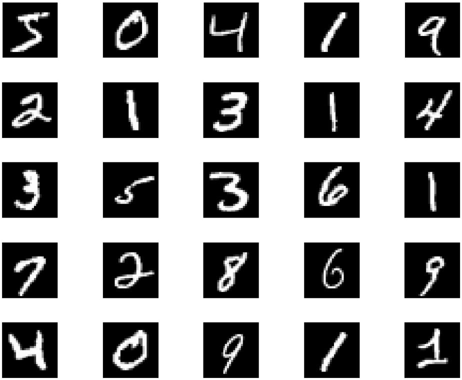
Iteration 10




Final clustering

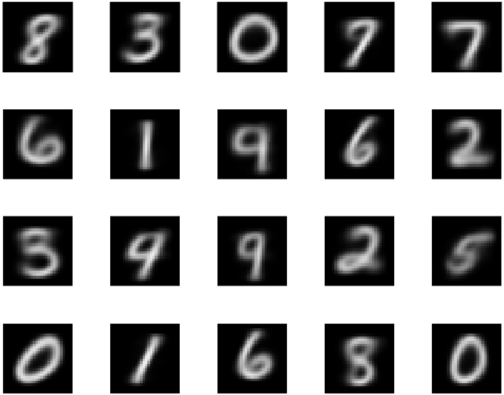
Handwritten digit image set

- MNIST images of handwritten digits
- $N = 60,000$ size 28×28 images, represented as 784-vectors x_i
- 25 image samples shown below



Group representatives, best clustering

$k = 20$ group representatives, z_j



Document topic discovery

- $N = 500$ Wikipedia articles
- dictionary of $n = 4423$ words
- each document is represented by a word histogram vector of length $n = 4423$
- $k = 9$, run 20 times with different initial assignments

Topics discovered (clusters 1-3)

words with largest representative coefficients of the word histogram

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

titles of articles closest to cluster representative of the word histogram

1. ”Floyd Mayweather, Jr”, ”Kimbo Slice”, ”Ronda Rousey”, ”José Aldo”, “Joe Frazier”, ”Wladimir Klitschko”, ”Saul Álvarez”, ”Gennady Golovkin”, ”Nate Diaz”, ...
2. ”Halloween”, ”Guy Fawkes Night” ”Diwali”, ”Hanukkah”, ”Groundhog Day”, ”Rosh Hashanah”, ”Yom Kippur”, ”Seventh-day Adventist Church”, ”Remembrance Day”, ...
3. ”Mahatma Gandhi”, ”Sigmund Freud”, ”Carly Fiorina”, ”Frederick Douglass”, ”Marco Rubio”, ”Christopher Columbus”, “Fidel Castro”, “Jim Webb”, ...

Topics discovered (clusters 4-6)

words with largest representative coefficients of the word histogram

Cluster 4		Cluster 5		Cluster 6	
Word	Coef.	Word	Coef.	Word	Coef.
album	0.031	game	0.023	series	0.029
release	0.016	season	0.020	season	0.027
song	0.015	team	0.018	episode	0.013
music	0.014	win	0.017	character	0.011
single	0.011	player	0.014	film	0.008

titles of articles closest to cluster representative

4. ”David Bowie”, ”Kanye West” ”Celine Dion”, ”Kesha”, ”Ariana Grande”, ”Adele”, ”Gwen Stefani”, ”Anti (album)”, ”Dolly Parton”, ”Sia Furler”, ...
5. ”Kobe Bryant”, ”Lamar Odom”, ”Johan Cruyff”, ”Yogi Berra”, ”José Mourinho”, ”Halo 5: Guardians”, ”Tom Brady”, ”Eli Manning”, ”Stephen Curry”, ”Carolina Panthers”, ...
6. ”The X-Files”, ”Game of Thrones”, ”House of Cards (U.S. TV series)”, ”Daredevil (TV series)”, ”Supergirl (U.S. TV series)”, ”American Horror Story”, ...

Topics discovered (clusters 7-9)

words with largest representative coefficients

Cluster 7		Cluster 8		Cluster 9	
Word	Coef.	Word	Coef.	Word	Coef.
match	0.065	film	0.036	film	0.061
win	0.018	star	0.014	million	0.019
championship	0.016	role	0.014	release	0.013
team	0.015	play	0.010	star	0.010
event	0.015	series	0.009	character	0.006

titles of articles closest to cluster representative

7. ”Wrestlemania 32”, ”Payback (2016)”, ”Survivor Series (2015)”, ”Royal Rumble (2016)”, ”Night of Champions (2015)”, ”Fastlane (2016)”, ”Extreme Rules (2016)”, ...
8. ”Ben Affleck”, ”Johnny Depp”, ”Maureen O’Hara”, ”Kate Beckinsale”, ”Leonardo DiCaprio”, ”Keanu Reeves”, ”Charlie Sheen”, ”Kate Winslet”, ”Carrie Fisher”, ...
9. ”Star Wars: The Force Awakens”, ”Star Wars Episode I: The Phantom Menace”, ”The Martian (film)”, ”The Revenant (2015 film)”, ”The Hateful Eight”, ...

Applications

Classification: determine vector belongs to which group

- cluster a large collection of vectors into k groups
- label the groups by hand
- assign *new* vectors to one of the k groups by choosing the nearest group representative

Recommendation engine: suggest items that user might be interested in

- example: vectors give the number of times a user has listened to or streamed each song from a library of n songs over some period
- clustering the vectors reveals groups of users with similar musical taste
- allows us to suggest new songs from those with similar tastes

References and further readings

- S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares,* Cambridge University Press, 2018.
- L. Vandenberghe. *EE133A lecture notes,* Univ. of California, Los Angeles. (`http://www.seas.ucla.edu/~vandenbe/ee133a.html`)