2. Norm, distance, angle

- norm
- distance
- $k$-means algorithm
- angle
- complex vectors
Euclidean norm

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

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

- if \( n = 1 \), \( \|a\| \) reduces to absolute value \( |a| \)
- measures the magnitude of \( a \)
- sometimes written as \( \|a\|_2 \) to distinguish from other norms, e.g.,

\[
\|a\|_1 = |a_1| + |a_2| + \cdots + |a_n|
\]
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 (proved on page 2.7)

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

Norm of block vector: if \( a, b \) are vectors,

\[ \left\| \begin{bmatrix} a \\ b \end{bmatrix} \right\| = \sqrt{\|a\|^2 + \|b\|^2} \]
Cauchy–Schwarz inequality

\[ |a^T b| \leq \|a\| \|b\| \quad \text{for all } a, b \in \mathbb{R}^n \]

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

- \( a = 0 \) or \( b = 0 \); in this case \( a^T b = 0 = \|a\| \|b\| \)

- \( a \neq 0 \) and \( b \neq 0 \), and \( b = \gamma a \) for some \( \gamma > 0 \); in this case

\[ 0 < a^T b = \gamma \|a\|^2 = \|a\| \|b\| \]

- \( a \neq 0 \) and \( b \neq 0 \), and \( b = -\gamma a \) 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 \(-1 \leq a^T b \leq 1\)

\[
0 \leq \|a - b\|^2 = (a - b)^T(a - b) = \|a\|^2 - 2a^Tb + \|b\|^2 = 2(1 - a^Tb)
\]

with equality only if \( a = b \)

\[
0 \leq \|a + b\|^2 = (a + b)^T(a + b) = \|a\|^2 + 2a^Tb + \|b\|^2 = 2(1 + a^Tb)
\]

with equality only if \( a = -b \)

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

\[
\frac{1}{\|a\|}a, \quad \frac{1}{\|b\|}b
\]
Average and RMS value

let \( a \) be a real \( n \)-vector

- the average of the elements of \( a \) is

\[
\text{avg}(a) = \frac{a_1 + a_2 + \cdots + a_n}{n} = \frac{1^T a}{n}
\]

- the root-mean-square value is the root of the average squared entry

\[
\text{rms}(a) = \sqrt{\frac{a_1^2 + a_2^2 + \cdots + a_n^2}{n}} = \frac{\|a\|}{\sqrt{n}}
\]

Exercises

- show that \(|\text{avg}(a)| \leq \text{rms}(a)|

- show that average of \( b = (|a_1|, |a_2|, \ldots, |a_n|) \) satisfies \( \text{avg}(b) \leq \text{rms}(a) \)
Triangle inequality from Cauchy–Schwarz inequality

for vectors $a$, $b$ of equal size

\[
\|a + b\|^2 = (a + b)^T(a + b)
\]
\[
= a^T a + b^T a + a^T b + b^T b
\]
\[
= \|a\|^2 + 2a^T b + \|b\|^2
\]
\[
\leq \|a\|^2 + 2\|a\|\|b\| + \|b\|^2 \quad \text{(by Cauchy–Schwarz)}
\]
\[
= (\|a\| + \|b\|)^2
\]

- taking squareroots gives the triangle inequality

- triangle inequality is an equality if and only if $a^T b = \|a\|\|b\|$ (see page 2.4)

- also note from line 3 that $\|a + b\|^2 = \|a\|^2 + \|b\|^2$ if $a^T b = 0$
Outline

• norm

• distance

• $k$-means algorithm

• angle

• complex vectors
Distance

the (Euclidean) distance between vectors $a$ and $b$ is defined as $\|a - b\|$

- $\|a - b\| \geq 0$ for all $a$, $b$ and $\|a - b\| = 0$ only if $a = b$

- triangle inequality

$$\|a - c\| \leq \|a - b\| + \|b - c\|$$  for all $a$, $b$, $c$

- RMS deviation between $n$-vectors $a$ and $b$ is $\text{rms}(a - b) = \frac{\|a - b\|}{\sqrt{n}}$
Standard deviation

let $a$ be a real $n$-vector

- the *de-meaned* vector is the vector of deviations from the average

$$
\begin{align*}
    a - \operatorname{avg}(a)1 &= \begin{bmatrix} a_1 - \operatorname{avg}(a) \\ a_2 - \operatorname{avg}(a) \\ \vdots \\ a_n - \operatorname{avg}(a) \end{bmatrix} = \begin{bmatrix} a_1 - (1^Ta)/n \\ a_2 - (1^Ta)/n \\ \vdots \\ a_n - (1^Ta)/n \end{bmatrix} \\
    
\end{align*}
$$

- the *standard deviation* is the RMS deviation from the average

$$
\operatorname{std}(a) = \operatorname{rms}(a - \operatorname{avg}(a)1) = \frac{\|a - ((1^Ta)/n)1\|}{\sqrt{n}}
$$

- the de-meaned vector in *standard units* is

$$
\frac{1}{\operatorname{std}(a)}(a - \operatorname{avg}(a)1)
$$
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
Exercise

show that

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

Solution

$$\text{std}(a)^2 = \frac{\|a - \text{avg}(a)1\|^2}{n}$$

$$= \frac{1}{n} \left( a - \frac{1^T a}{n} 1 \right)^T \left( a - \frac{1^T a}{n} 1 \right)$$

$$= \frac{1}{n} \left( a^T a - \frac{(1^T a)^2}{n} - \frac{(1^T a)^2}{n} + \left( \frac{1^T a}{n} \right)^2 n \right)$$

$$= \frac{1}{n} \left( a^T a - \frac{(1^T a)^2}{n} \right)$$

$$= \text{rms}(a)^2 - \text{avg}(a)^2$$
Exercise: nearest scalar multiple

given two vectors \(a, b \in \mathbb{R}^n\), with \(a \neq 0\), find scalar multiple \(ta\) closest to \(b\)

\[
\begin{align*}
\text{line } \{ta \mid t \in \mathbb{R}\} \\
\hat{t}a \\
b
\end{align*}
\]

Solution

- squared distance between \(ta\) and \(b\) is

\[
\|ta - b\|^2 = (ta - b)^T (ta - b) = t^2 a^T a - 2ta^T b + b^T b
\]

a quadratic function of \(t\) with positive leading coefficient \(a^T a\)

- derivative with respect to \(t\) is zero for

\[
\hat{t} = \frac{a^T b}{a^T a} = \frac{a^T b}{\|a\|^2}
\]
Exercise: average of collection of vectors

given \( N \) vectors \( x_1, \ldots, x_N \in \mathbf{R}^n \), find the \( n \)-vector \( z \) that minimizes

\[
\| z - x_1 \|^2 + \| z - x_2 \|^2 + \cdots + \| z - x_N \|^2
\]

\( z \) is also known as the *centroid* of the points \( x_1, \ldots, x_N \)
**Solution:** sum of squared distances is

\[ \| z - x_1 \|^2 + \| z - x_2 \|^2 + \cdots + \| z - x_N \|^2 \]

\[ = \sum_{i=1}^{n} \left( (z_i - (x_1)_i)^2 + (z_i - (x_2)_i)^2 + \cdots + (z_i - (x_N)_i)^2 \right) \]

\[ = \sum_{i=1}^{n} \left( N z_i^2 - 2 z_i ((x_1)_i + (x_2)_i + \cdots + (x_N)_i) + (x_1)^2_i + \cdots + (x_N)^2_i \right) \]

here \((x_j)_i\) is \(i\)th element of the vector \(x_j\)

- term \(i\) in the sum is minimized by

\[ z_i = \frac{1}{N} ((x_1)_i + (x_2)_i + \cdots + (x_N)_i) \]

- solution \(z\) is component-wise average of the points \(x_1, \ldots, x_N\):

\[ z = \frac{1}{N} (x_1 + x_2 + \cdots + x_N) \]
Outline

• norm

• distance

• $k$-means algorithm

• angle

• complex vectors
$k$-means clustering

a popular iterative algorithm for partitioning $N$ vectors $x_1, \ldots, x_N$ in $k$ clusters
Algorithm

choose initial ‘representatives’ $z_1, \ldots, z_k$ for the $k$ groups and repeat:

1. assign each vector $x_i$ to the nearest group representative $z_j$
2. set the representative $z_j$ to the mean of the vectors assigned to it

• as a variation, choose a random initial partition and start with step 2
• initial representatives are often chosen randomly
• solution depends on choice of initial representatives or partition
• can be shown to converge in a finite number of iterations
• in practice, often restarted a few times, with different starting points

Norm, distance, angle
Example: first iteration

assignment to groups

updated representatives

Norm, distance, angle 2.17
Iteration 2

assignment to groups

updated representatives
Iteration 3

assignment to groups

updated representatives

Norm, distance, angle 2.19
Iteration 11

assignment to groups  
updated representatives

Norm, distance, angle 2.20
Iteration 12

assignment to groups

updated representatives

Norm, distance, angle

2.21
Iteration 13

assignment to groups
updated representatives

Norm, distance, angle
Iteration 14

assignment to groups

updated representatives

Norm, distance, angle 2.23
Final clustering

Norm, distance, angle 2.24
Image clustering

- MNIST dataset of handwritten digits
- \( N = 60000 \) grayscale images of size \( 28 \times 28 \) (vectors \( x_i \) of size \( 28^2 = 784 \))
- 25 examples:
Group representatives \((k = 20)\)

- \(k\)-means algorithm, with \(k = 20\) and randomly chosen initial partition
- 20 group representatives

![Images of group representatives]
Group representatives \((k = 20)\)

result for another initial partition
Document topic discovery

- \( N = 500 \) Wikipedia articles, from weekly most popular lists (9/2015–6/2016)
- dictionary of 4423 words
- each article represented by a word histogram vector of size 4423
- result of \( k \)-means algorithm with \( k = 9 \) and randomly chosen initial partition

Cluster 1

- largest coefficients in cluster representative \( z_1 \)

<table>
<thead>
<tr>
<th>word</th>
<th>fight</th>
<th>win</th>
<th>event</th>
<th>champion</th>
<th>fighter</th>
<th>…</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.038</td>
<td>0.022</td>
<td>0.019</td>
<td>0.015</td>
<td>0.015</td>
<td>…</td>
</tr>
</tbody>
</table>

- documents in cluster 1 closest to representative

  “Floyd Mayweather, Jr”, “Kimbo Slice”, “Ronda Rousey”, “José Aldo”, “Joe Frazier”, …
Cluster 2

- largest coefficients in cluster representative $z_2$

<table>
<thead>
<tr>
<th>word</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>holiday</td>
<td>0.012</td>
</tr>
<tr>
<td>celebrate</td>
<td>0.009</td>
</tr>
<tr>
<td>festival</td>
<td>0.007</td>
</tr>
<tr>
<td>celebration</td>
<td>0.006</td>
</tr>
<tr>
<td>calendar</td>
<td>0.006</td>
</tr>
</tbody>
</table>

- documents in cluster 2 closest to representative


Cluster 3

- largest coefficients in cluster representative $z_3$

<table>
<thead>
<tr>
<th>word</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>united</td>
<td>0.004</td>
</tr>
<tr>
<td>family</td>
<td>0.003</td>
</tr>
<tr>
<td>party</td>
<td>0.003</td>
</tr>
<tr>
<td>president</td>
<td>0.003</td>
</tr>
<tr>
<td>government</td>
<td>0.003</td>
</tr>
</tbody>
</table>

- documents in cluster 3 closest to representative

  “Mahatma Gandhi”, “Sigmund Freund”, “Carly Fiorina”, “Frederick Douglass”, “Marco Rubio”, …
Cluster 4

- largest coefficients in cluster representative $z_4$

<table>
<thead>
<tr>
<th>word</th>
<th>album</th>
<th>release</th>
<th>song</th>
<th>music</th>
<th>single</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.031</td>
<td>0.016</td>
<td>0.015</td>
<td>0.014</td>
<td>0.011</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 4 closest to representative


Cluster 5

- largest coefficients in cluster representative $z_5$

<table>
<thead>
<tr>
<th>word</th>
<th>game</th>
<th>season</th>
<th>team</th>
<th>win</th>
<th>player</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.023</td>
<td>0.020</td>
<td>0.018</td>
<td>0.017</td>
<td>0.014</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 5 closest to representative

  “Kobe Bryant”, “Lamar Odom”, “Johan Cruyff”, “Yogi Berra”, “José Mourinho”, …
Cluster 6

- largest coefficients in representative \( z_6 \)

<table>
<thead>
<tr>
<th>word</th>
<th>series</th>
<th>season</th>
<th>episode</th>
<th>character</th>
<th>film</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.029</td>
<td>0.027</td>
<td>0.013</td>
<td>0.011</td>
<td>0.008</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 6 closest to cluster representative

  “The X-Files”, “Game of Thrones”, “House of Cards”, “Daredevil”, “Supergirl”, …

Cluster 7

- largest coefficients in representative \( z_7 \)

<table>
<thead>
<tr>
<th>word</th>
<th>match</th>
<th>win</th>
<th>championship</th>
<th>team</th>
<th>event</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.065</td>
<td>0.018</td>
<td>0.016</td>
<td>0.015</td>
<td>0.015</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 7 closest to cluster representative

  “Night of Champions (2015)”, …
Cluster 8

- largest coefficients in representative $z_8$

<table>
<thead>
<tr>
<th>word</th>
<th>film</th>
<th>star</th>
<th>role</th>
<th>play</th>
<th>series</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.036</td>
<td>0.014</td>
<td>0.014</td>
<td>0.010</td>
<td>0.009</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 8 closest to cluster representative


Cluster 9

- largest coefficients in representative $z_9$

<table>
<thead>
<tr>
<th>word</th>
<th>film</th>
<th>million</th>
<th>release</th>
<th>star</th>
<th>character</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>coefficient</td>
<td>0.061</td>
<td>0.019</td>
<td>0.013</td>
<td>0.010</td>
<td>0.006</td>
<td>...</td>
</tr>
</tbody>
</table>

- documents in cluster 9 closest to cluster representative

Outline

• norm

• distance

• $k$-means algorithm

• angle

• complex vectors
Angle between vectors

the angle between nonzero real vectors \( a, b \) is defined as

\[
\arccos \left( \frac{a^T b}{\|a\| \|b\|} \right)
\]

- this is the unique value of \( \theta \in [0, \pi] \) that satisfies \( a^T b = \|a\| \|b\| \cos \theta \)

- Cauchy–Schwarz inequality guarantees that

\[
-1 \leq \frac{a^T b}{\|a\| \|b\|} \leq 1
\]
Terminology

\[ \theta = 0 \quad \Rightarrow \quad a^T b = \|a\|\|b\| \quad \text{vectors are aligned or parallel} \]

\[ 0 \leq \theta < \pi/2 \quad \Rightarrow \quad a^T b > 0 \quad \text{vectors make an acute angle} \]

\[ \theta = \pi/2 \quad \Rightarrow \quad a^T b = 0 \quad \text{vectors are orthogonal} \quad (a \perp b) \]

\[ \pi/2 < \theta \leq \pi \quad \Rightarrow \quad a^T b < 0 \quad \text{vectors make an obtuse angle} \]

\[ \theta = \pi \quad \Rightarrow \quad a^T b = -\|a\|\|b\| \quad \text{vectors are anti-aligned or opposed} \]
**Correlation coefficient**

The *correlation coefficient* between non-constant vectors \(a, b\) is

\[
\rho_{ab} = \frac{\tilde{a}^T \tilde{b}}{||\tilde{a}|| \ ||\tilde{b}||}
\]

where \(\tilde{a} = a - \text{avg}(a)\mathbf{1}\) and \(\tilde{b} = b - \text{avg}(b)\mathbf{1}\) are the de-meaned vectors

- only defined when \(a\) and \(b\) are not constant (\(\tilde{a} \neq 0\) and \(\tilde{b} \neq 0\))
- \(\rho_{ab}\) is the cosine of the angle between the de-meaned vectors
- a number between \(-1\) and \(1\)
- \(\rho_{ab}\) is the average product of the deviations from the mean in standard units

\[
\rho_{ab} = \frac{1}{n} \sum_{i=1}^{n} \frac{(a_i - \text{avg}(a))}{\text{std}(a)} \cdot \frac{(b_i - \text{avg}(b))}{\text{std}(b)}
\]
Examples

\[ \rho_{ab} = 0.968 \]

\[ \rho_{ab} = -0.988 \]

\[ \rho_{ab} = 0.004 \]
Regression line

- scatter plot shows two $n$-vectors $a$, $b$ as $n$ points $(a_k, b_k)$
- straight line shows affine function $f(x) = c_1 + c_2x$ with

$$f(a_k) \approx b_k, \quad k = 1, \ldots, n$$
Least squares regression

use coefficients $c_1, c_2$ that minimize $J = \frac{1}{n} \sum_{k=1}^{n} (f(a_k) - b_k)^2$

- $J$ is a quadratic function of $c_1$ and $c_2$:

\[
J = \frac{1}{n} \sum_{k=1}^{n} (c_1 + c_2 a_k - b_k)^2 \\
= \left( nc_1^2 + 2n \text{avg}(a)c_1c_2 + \|a\|^2c_2^2 - 2n \text{avg}(b)c_1 - 2a^T bc_2 + \|b\|^2 \right) / n
\]

- to minimize $J$, set derivatives with respect to $c_1, c_2$ to zero:

\[
c_1 + \text{avg}(a)c_2 = \text{avg}(b), \quad n \text{avg}(a)c_1 + \|a\|^2c_2 = a^T b
\]

- solution is

\[
c_2 = \frac{a^T b - n \text{avg}(a) \text{avg}(b)}{\|a\|^2 - n \text{avg}(a)^2}, \quad c_1 = \text{avg}(b) - \text{avg}(a)c_2
\]
Interpretation

Slope $c_2$ can be written in terms of correlation coefficient of $a$ and $b$:

$$
c_2 = \frac{(a - \text{avg}(a)1)^T(b - \text{avg}(b)1)}{\|a - \text{avg}(a)1\|^2} = \rho_{ab} \frac{\text{std}(b)}{\text{std}(a)}
$$

- Hence, the expression for regression line can be written as

$$
f(x) = \text{avg}(b) + \rho_{ab} \frac{\text{std}(b)}{\text{std}(a)} (x - \text{avg}(a))
$$

- Correlation coefficient $\rho_{ab}$ is the slope after converting to standard units:

$$
\frac{f(x) - \text{avg}(b)}{\text{std}(b)} = \rho_{ab} \frac{x - \text{avg}(a)}{\text{std}(a)}
$$
Examples

\[ \rho_{ab} = 0.91 \quad \rho_{ab} = -0.89 \quad \rho_{ab} = 0.25 \]

- dashed lines in top row show average ± standard deviation
- bottom row shows scatter plots of top row in standard units
Outline

- norm
- distance
- $k$-means algorithm
- angle
- complex vectors
Norm

norm of vector \( a \in \mathbb{C}^n \):

\[
\|a\| = \sqrt{|a_1|^2 + |a_2|^2 + \cdots + |a_n|^2} = \sqrt{a^H a}
\]

- Positive definite:

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

- Homogeneous:

\[
\|\beta a\| = |\beta|\|a\| \quad \text{for all vectors } a, \text{ complex scalars } \beta
\]

- Triangle inequality:

\[
\|a + b\| \leq \|a\| + \|b\| \quad \text{for all vectors } a, b \text{ of equal size}
\]
Cauchy–Schwarz inequality for complex vectors

\[ |a^H b| \leq ||a|| ||b|| \quad \text{for all } a, b \in \mathbb{C}^n \]

moreover, equality \( |a^H b| = ||a|| ||b|| \) holds if:

- \( a = 0 \) or \( b = 0 \)
- \( a \neq 0 \) and \( b \neq 0 \), and \( b = \gamma a \) for some (complex) scalar \( \gamma \)

- exercise: generalize proof for real vectors on page 2.4
- we say \( a \) and \( b \) are orthogonal if \( a^H b = 0 \)
- we will not need definition of angle, correlation coefficient, \( \ldots \) in \( \mathbb{C}^n \)