Norm and Distance

CE40282-1: Linear Algebra
Hamid R. Rabiee and Maryam Ramezani
Sharif University of Technology

The reason to use norms

- Machine learning uses vectors, matrices, and tensors as the basic units of representation
- Two reasons to use norms
  - To estimate how **big** a vector/matrix/tensor is
    - How big is the difference between two tensors is
  - To estimate how **close** one tensor is to another
    - How close is one image to another

Euclidean Norm

- Euclidean Norm (2-norm, l_2 norm, length)
  - Corresponds to our usual notation of distance

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x^T x}$$

  - It is a nonnegative scalar
  - In R^2follows from the Pythagorean Theorem.
  - What about R^3?

Euclidean Norm

- Euclidean Norm (2-norm, l_2 norm, length)
  - A vector whose length is 1 is called a unit vector
  - Normalizing: divide a nonzero vector by its length which is a unit vector in the same direction of original vector
  - What is the shape of $||x||_2 = 1$?

Vector Norms

- p-norm:

$$||x||_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

$$p \geq 1$$

- What is the shape of $||x||_p = 1$?

Vector Norms Properties

- Absolute homogeneity/Linearity:
  - $||\alpha x|| = |\alpha| x$
- Subadditivity/Triangle inequality
  - $||x + y|| \leq ||x|| + ||y||$
- Positive definiteness/Point separating
  - If $||x|| = 0$ then $x = 0$
  - For every x, $||x|| = 0$ if and only if $x = 0$
- Non-negativity
  - $||x|| \geq 0$

Root-mean-square value

- Mean-square (MS) value of n-vector x is:

$$\frac{x_1^2 + \cdots + x_n^2}{n} = \frac{\|x\|^2}{n}$$

- Root-mean-square value (RMS)

$$\mathbf{rms}(x) = \sqrt{\frac{x_1^2 + \cdots + x_n^2}{n}} = \frac{\|x\|}{\sqrt{n}}$$

- The RMS value of a vector x is useful when comparing norms of vectors with different dimensions
- rms(x) gives 'typical' value of $|x_i|$
  - e.g., rms(**1**) = 1 (independent of n)
  - if all the entries of a vector are the same, (a) then the RMS value of the vector is |a|

Nom of sum

- If x and y are vectors:

$$\|x + y\| = \sqrt{\|x\|^2 + 2x^T y + \|y\|^2}.$$

  - Proof:

$$\begin{aligned}
\|x + y\|^2 &= (x + y)^T (x + y) \\
&= x^T x + x^T y + y^T x + y^T y \\
&= \|x\|^2 + 2x^T y + \|y\|^2.
\end{aligned}$$

Norm of block vectors

- suppose a, b, c are vectors
- $\|(a, b, c)\|^2 = a^T a + b^T b + c^T c = \|a\|^2 + \|b\|^2 + \|c\|^2$
- so we have

$$\|(a, b, c)\| = \sqrt{\|a\|^2 + \|b\|^2 + \|c\|^2} = \|(\|a\|, \|b\|, \|c\|)\|$$

(parse RHS very carefully!)

- The norm of a stacked vector is the norm of the vector formed from the norms of the sub vectors.

Chebyshev inequality

- suspose that k of the numbers $|x_1|, \ldots, |x_n|$ are $\geq a$,

  then k of the numbers $x_1^2, \ldots, x_n^2$ are $\geq a^2$

  so $\|x\|^2 = x_1^2 + \cdots + x_n^2 \geq ka^2$

  so we have $k \leq \|x\|^2 / a^2$

  number of x_i with $|x_i| \geq a$ is no more than $\|x\|^2 / a^2$

  this is the *Chebyshev inequality*
- What happens when $\|x\|^2 / a^2 \geq n$?
- No entry of a vector can be larger in magnitude than the norm of the vector

Chebyshev inequality

- Chebyshev inequality is easier to interpret in terms of the RMS value of a vector.

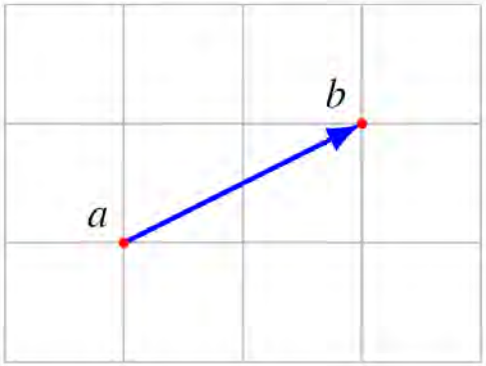
$$\frac{k}{n} \leq \left(\frac{\mathbf{rms}(x)}{a} \right)^2$$

- How many entries of x can have value more than 5rms(x)?
- The Chebyshev inequality partially justifies the idea that the RMS value of a vector gives an idea of the size of a typical entry: It states that not too many of the entries of a vector can be much bigger (in absolute value) than its RMS value

Euclidean distance

- Distance

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



- RMS deviation between the two vectors

$$\mathbf{rms}(a - b) \qquad \|a - b\| / \sqrt{n}$$

Euclidean distance

- Distance between two n-vectors shows the vectors are “`close' or `nearby’” or “far”.

As an example, consider the 4-vectors

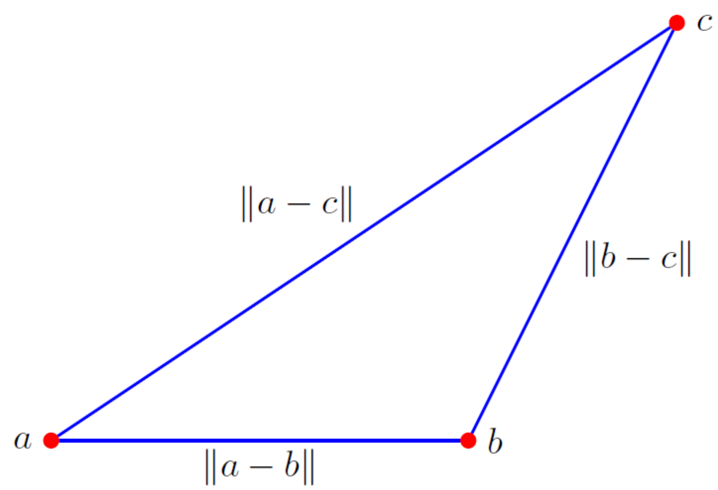
$$u = \begin{bmatrix} 1.8 \\ 2.0 \\ -3.7 \\ 4.7 \end{bmatrix}, \qquad v = \begin{bmatrix} 0.6 \\ 2.1 \\ 1.9 \\ -1.4 \end{bmatrix}, \qquad w = \begin{bmatrix} 2.0 \\ 1.9 \\ -4.0 \\ 4.6 \end{bmatrix}.$$

The distances between pairs of them are

$$\|u - v\| = 8.368, \qquad \|u - w\| = 0.387, \qquad \|v - w\| = 8.533,$$

Triangle inequality

- Consider a triangle in two or three dimensions, whose vertices have coordinates a, b, and c.

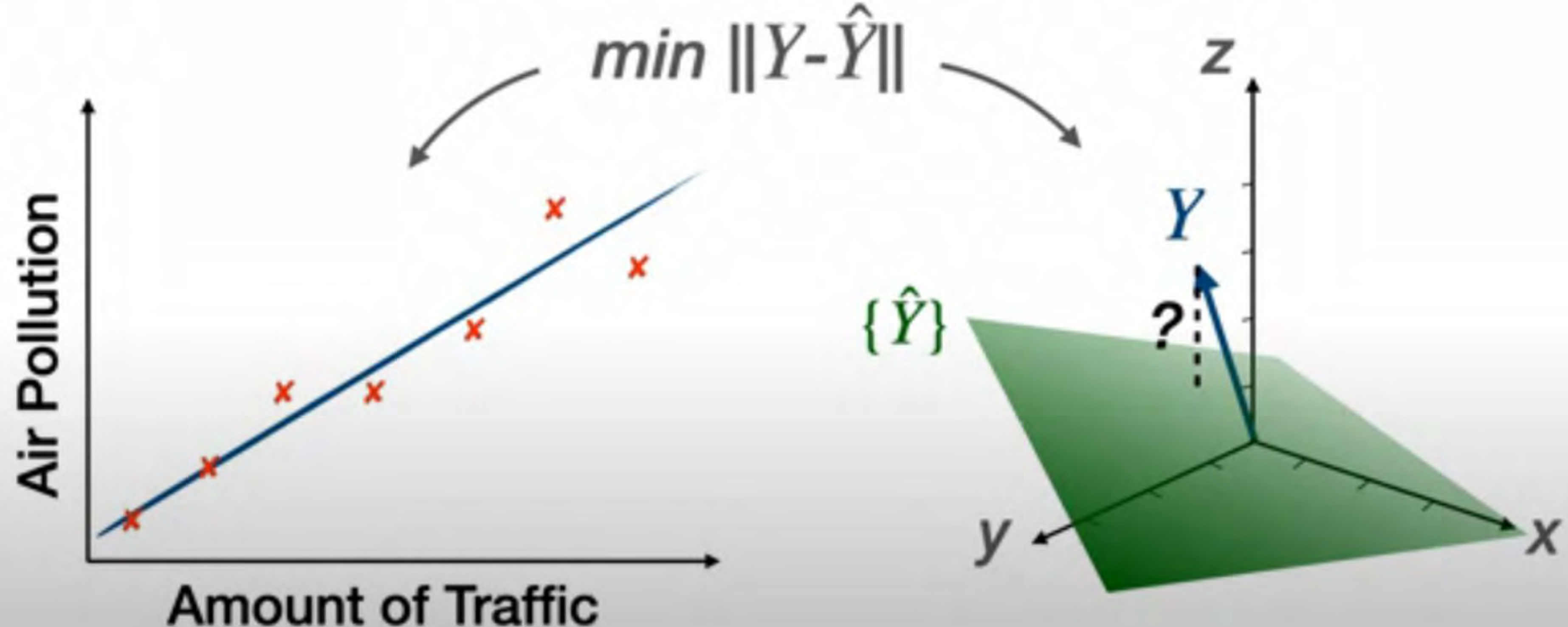


Compare norm and distance

Norm (Normed Linear Space)	*Distance function (Metric Space)*
1. $\|x-y\| \geq 0$	1. $d(x,y) \geq 0$
2. $\|x-y\| = 0 \implies x = y$	2. $d(x,y) = 0 \implies x = y$
3. $\|\lambda(x-y)\| = \|\lambda\| \, \|x-y\|$	3. $d(x,y) = d(y,x)$

ML application

The best linear regression model comes from choosing the closest $\hat{Y}$ to Y based on

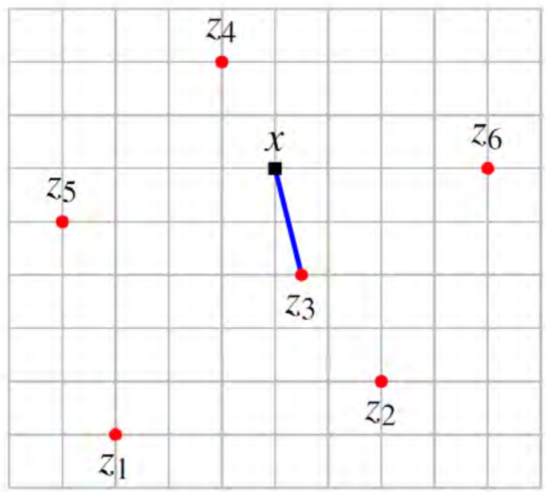


ML Application

Feature distance and nearest neighbors

- if x and y are feature vectors for two entities, $\|x - y\|$ is the *feature distance*
- if $z_1, \ldots, z_m$ is a list of vectors, z_j is the *nearest neighbor* of x if

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



- these simple ideas are very widely used

- Number of flops and order?

ML Application

Document dissimilarity

- 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

Standard deviation

- for n-vector x, $\mathbf{avg}(x) = \mathbf{1}^T x/n$
- *de-meaned vector* is $\tilde{x} = x - \mathbf{avg}(x)\mathbf{1}$ (so $\mathbf{avg}(\tilde{x}) = 0$)
- *standard deviation* of x is

$$\mathbf{std}(x) = \mathbf{rms}(\tilde{x}) = \frac{\|x - (\mathbf{1}^T x/n)\mathbf{1}\|}{\sqrt{n}}$$

- $\mathbf{std}(x)$ gives 'typical' amount x_i vary from $\mathbf{avg}(x)$
- $\mathbf{std}(x) = 0$ only if $x = \alpha\mathbf{1}$ for some α
- greek letters μ, σ commonly used for mean, standard deviation
- a basic formula:

$$\mathbf{rms}(x)^2 = \mathbf{avg}(x)^2 + \mathbf{std}(x)^2$$

Chebyshev inequality for standard deviation

x is an n-vector with mean $\mathbf{avg}(x)$, standard deviation $\mathbf{std}(x)$

rough idea: most entries of x are not too far from the mean

by Chebyshev inequality, fraction of entries of x with

$$|x_i - \mathbf{avg}(x)| \geq \alpha \ \mathbf{std}(x)$$

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

- The fraction of entries of x within θ standard deviations of avg(x) is at least $(1 - \frac{1}{\theta^2})$ for $\theta > 1$

Properties of standard deviation

- *Adding a constant.* For any vector x and any number a, we have $\mathbf{std}(x+a\mathbf{1}) = \mathbf{std}(x)$. Adding a constant to every entry of a vector does not change its standard deviation.

- *Multiplying by a scalar.* For any vector x and any number a, we have $\mathbf{std}(ax) = |a|\,\mathbf{std}(x)$. Multiplying a vector by a scalar multiplies the standard deviation by the absolute value of the scalar.

Vector Standardization

$$z = \frac{1}{\mathbf{std}(x)}(x - \mathbf{avg}(x)\mathbf{1}).$$

- It has mean zero, and standard deviation one.
- Its entries are sometimes called the z-scores associated with the original entries of x.
- The standardized values for a vector give a simple way to interpret the original values in the vectors.

Vector Norms

- 1-norm: $(\boldsymbol{l_1})$

$$||x||_1 = (|x_1| + |x_2| + \cdots + |x_n|)$$

- What is the shape of $||x||_1 = 1$?

Vector Norms

- ∞-norm: ($\boldsymbol{l}_\infty$) (max norm)

$$\infty = max(|x_1|, |x_2|, \dots, |x_n|)$$

- What is the shape of $||x||_\infty = 1$?

Vector Norms

- $\frac{\mathbf{1}}{\mathbf{2}}$-norm: $(\boldsymbol{l}_{\frac{1}{2}})$

- What is the shape of $||x||_{\frac{1}{2}} = 1$?

Vector Norms

- $\boldsymbol{zero}$-norm: $(\boldsymbol{l_0})$

$$||x||_0 = \lim_{\alpha \to 0^+} ||x||_\alpha = \left(\sum_{k=1}^{n} |x|^\alpha \right)^{1/\alpha} = \sum_{k=1}^{n} 1_{(0,\infty)}(|x|)$$

- Zero-norm, defined as **the number of non-zero elements in a vector**, is an ideal quantity for feature selection. However, minimization of zero-norm is generally regarded as a combinatorically difficult optimization
- $||x||_0 = \sum_{x_i \neq 0} 1$
- Normalizing $||.||_0$ leads to $0 \leq ||.||_0 \leq 1$

Feedback ☺

- https://forms.gle/9QgxS98RJDSjhHCu8

Reference

- Linear Algebra and Its Applications, David C. Lay, Chapter 6.

- Introduction to Applied Linear Algebra Vectors, Matrices, and Least Squares, Stephen Boyd, Chapter 3.