COMP229: Introduction to Data Science
Lecture 16: Metric axioms

Olga Anosova, O.Anosova@liverpool.ac.uk
Autumn 2023, Computer Science department
University of Liverpool, United Kingdom

Lecture plan

- Metric
- Euclidean metric in $\mathbb{R}$
- Euclidean metric in $\mathbb{R}^n$
- L_s metric
- Shortest path distance in graphs

Reminder: SLR

- The *least-squares regression line* minimises the sum of squared *vertical* distances from points.
- The regression line $y = ax + b$ has $b = \bar{y} - a\bar{x}$, $a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ and passes through the point $(\bar{x}, \bar{y})$, where $\bar{x}, \bar{y}$ are the sample means.
- A regression line $y = ax + b$ may not be symmetric with respect to x, y, i.e. swapping x, y may give another regression $x = cy + d$.
- SLR predictions can be useful, but misleading.

The key questions about data

The real-life question 'what is data?' can be split into the following important subquestions:

What is a single data point? Often it is a sequence of numbers. How many coordinates?

What data points are considered equivalent? Is it an equivalence relation?

If data points are different, how different they are? How can a similarity between points be measured?

The Euclidean metric in $\mathbb{R}^n$

If data are given as points in $\mathbb{R}^n$, one of *infinitely many ways* to measure a similarity or a distance between $p = (p_1, \dots, p_n)$ and $q = (q_1, \dots, q_n)$ is the *Euclidean metric* $L_2(p, q)$

$$= \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

$$= \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}$$

The key message of this lecture: there are *infinitely many other ways* to measure a distance, which are often better suited for specific applications.

Axioms for a metric (distance)

Definition 16.1. For any set C of arbitrary elements, a *metric* (distance) is a real-valued function $d : C \times C \to \mathbb{R}$ satisfying the axioms:

(1) *identity* : $d(p, q) = 0$ if and only if $p = q$;

(2) *symmetry* : $d(p, q) = d(q, p)$ for any $p, q \in C$;

(3) *triangle inequality* (draw a triangle on p, q, r) : $d(p, q) + d(q, r) \geqslant d(p, r)$ for any $p, q, r \in C$.

The positivity of a metric

Claim 16.2. Any metric from Definition 14.1 satisfies *positivity* : $d(p,q) \geqslant 0$ for any $p,q \in C$.

Proof. If we set $r = p$, then the triangle inequality is $d(p,q) + d(q,p) \geqslant d(p,p)$. Apply the symmetry and identity: $d(p,q) + d(p,q) \geqslant 0$, $d(p,q) \geqslant 0$.

Often the positivity is included in the 1st axiom, but the identity condition can't be missed. Why not?
Without it, the trivial example is $d(p,q) = 0$ for all p,q collapses all points together and does not allow any measurement of difference between points.

A 1-dimensional case

Data points can be real numbers, points in $\mathbb{R}^n$, matrices, images, molecules, people or anything.

In the simplest case when data points are real numbers, e.g. ages of students, how would you measure a distance between numbers $p, q \in \mathbb{R}$?

The Euclidean metric is $L_2(p, q) = |p - q|$. The absolute value of $x \in \mathbb{R}$ is $|x| = \begin{cases} x & \text{for } x \geqslant 0, \\ -x & \text{for } x < 0. \end{cases}$

Examples

Problem 16.3. Is $d(p,q) = p - q$ a metric in $\mathbb{R}$?
Solution 16.3. $d(p,q) = p - q$ is not a metric, because the symmetry axiom fails: $d(0,1) \neq d(1,0)$ shows that $d(p,q) = p - q$ isn't a metric on $\mathbb{R}$.

Problem 16.4. Is $d(p,q) = (p-q)^2$ a metric?

Solution 16.4. No since the triangle axiom fails:
$d(-1,0) + d(0,1) = 1^2 + 1^2 = 2 < d(-1,1) = 4$, though the identity and symmetry axioms hold.

The Euclidean metric on $\mathbb{R}$

Problem 16.5. Is $d(p, q) = |p - q|$ a metric on $\mathbb{R}$?

Solution 16.5. Check all the axioms for any real $p, q, r \in \mathbb{R}$.

(1) $|p - q| = 0$ if and only if $p = q$, true.

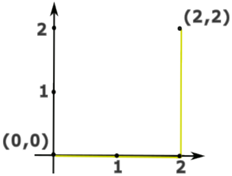
(2) symmetry: $|p - q| = |q - p|$, true.

(3) triangle inequality: if $p \geqslant q \geqslant r$, then
$|p - q| + |q - r| = (p - q) + (q - r) = p - r = |p - r|$,
(sketch 3 points in $\mathbb{R}$). Other cases are easy: for $p \geqslant r \geqslant q$,
$|p - q| = p - q \geqslant p - r = |p - r|$.

From $\mathbb{R}$ to $\mathbb{R}^2$

How to define a metric in $\mathbb{R}^2$?

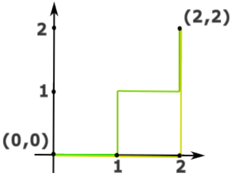
Let's try $|p_1 - q_1| + |p_2 - q_2|$ for $p = (p_1, p_2)$ and $q = (q_1, q_2)$.

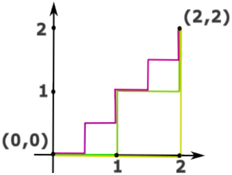


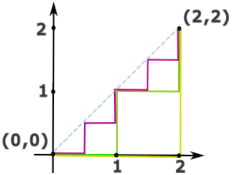
This distance between $(0, 0)$ and $(2, 2)$ is

$|2 - 0| + |2 - 0| = 4$.

Is it the shortest path?







The metric axioms of L_2

Claim 16.6. The Euclidean metric $L_2(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$ satisfies the metric axioms.

Proof. The identity and symmetry axioms are easy. The triangle inequality for $\triangle pqr$ in terms of vectors $\vec{u} = \overrightarrow{pq}, \vec{v} = \overrightarrow{qr}$ says that $|\vec{u} + \vec{v}| \leqslant |\vec{u}| + |\vec{v}|$. Any vector $\vec{w}$ has angle 0 with itself, so $\vec{w} \cdot \vec{w} = |\vec{w}|^2$.

Then $(\vec{u} + \vec{v}) \cdot (\vec{u} + \vec{v}) \leqslant |\vec{u}|^2 + 2|\vec{u}| \cdot |\vec{v}| + |\vec{v}|^2$ follows from Cauchy's inequality $\vec{u} \cdot \vec{v} \leqslant |\vec{u}| \cdot |\vec{v}|$.

Other metrics on $\mathbb{R}^n$

Definition 16.7. For any real $s \geqslant 1$ and $p, q \in \mathbb{R}^n$, the L_s-**metric** is $L_s(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^s\right)^{1/s}$.

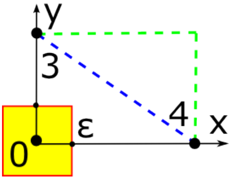
For $s = 1$, $L_1(p,q) = \sum_{i=1}^{n} |p_i - q_i|$ is also called the **Manhattan** metric.

When $s \to +\infty$, the limit case gives the **max** (or Chebyshev) metric $L_\infty(p,q) = \max_{i=1,\ldots,n} |p_i - q_i|$.

The balls in other metrics

Problem 16.8. $p = (4,0)$, $q = (0,3)$. Find L_1, L_∞. What is the unit ball $B = \{p \in \mathbb{R}^2 : L_\infty(p,0) \leqslant 1\}$?
$B = \{p \in \mathbb{R}^2 : L_1(p,0) \leqslant 1\}$?

Solution 16.8. Manhattan has vertical avenues (north-to-south), horizontal streets (east-to-west), hence it's known as *taxicab* (or *Manhattan*) L_1-metric.



$$L_1(p,q) = |4-0| + |0-3| = 7.$$
$$L_\infty(p,q) = \max\{|4-0|, |0-3|\} = 4.$$

The yellow square is the ball of radius $\varepsilon = 1$ in L_∞.
The unit ball in L_1 is the square $|x \pm y| \leqslant 1$.

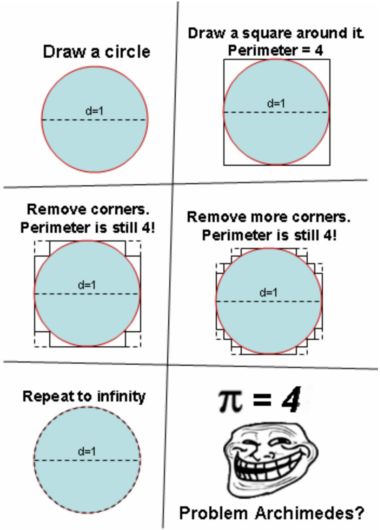
L_∞ as a chessboard distance

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

L_∞ gives the minimum number of moves a king requires to move between two chessboard squares.

Here are the L_∞ distances from the square F6.

Reminder: new formula for π



Indeed in L_1 the value of a geometric analog to π is 4.

L_1 is a handy tool to measure the differences in discrete frequency distributions.

Graphs with vertices and edges

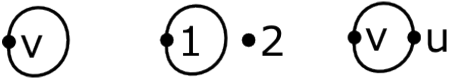
Definition 16.9. A (unoriented) **graph** is a pair (vertices,edges), where **vertices** V form a finite set of $|V|$ elements, and **edges** form a set E defined by unordered pairs of vertices.

$|V|$ = number of vertices, $|E|$ = number of edges.
Other names: graph = network, vertex = node, edge = link (or connection), oriented = directed.

The graph with a single edge connecting two vertices (labelled by 1,2) can be described by the pair $(\{1,2\},\{(1,2)\})$ or by the pair $(\{1,2\},\{(2,1)\})$. Sometimes only the number of vertices with the list of edges is used: $|V| = 2, \{(1,2)\}$.

More conventions and examples

For any vertex v, the pair (v, v) represents a **loop** at v (one edge connecting a vertex to itself).



For $|V| = 2$, the list $\{(1, 1)\}$ denotes the graph consisting of one loop and the isolated vertex 2.

For vertices u, v, the repeated pair $(u, v), (u, v)$ in a list represents a **double** edge between u, v. The list $(1, 2), (2, 3), (3, 1)$ represents a triangular cycle.

A metric graph

Definition 16.10. A graph G is **connected** if any two vertices are connected by a **path** (sequence) of edges $e_1, \dots, e_k$ such that any successive edges e_i, e_{i+1} share a vertex for $i = 1, \dots, k-1$.

Let associate to every edge a non-negative length or weight. The length of any path is the sum of lengths of its edges. For any vertices u, v, the **shortest path distance** (or **graph geodesic**) is the length of a shortest path from u to v.

Time to revise and ask questions

- A metric (distance) should satisfy the axioms of identity, symmetry, triangle inequality.
- $L_s(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^s\right)^{1/s}$ in $\mathbb{R}^n$, $s \geqslant 1$.

Problem 16.11. Is the shortest path distance of any connected graph a metric?