COMP229: Introduction to Data Science
Lecture 2: data clouds and metric spaces

Vitaliy Kurlin, vitaliy.kurlin@liverpool.ac.uk
Autumn 2018, Computer Science department
University of Liverpool, United Kingdom

The scanner for registrations will be from week 2.

What is data in COMP229?

In many applications, *data* (often written in a singular form, not plural) is given as a cloud.

Definition 2.1. A *point cloud* is a finite set C of points with a real-value function $d : C \times C \to \mathbb{R}$ (called a *metric*) satisfying the axioms below:

(1) *positivity*: $d(p, q) \geq 0$ for any $p, q \in C$, and $d(p, q) = 0$ if and only if $p = q$ (points coincide);

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

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

1-dimensional non-example

Data points can be real numbers, points in $\mathbb{R}^m$, 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 $p, q \in \mathbb{R}$?

Is $d(p, q) = p - q$ or $d = q - p$ a metric in $\mathbb{R}$?

No, because the positivity axiom isn't satisfied, e.g. the example $d(1, 2) = 1 - 2 < 0$ disproves the conjecture that $d(p, q) = p - q$ is a metric.

1-dimensional example

Claim 2.2. $d(p, q) = |p - q|$ is a metric on $\mathbb{R}$.

Proof. We check all the axioms for any real $p, q, r \in \mathbb{R}$. When conclusions are simple, it's enough to explicitly write them as below.

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

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

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

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

Definition 2.3. For points $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$, the *Euclidean metric* is

$$L_2(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}. \tag{2.3}$$

Claim 2.4. The function L_2 in (2.3) is a metric.

The positivity and symmetry axioms are simple, the triangle inequality follows from Pythagoras' theorem (to be discussed in Lecture 6).

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

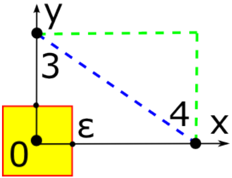
Definition 2.5. For any real $s \geq 1$, points $p = (p_1, \ldots, p_n)$, $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$, the *$L_s$-metric* is $L_s(p,q) = \left(\sum_{i=1}^{n} |p_i - q_i|^s \right)^{1/s}$ (2.5).

For $s = 1$, $L_1(p,q) = \sum_{i=1}^{n} |p_i - q_i|$ is also called a *Manhattan* metric. When $s \to +\infty$, the limit case gives the *max* metric $L_\infty(p,q) = \max_{i=1,\ldots,n} |p_i - q_i|$.

Compute L_2, L_1, L_∞ for $p = (4,0)$, $q = (0,3) \in \mathbb{R}^2$.

Example computations

$p = (4, 0)$, $q = (0, 3)$. Compute by the definitions:



$$L_2(p, q) = \sqrt{(4-0)^2 + (0-3)^2} = 5,$$
$$L_1(p, q) = |4-0| + |0-3| = 7,$$
$$L_\infty(p, q) = \max\{|4-0|, |0-3|\} = 4.$$

Any point cloud C of n ordered points can be represented by its *distance* $n \times n$ *matrix* d_{ij} = distance between i-th and j-th points of C.

Most general algorithms accept any distance matrix as an input, many metrics can be tried.

A metric space

Definition 2.6. A *metric space* is any set C with $d : C \times C \to \mathbb{R}$ (called a *metric*) such that

(1) *positivity*: $d(p,q) \geq 0$ for any $p,q \in C$, and $d(p,q) = 0$ if and only if $p = q$ (points coincide);

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

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

A point cloud in Def 2.1 is a finite metric space.

Claim 2.7. For any metric d, $t > 0$, td is a metric.

Outline. All axioms hold for d, hence for td. $\square$

Examples and non-examples

Claim 2.8. $d(p,q) = \begin{cases} 0 & \text{for } p = q, \\ 1 & \text{for } p \neq q \end{cases}$ is a metric.

Outline. Axiom (3) $d(p,q) + d(q,r) \geq d(p,r)$ can fail only if $d(p,q) = 0 = d(q,r)$, so $p = q = r$. $\square$

For real p, q, are these functions metrics?

1) $d(p,q) = |p^2 - q^2|$. No, because $d(1,-1) = 0$.

2) $d(p,q) = |p - 2q|$. No since $d(0,1) \neq d(1,0)$.

3) $d(p,q) = (p-q)^2$. No, because axiom (3) fails: $d(1,2) + d(2,3) = 1 + 1 < d(1,3) = 4$.

Other distances (non-metrics)

If positivity axiom (1) is replaced by the weaker (1') $d(p,q) \geq 0$ and $d(p,p) = 0$, possibly $d(p,q) = 0$ for $p \neq q$, we get a *pseudometric*, which are common in applications when objects are compared only by their partial descriptors, e.g. distance(people) = abs. difference of ages.

If symmetry axiom (2) is dropped, we get a *quasimetric*, e.g. the time to get from one place to another (via hills or 1-way roads in a town).

More examples are in Encyclopedia of Distances.

Your questions and the quiz

To benefit from the lecture, now you could

- ask or submit your anonymous questions to the COMP229 folder after the lecture;
- write down your summary in 2-3 phrases, e.g. list key concepts you have learned;
- talk to your classmates to revise the lecture.

Question. Is the function $d(p, q) = p^2 - pq + q^2$ a metric on real numbers?

Answer to the quiz and summary

Answer. No. Axiom (1) fails: $d(1,1) = 1$.
Axiom (2) holds: $d(p,q) = p^2 - pq + q^2 = d(q,p)$.
Axiom (3) fails: $d(0, \pm 1) = 1$, $d(1, -1) = 3$.

- A metric space has a metric satisfying the positivity, symmetry, triangle inequality.
- The common metrics on $\mathbb{R}^n$ are L_1, L_2, L_∞.
- There are many useful non-metric functions.
- More results can be proved for metric spaces, e.g. on convergence of iterative algorithms.