Lecture notes - Clustering and Persistence SF2704

Teacher: Wojciech Chachólski

Jim Holmström
Ariel Ekgren
Gabriel Isheden
{jimho, ekgren, ...}@kth.se

October 13, 2013

1 introduction

The big goal for this lecture series is to understand metrics by searching for a metric between metric spaces. This has been proven to be impossible (ref), but dispite the lack of a grand metric of metrics we can still try to search within (common word for ultrametrics, barcodes etc) to at least get a grasp on certain parts.

2 notes

note 1

Skeleton for note 1

note 2

Metric space (X, d) where X is a finite set and d is a distance between the points in this finite set.

$$|X| < \infty \tag{1}$$

$$d(x,x) = 0, d(x,y) = d(y,x), d(x,y) + d(y,z) \leq d(x,z) \forall x,y,z \in X \tag{2}$$

An intresting submetric is the *ultrametric* which has a strong triangle inequality contraint

$$\max\{d(x,y), d(y,z)\} \leq d(x,z) \forall x,y,z \in X \tag{3}$$

also called ultrametric inequality.

Stated in another way, for an ultrametric any 3 points, i.e. a triangle, will have the following property

<image of triangle with the sides (a, a, b <= a)>

P (X) is the set of partitions of X. A partition of X

$$\sigma = \{u_i\} = \bigsqcup_i u_i \in \mathrm{P}\ (X) \tag{4}$$

where $u_i \in \sigma$ is called a block.

To be able to order partitionings we define

$$\sigma \leq \tau \Leftrightarrow \forall u \in \sigma \exists v \in \tau : u \subset v \tag{5}$$

A *clustering* is just a function Ψ in the form of a algorithm or procedure which maps a metric space (X, d) to a partition of X

$$\Psi : (X, d) \to \mathrm{P}\ (X) \tag{6}$$

The function $\Phi(X, d) \in \mathrm{P}\ (X)$ can have 3 important properties

Scale invariant Changing the scale for the distance does not change the partitioning

$$\Psi(X, d) = \Psi(X, \alpha d) \forall \alpha \in \mathbb{R}_{++} \tag{7}$$

Rich $\Psi(X, d) \twoheadrightarrow \mathrm{P}\ (X)$ i.e. surjective or onto the set of partitions of X.

$$\forall \sigma \mathrm{P}\ (X)\, \exists d : \Psi(X, d) = \sigma \tag{8}$$

Consistent d' is such that you decrease the intrablock distance and increase the extrablock distance for all blocks. This will correspond to making the clusters more distinct. Let $x \sim_{\Psi(X,d)} y$ denote that x and y belongs to the same block of the clustering $\Psi(X, d)$. Then d' is a transformation such that:

$$d'(x, y) : \begin{cases} \leq d(x, y) & x \sim_{\Psi(X,d)} y \\ \geq d(x, y) & x \not\sim_{\Psi(X,d)} y \end{cases} \tag{9}$$

(not in notes; but it should be that $\Psi(X, d) = \Psi(X, d') \forall d'$ fullfilling the above property)

According to the Kleinberg theorem¡ref¿ a Φ satisfying all these properties does not exists.

Consider the set of 3 points $\{a, b, c\}$, then a metric can be represented by a matrix

	a	b	c
a	0	x	y
b	x	0	z
c	y	z	0

(10)

which satisfies the triangle inequality

$$\begin{cases} d(a,b) + d(b,c) \leq d(a,c) \\ d(a,c) + d(c,b) \leq d(a,b) \\ d(b,a) + d(a,c) \leq d(b,c) \end{cases} \Rightarrow \begin{cases} x + z \leq y \\ y + z \leq x \\ x + y \leq z \end{cases} \tag{11}$$

That is the distances is

$$\left\{ (x,y,z) \middle| \begin{matrix} x + z \leq y \\ y + z \leq x \\ x + y \leq z \end{matrix} \right\} \tag{12}$$

which is an intersection of 3 halfspaces.

¡def¿ An isometry

$$f : (X, d_x) \to (Y, d_y) \tag{13}$$

is a bijection such that

$$d_x(x_i, x_j) = d_y(f(x_i), f(x_j)) \tag{14}$$

in other words a distance preserving map.

$$\mapsto \begin{array}{ccc} & \text{c} & \\ \text{z} & & \text{y} \\ \text{a} & \text{x} & \text{b} \end{array} \tag{15}$$

note 9

Skeleton for note 9.

3 exercises

exercise 1

4 code