Distances & vectorization methods for persistent homology

finite

Persistent diagram is a multiset of points in $\mathbb{R}^2$ such that first coordinate (birth time) $\leq$ second coordinate (death time).

Space of all persistence diagrams is equipped with a number of standard metrics. We will introduce a few of them; To do that let;

P_1 - one persistence diagram, P_2 - second persistence diagram.

For technical reason that will soon become clear we add all the points from $y=x$, with infinite multiplicity, both to P_1 & P_2.

That trick make P_1 & P_2 having the same (infinite) size.

Let us then consider all bijections

$$b : P_1 \to P_2$$

For a fixed b we can compute;

$$\max_{p \in P_1} \left(\text{dist}(p, b(p)) \right) =: W_b^\infty$$

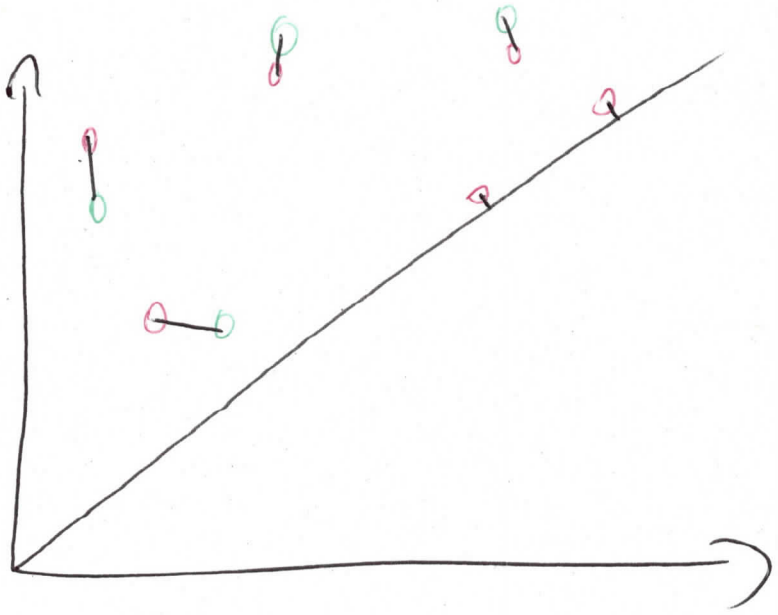
or

$$\left(\sum_{p \in P_1} \left(\text{dist}(p, b(p)) \right)^q \right)^{\frac{1}{q}} =: W_b^q$$

Taking $\inf_b W_b^\infty$ gives a bottleneck distance, and

$\inf_b W_b^q$ gives a Wasserstein distance.

There is an obvious geometric interpretation.
Let green circles denote points in P_1 and red squares, points from P_2. Every matching b can be visualized as a line segment joining $P_1 \ni p \to b(p) \in P_2$



Taking a bijection b it can be visualized as a line segment. W_b^∞ correspond to the length of the longest edge in b, while W_b^q, to (the sum of all lengths (to power q)) to power $\frac{1}{q}$.

Note that some of the points are matched to digonal and of course almost all point in digonal are matched to the corresponding points of the second diagram.

When speaking about space of persistence diagrams we often require that all diagrams are finite, or sometimes that sum of persistence of points in each diagram is finite. It is then complete, separable metric space.

However this is not a fixed-size representation, which makes it difficult to apply statistics & machine learning methods on persistence diagrams.

Therefore a number of vectorization methods like persistence landscapes, images, by of warbs, has been introduced.

With them one can perform most of statistical and machine learning options on persistence diagrams.