Topological Data Analysis

3 December 2019

8 Stability of persistence diagrams

8.1 Hausdorff distance

Suppose that X and Y are nonempty subsets of a metric space M with distance d. The *diameter* of X is defined as

$$\mathrm{diam}(X) = \sup\{d(x, x') \mid x, x' \in X\},$$

and X is *bounded* if its diameter is finite.

For a point $x \in X$, define

$$d(x, Y) = \inf\{d(x, y) \mid y \in Y\},$$

and then

$$d(X, Y) = \sup\{d(x, Y) \mid x \in X\}.$$

If X is bounded, then $d(X, Y)$ is finite; in fact, $0 \le d(X, Y) \le \mathrm{diam}(X) + d(x_0, y_0)$ for arbitrary points $x_0 \in X$ and $y_0 \in Y$. However, $d(X, Y) \ne d(Y, X)$ in general.

Now suppose that X and Y are compact (hence bounded and closed in the metric space M). The *Hausdorff distance* between X and Y is defined as

$$d_H(X, Y) = \max\{d(X, Y),\, d(Y, X)\}.$$

Note that if $d(X, Y) = 0$ then $X \subseteq Y$, since Y is closed in M. Consequently, $d_H(X, Y) = 0$ if and only if $X = Y$. Moreover, d_H satisfies the triangle inequality and therefore d_H is indeed a distance on the set of nonempty compact subsets of M. However, if $M = \mathbb{R}$, $X = \mathbb{Q}$ and $Y = \mathbb{R} \smallsetminus \mathbb{Q}$ (not compact) then $d_H(X, Y) = 0$.

8.2 Gromov–Hausdorff distance

For X and Y nonempty compact metric spaces, the *Gromov–Hausdorff distance* between X and Y is defined as

$$d_{GH}(X, Y) = \inf\{d_H(f(X), g(Y)) \mid f\colon X \hookrightarrow M,\, g\colon Y \hookrightarrow M\},$$

where the infimum is taken over all isometric embeddings $f\colon X \hookrightarrow M$, $g\colon Y \hookrightarrow M$ into some common metric space M. Hence $d_{GH}(X, Y) = 0$ if and only if X and Y are isometric.

An alternative description of the Gromov–Hausdorff distance is as follows. For nonempty compact metric spaces X and Y, a *correspondence* between X and Y is a surjective multivalued function from X to Y, that is, a subset $C \subseteq X \times Y$ where for all $x_0 \in X$ there is some $(x_0, y) \in C$ and for all $y_0 \in Y$ there is some $(x, y_0) \in C$.

If C is a correspondence between X and Y, then the inverse correspondence C^{-1} is the set of points $(y,x) \in Y \times X$ for which $(x,y) \in C$.

The *distortion* of a correspondence $C \subseteq X \times Y$ is defined as

$$\mathrm{dis}(C) = \max\{|d_X(x,x') - d_Y(y,y')| \,:\, (x,y),(x',y') \in C\}.$$

For example, If $C = \{(x, f(x)) \mid x \in X\}$ for some surjective function $f\colon X \to Y$, then $\mathrm{dis}(C) = 0$ if and only if f is an isometry.

Then the following holds:

$$d_{GH}(X,Y) = \frac{1}{2} \inf\{\mathrm{dis}(C) \mid C \subseteq X \times Y\},$$

where the infimum is taken over all correspondences between X and Y.

8.3 Bottleneck distance

Suppose given persistence modules (V,π) and (V',π') with persistence diagrams D and D', viewed as sets of points with multiplicities in a birth-death coordinate plane. As customary, points in the diagonal $b = d$ are included with infinite multiplicity. In what follows, we will assume that the barcodes of (V,π) and (V',π') have the same number of infinite rays, and include each infinite ray $[a,\infty)$ in the corresponding persistence diagram as a point (a, y_∞) for an arbitrary but fixed value y_∞ far above the spectral values of V and V'.

A bijective function $\varphi\colon D \to D'$ is called a *matching* between D and D' if each diagonal point $(b,b) \in D$ is either matched with itself or with an off-diagonal point $(x,y) \in D'$, that is, with $x \neq y$. For each such matching, write

$$\|\varphi\| = \sup\{d_\infty((x,y),\, \varphi(x,y)) \mid (x,y) \in D\},$$

where d_∞ is the ℓ_∞-distance on $\mathbb{R}^2$, namely

$$d_\infty((x,y),(x',y')) = \max\{|x-x'|,\, |y-y'|\}.$$

The *bottleneck distance* between D and D' is defined as

$$W_\infty(D,D') = \inf\{\|\varphi\| \mid \varphi\colon D \to D'\},$$

where the infimum is taken over all matchings $\varphi\colon D \to D'$.

Hence the bottleneck distance between D and D' is the smallest $\varepsilon \geq 0$ for which there exists a matching between the points of D and D' (including all their diagonal points) such that any couple of matched points are at distance at most ε, where the distance between points in $\mathbb{R}^2$ is the ℓ_∞-distance.

The bottleneck distance is the case $p = \infty$, $q = \infty$ of the *Wasserstein distances*, defined for $p, q \geq 1$ as

$$W_p[q](D,D') = \inf_{\varphi\colon D \to D'} \left[\sum_{(x,y)\in D} d_q((x,y),\, \varphi(x,y))^p \right]^{1/p},$$

where d_q is the ℓ_q-distance on $\mathbb{R}^2$:

$$d_q((x,y),(x',y')) = \Big(|x-x'|^q + |y-y'|^q\Big)^{1/q}.$$

8.4 Interleaving distance

For a persistence module (V, π) and $\delta \in \mathbb{R}$, define another persistence module by $V[\delta]_t = V_{t+\delta}$ and $\pi[\delta]_{s,t} = \pi_{s+\delta,\,t+\delta}$. This is called a *$\delta$-shift* of (V, π). If $\delta \geq 0$, then there is a morphism of persistence modules $\sigma_\delta \colon V \to V[\delta]$ given by $(\sigma_\delta)_t = \pi_{t,\,t+\delta}$ for all $t \in \mathbb{R}$. Moreover, each morphism $f \colon V \to V'$ of persistence modules yields a morphism $f[\delta] \colon V[\delta] \to V'[\delta]$ for all $\delta \in \mathbb{R}$, namely $f[\delta]_t = f_{t+\delta}$ for all $t \in \mathbb{R}$.

For $\delta > 0$, two persistence modules (V, π) and (V', π') are *δ-interleaved* if there exist morphisms $f \colon V \to V'[\delta]$ and $g \colon V' \to V[\delta]$ such that

$$g[\delta] \circ f = \sigma_{2\delta} \quad \text{and} \quad f[\delta] \circ g = \sigma'_{2\delta}.$$

If (V, π) and (V', π') are δ-interleaved for some $\delta > 0$, then $\dim_{\mathbb{F}} V_\infty = \dim_{\mathbb{F}} V'_\infty$ and hence $V_\infty \cong V'_\infty$.

The *interleaving distance* between two persistence modules (V, π) and (V', π') with $V_\infty \cong V'_\infty$ is defined as

$$d_{\mathrm{int}}(V, V') = \inf\{\delta > 0 \mid (V, \pi) \text{ and } (V', \pi') \text{ are } \delta\text{-interleaved}\}.$$

It follows that, if $a < b$ and $c < d$, then

$$d_{\mathrm{int}}(\mathbb{F}[a, b), \mathbb{F}[c, d)) = \min\left\{\max\left\{\tfrac{1}{2}(b-a),\, \tfrac{1}{2}(d-c)\right\},\, \max\left\{|a-c|, |b-d|\right\}\right\}$$

while

$$d_{\mathrm{int}}(\mathbb{F}[a, \infty), \mathbb{F}[c, \infty)) = |a - c|.$$

8.5 Isometry Theorem

The interleaving distance between two persistence modules (V, π) and (V', π') with $V_\infty \cong V'_\infty$ is equal to the bottleneck distance between their persistence diagrams:

$$d_{\mathrm{int}}(V, V') = W_\infty(D(V), D(V')).$$

8.6 Stability Theorem

In this section we state the Stability Theorem for point clouds. Other versions will be discussed later. Thus, let X and Y be point clouds in $\mathbb{R}^N$ for some N. Let $V_t(X) = H_*(R_t(X))$ be the Vietoris–Rips persistence module of X, and similarly for Y. Then the interleaving distance between $V(X)$ and $V(Y)$ is related with the Gromov–Hausdorff distance between X and Y as follows:

$$\frac{1}{2}\, d_{\mathrm{int}}(V(X), V(Y)) \leq d_{GH}(X, Y).$$

As a consequence of this fact and the Isometry Theorem, since $d_{GH}(X, Y) \leq d_H(X, Y)$, we have that

$$W_\infty(D(X), D(Y)) \leq 2\, d_H(X, Y),$$

where $D(X)$ and $D(Y)$ denote the Vietoris–Rips persistence diagrams of X and Y.

Short exercise

(1) Consider the following point clouds in $\mathbb{R}^2$:

$$X = \{(0.81, 2.87), (2.15, 1.18), (3.19, 3.62), (4.17, 2.01), (5.32, 4.88), (6.21, 3.13)\},$$

$$Y = \{(0.75, 2.80), (2.33, 1.25), (3.28, 3.66), (4.15, 2.15), (5.24, 4.78), (6.34, 3.12)\}.$$

(a) Compute the Hausdorff distance $d_H(X, Y)$.

(b) Prove that $W_\infty(D(X), D(Y)) < 2\, d_H(X, Y)$, where $W_\infty(D(X), D(Y))$ is the bottleneck distance between the Vietoris–Rips persistence diagrams of X and Y.

Longer exercise

(1) Prove the formula

$$d_{GH}(X, Y) = \frac{1}{2} \inf\{\mathrm{dis}(C) \mid C \subseteq X \times Y\},$$

where the infimum is taken over all correspondences between X and Y. For this, see [F. Mémoli, G. Sapiro, A theoretical and computational framework for isometry invariant recognition of point cloud data, *Found. Comput. Math.* 5 (2005), 313–347].