Topological Data Analysis

19 December 2019

13 Statistical inference using landscapes

Hypothesis testing for point clouds in $\mathbb{R}^N$ can be carried out by means of several kinds of estimators. Here we describe one option, based on norms of landscapes.

For a point cloud X, let $\Lambda(X) = \{\lambda_k(X)\}_{k\geq 1}$ denote the sequence of persistence landscapes associated to the Vietoris–Rips barcode of X. Thus, we have a piecewise linear function $\lambda_k(X)\colon \mathbb{R} \to \mathbb{R}$ with compact support for each $k \in \mathbb{N}$, and $\Lambda(X)$ may be viewed as an element of the Banach space $L^p(\mathbb{N} \times \mathbb{R})$ for every $p \geq 1$.

Now suppose that we treat X as a random variable (for example, a random point cloud on a sphere or a torus). Then $\Lambda(X)$ is a random variable with values in $L^p(\mathbb{N} \times \mathbb{R})$. If $X_1, \ldots, X_n$ are independent, identically distributed copies of X, we may consider the *average landscape* $\overline{\Lambda(X)} \in L^p(\mathbb{N} \times \mathbb{R})$, which is defined as

$$\overline{\Lambda(X)}(k,t) = \frac{1}{n}\sum_{i=1}^{n} \lambda_k(X_i)(t).$$

The Central Limit Theorem implies that, for $p \geq 2$, if the expected values $E\|\Lambda(X)\|_p$ and $E\|\Lambda(X)\|_p^2$ are finite, then

$$\sqrt{n}\,[\,\overline{\Lambda(X)} - E(\Lambda(X))]$$

converges weakly to a Gaussian random variable.

As a consequence of this fact, if we define

$$Y = \|\Lambda(X)\|_1 = \int_{\mathbb{N}\times\mathbb{R}} \Lambda(X) = \sum_{k=1}^{\infty} \int_{-\infty}^{\infty} \lambda_k(X)(t)\,dt, \tag{13.1}$$

then Y has the property that $\sqrt{n}\,[\,\bar{Y} - E(Y)]$ converges to a normal distribution with zero mean. This allows us to use the estimator

$$z = \frac{\bar{Y}\sqrt{n}}{S_n}$$

for confidence intervals and hypothesis testing, where S_n^2 is the sample variance. Hence,

$$\bar{Y} \pm z_0\,\frac{S_n}{\sqrt{n}}$$

is a $(1-\alpha)$ confidence interval for $E(Y)$ if z_0 is the upper $\alpha/2$ critical value for a $N(0,1)$ distribution.

References: Further details can be found in [P. Bubenik, Statistical topological data analysis using persistence, *J. Machine Learning Res.* 16 (2015), 77–102]. Other estimators are described in [F. Chazal et al., Robust topological inference: distance to a measure and kernel distance, *J. Machine Learning Res.* 18 (2018), 1–40].