APPLIED TOPOLOGY LECTURE NOTES

CHAD GIUSTI

UPDATED: NOVEMBER 6, 2017

Topology of Data

When we move to the realm of science and engineering one of the fundamental changes from that of mathematics is that we only get to measure things, rather than assuming we know what we're studying – to put it another way, we move from topological spaces to a finite sample of points from those spaces. However, with a bit of cleverness, we can still use our topological tools to recover information about the underlying spaces in a sensible way.

Summary statistics for persistent homology

The bottleneck distance is provides a solid theoretical foundation for understanding persistence, as a notion of stability is absolutely necessary for any serious application: if small perturbations to the data created large changes in the persistence diagram, it would be impossible to interpret answers. However, the resulting geometry on the space of persistence diagrams is not as straightforward as we might like it to be.

Example.

Let $D_1 = \{(2,4),(2,6)\}$ and $D_2 = \{(1,5),(3,5)\}$. There are two partial matchings between D_1 and D_2 which induce the bottleneck distance: matching the cycle at $(2,4)$ to that at $(1,5)$, and $(2,6)$ to $(3,5)$ gives the correct bottleneck distance of 1, as does matching $(2,4)$ to $(3,5)$ and $(2,6)$ to $(1,5)$. We could use either matching to give a sensible notion of the "mean" of these two diagrams[1], and those two choices disagree.

[1] In particular, the Fr'echet mean, which is a choice of a centroid point for each matched pair which minimizes the total variance.

If we can't even take means of diagrams, we are in trouble with regard to doing statistical analyses using persistent homology. One of the most promising ways to fix this problem is to embed the diagrams in a function space where such problems can be easily overcome. The down-side is that we move even further away from the persistence module, often losing most or all of the underlying topological structure. Nonetheless, the trade-off often provides discriminatory power that is useful in applications. We'll consider three such approaches.

Definition. Let $H_p(S)$ be the pth persistent homology of a filtered simplicial complex. The *pth Betti curve* for S is the piecewise constant function $\beta_p(t) : \mathbb{R} \to \mathbb{R}_{\geq 0}$ given by $\beta_p(t) = \beta_p(S_{t(i)})$ for $t \in (t(i), t(i+1)]$.

The Betti curves of a complex provide a portrait of the number of classes that are present at any given filtration, but throw away information on which cycles are persisting from filtration to filtration – for example, imagine a structure with a few very long lived cycles having the same Betti curves as one with very many short-lived cycles that don't overlap much. However, as an aggregate measure of "how much homology" is present, they provide a useful summary. Further, it is straightforward to compute means (pointwise on the t-axis), norms (standard integral norms), variance/standard deviation and other statistics of curves. If we don't want to throw away lifetime information for individual cycles, we can use a richer embedding.

Definition. Let $H_p(S)$ be the pth persistent homology of a filtered simplicial complex, $GC(S)$ a collection of generating cycles for S. Transform the birth-death coordinates $(b_{[x_a]}, d_{[x_a]})$ for each $[x_a] \in GC(S)$ into $(m_{[x_a]}, h_{[x_a]}) = (\frac{1}{2}(d_{[x_a]} + b_{[x_a]}), \frac{1}{2}(d_{[x_a]} - b_{[x_a]}))$ coordinates, and define $R_{[x_a]} \subseteq \mathbb{R}^2$ to be the triangular region bounded by the m-axis, the slope 1 line segment $(b_{[x_a]}, 0)$ to $(m_{[x_a]}, h_{[x_a]})$, and the slope -1 line segment $(d_{[x_a]}, 0)$ to $(m_{[x_a]}, h_{[x_a]})$. Define a function $r(m,h) = |\{[x_a] \mid (m,h) \in R_{[x_a]}\}|$ which counts how many regions R a point lies inside, and define a sequence of functions $\lambda_k(m) = \sup_h\{r(m,h) \leq k\}$, which record the "top boundary" of the region of the plane with $r(m,h) \leq k$. The *pth persistence landscape* for S is $\Lambda(S) = \{\lambda_k(m) \mid k = 1, 2, \dots\}$.

Since only finitely many regions exist for any choice of S, the λ_k are uniformly zero for $k >> 0$. Thus, we can assume that for any finite family of diagrams there is a maximum k and truncate the sequence of functions at that value. Now, as with the Betti curves, we can compute statistics on each of the individual functions λ_k pointwise as a function of m, or using statistical measures build specifically for functions.

Landscapes retain a great deal more information about the lifetimes of individual cycles at the price of being much larger and potentially computationally cumbersome. They also fail to have a fixed size: the maximum k might vary from data set to data set, which may in turn complicate input into machine learning tools, etc. The following is a potential remedy.

Definition. Let $H_p(S)$ be the pth persistent homology of a filtered simplicial complex, $GC(S)$ its persistence diagram. Transform the birth-death coordinates $(b_{[x_a]}, d_{[x_a]})$ for each $[x_a] \in GC(S)$ into $(b_{[x_a]}, \ell_{[x_a]}) =$

$(b_{[x_a]}, (d_{[x_a]} - b_{[x_a]}))$ coordinates, and let $g(b, \ell)$ be a differentiable *kernel* function with $\int_{\mathbb{R}^2} g = 1$ and mean $(0, 0)$. The *persistence surface* for D and g is $\rho_D(b, \ell) = \frac{1}{|GC(S)|} \sum_{[x_a] \in GC(S)} g(b - b_{[x_a]}, \ell - \ell_{[x_a]})$; with mild abuse of notation, this is just the convolution $D * g$. Given discretization parameters Δb and $\Delta \ell$, and a rectangular region $R = (B, B + k\Delta b) \times (L, L + m\Delta \ell)$ of the (b, ℓ)-plane, the *persistence image* of D and g on $(R, \Delta x)$ is the $(k \times m)$ matrix $PI(S)$ with

$$PI(S)_{i,j} = \int_{B+(i-1)\Delta b}^{B+i\Delta b} \int_{L+(j-1)\Delta \ell}^{L+j\Delta \ell} \rho_D d\ell db.$$

Thus, we divide the region of interest into rectangular 'pixels' and take the total value of the persistence surface on each pixel as a proxy for the number of "nearby" points in the persistence diagram. For a fixed choice of region and discretization parameters, this approach allows the comparison of diagrams with arbitrary numbers of points as a vector of fixed size, which is then appropriate input for many machine learning algorithms.