APPLIED TOPOLOGY LECTURE NOTES

CHAD GIUSTI

UPDATED: OCTOBER 13, 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.

Clustering, dendrograms and Mapper

The concept of *data* is much subtler than one might assume from first impressions. This course is, at its core, about uncovering some of those subtleties, and so we'd better start by centering ourselves a bit. Since we don't have time to dive deep into topics from probability and statistics, we'll have to skirt certain important details. However, we won't let that scare us away from thinking big.

Definition. Let X be a topological space. A *point cloud* in X is a finite collection P of (possibly non-distinct) points $(x_1, \ldots x_n) \in X^n$. A *proximity measure* on X is a real-valued function $\mu : X \times X \to \mathbb{R}_{\geq 0}$, and the *$\mu$-proximity matrix* for P is $(M(P)_{i,j}) = \mu(x_i, x_j)$.

Proximity measures usually come in one of two types: *dissimilarity measures*, which take large values when two objects are dissimilar, as with distances, and *similarity* measures, which behave inversely, as with correlations.

The most common context in applications is for X to be some Euclidean space $\mathbb{R}^d$, or some subspace thereof, μ to be the Euclidean metric, a dissimilarity measure, on that space. This usually corresponds to some collection of d measurements taken independently or as a whole for several objects – spatial coordinates, physical properties, sale prices, time stamps, etc. It is important to note that the choice of μ as the Euclidean distance is not always sensible, particularly when we have several different types of unrelated data being by the coordinates. For example, we can think of vectors in $\mathbb{R}^d$ as being time series, in which case pointwise dissimilarity may not indicate overall differences – consider sampling $\sin(x)$ and $\cos(x)$ at a range of x values, for example. In this case, it is often saner to consider a similarity measure like correlation or coherence.

One of the most common problems in scientific and engineering applications is as follows.

Problem. Let A be a topological space, $S = (s_1, \ldots, s_n)$ a point cloud in A of *samples* or *objects*, (X, μ) a topological *measurement space* equipped with a proximity measure, and L a topological space of *labels*. Suppose there is a continuous map $\ell : A \to L$, and say the *label* of $a \in A$ is $\ell(a)$, and a relation $r : A \to X$ of *measurements*[1]. Given some assumptions on r and the point cloud $r(S)$ (or the proximity matrix $M(r(S))$, how much of ℓ can we recover?

We tend to believe that r is a continuous function, but we also usually assume that measurements are noisy – that we add some small, random amount of error to the measurement. Thus r isn't even a function, since repeated measurements of the same object will give different answers. However, in well-behaved cases it is "nearly continuous", so we can try to make reasonable inferences in spite of things going a little wrong.

Let $A = \coprod_{i=1}^{k} A_i$ for some path-connected topological spaces A_i and $L = \pi_0(A)$ with the natural map $\ell : A \to L$ given by $\ell(a) = [a]$. The problem of recovering $\hat{\ell} \approx \ell$ from $r(S)$ is called *clustering*.

Suppose $r : A \to \mathbb{R}^d$ is also constant on each component, so $r(a) = r_i \in \mathbb{R}^d$ for all $a \in A_i$, but that we add noise to each evaluation. If $\mu(r_i, r_j)$ is much larger/smaller than the magnitude of the noise, it is usually possible to recover the distinction between ℓ_i and ℓ_j using standard techninques like k-means or linkage clustering, support vector machines, etc. For concreteness, let's work with (single-)linkage clustering under a dissimilarity measure, where $\hat{\ell}(r(s_i)) = \hat{\ell}(r(s_j))$ if $\mu(r(s_i), r(s_j)) \leq t$ for some $t \in \mathbb{R}_{\geq 0}$, extended by transitivity. These techniques extend nicely to $r(a)$ being non-constant, so long as the images of the components are well-separated by μ; in this case, kernel-based or deep-learning methods are often employed to overcome the potential geometric complexity introduced by r.

However, as separation decreases or happens across multiple scales, it often becomes less clear when to give two points distinct labels. One common way to deal with this is *hierarchical clustering*, where we consider the outcome of our chosen clustering algorithm as we vary the criterion for when to assign distinct labels to different subsets. Thus, we obtain a continuous sequence of imputed labelings of points, $\hat{\ell}_t$. If we draw the codomain of $\hat{\ell}_t$ as a function of t, we obtain a tree called a *dendrogram* or *merge tree*,

[1] By abuse, we write r for the induced map on point clouds, as well.

which describes the proximity (or other parameter) at which the number of connected components changes. Each component of the level set of the tree at a given proximity corresponds to a cluster – this is the point cloud version of a Reeb graph for a space Q with a function $f : Q \to \mathbb{R}_{\geq 0}$ for which the level sets are $f^{-1}(t) = \bigcup_{i=1}^{m} \{x \in X \mid \min(\mu(r(s_i), x), \mu(x, r(s_i))) < t\}$.

Restricting ourselves to a discrete label space provides a lot of structure, and thus provides powerful tools, but is extremely limiting. If we replace our label space by $L = \mathbb{R}^d$ – suppose that each object is labeled by its weight or position, for example – a range of techniques like principal component analysis (for maps where a linear $\hat{\ell}$ make sense) and multidimensional scaling (where $\hat{\ell}$ can be non-linear) are employed. In exchange, we have to assume certain things about X, usually that it is $\mathbb{R}^D$, $D >> d$, with μ the standard Euclidean distance, but we can get a very good recovery of the lower-dimensional labels if this is all true, however these come at the cost of distortion of distances and sometimes other feature loss.