UMAP – background

UMAP is short for *Uniform Manifold Approximation and Projection* and is a dimension reduction algorithm. The algorithm builds on three assumptions: the data under consideration is uniformly distributed on a Riemannian manifold[[1]](#footnote-1), this manifold is locally connected and its metric is locally constant. The aim is to reconstruct (or: model) this manifold and derive a lower dimensional projection which is as closely equivalent as possible.

As a first step, an open ball of fixed radius is positioned around each datapoint. The union of these balls should be an open cover of the manifold. In general, this will not be the case, since in reality, data will not look uniformly distributed. Thus, some parts will not be covered while other parts may be overly covered (by too many balls). Since we assumed the data to be uniformly distributed on the manifold, this means that the notion of distance is not constant on the manifold. Instead of fixing the radius, a unit ball will therefore be defined as stretching to the k-th nearest neighbor, where k is a hyper parameter (set by the user). The algorithm is refined by working in a fuzzy topology, meaning that being in a set is no longer binary (yes or no), but this is described as a fuzzy value between zero and one. To guarantee local connectedness (no point being completely isolated), the fuzzy confidence decays only after the first nearest neighbor.

Next, the constructed open cover of the manifold can be translated into a simplicial complex. To understand what this is, first consider the a k-simplex. A k-simplex is a convex hull of k+1 independent points. For example, a 0-simplex is a point, a 1-simplex is a line and a 2-simplex is a triangle. A simplicial complex is the result of gluing k-simplices together along their faces (their corners and sides). Now, for each set in the cover, a 0-simplex is created; when the intersection between two sets is non-empty, a 1-simplex between the two is created; when the triple-intersection between three sets is non-empty, a 2-simplex between them is created; and so on. If one only allows 0- and 1-simplices, one obtains a graph with weighted edges, where each weight represents the probability of the edge existing.

The lower dimensional representation will be constructed on a Euclidian manifold, in our case a two-dimensional one. Constructing it is considered an optimalisation problem: minimize the cross entropy between the high dimensional manifold with a fuzzy topological structure () and the low dimensional Euclidian manifold (). Cross entropy is defined as follows

Where is the set of edges and is the weight of edge . The first part represents an attractive force that is large when is large and which is minimized by being as large as possible (meaning the points are close to each other); the second part represents a repellent force that is large when is small and which is minimized by being as small as possible (the points being far away from each other). Solving this problem, gives a low-dimensional embedding of the manifold.

The distance as measured on the high dimensional manifold can be any type of distance. In our analysis, we used Mahalanobis distance, which is a multi-dimensional generalization of measuring the amount of standard deviations between a point and a distribution mean.

ROUGH TEXT

This will be done as follows. Around each datapoint an open ball is positioned.

As a first step, this manifold is modeled as a fuzzy topological structure. Then, a lower dimensional projection is derived as being the optimal solution to a minimization problem of cross entropy.

The dimension reduction is considered an optimalisation problem that minimizes the cross entropy between the reconstructed manifold and a lower dimensional projection.

UMAP is a dimension reduction algorithm. It assumes data is drawn from an underlying topological space, more specifically a manifold (which is a topological space wherein each point has a neighborhood that is homeomorphic to an open subset of Euclidian space).

We want to reconstruct the manifold and do this as follows: we assume the data is uniformly distributed along the manifold. We place a ball around each datapoint. If we do this in Euclidean space in Euclidian distance, we will likely find the data to be not uniformly distributed: some parts are not covered and some parts are overly covered (by a lot of balls). This means, the noting of distance on the manifold was wrong -> it is not constant. We will define a unit ball as stretching to the k-th nearest neighbour (where k is a hyper parameter of the algorithm). We now improve by:

* Being in a set or not is no longer considered binary (yes or no), but described as a fuzzy value between zero and one.
* To capture connectedness of the manifold (not as a whole, but any point has a connect neighborhood) -> no point should be completely alone -> the fuzzy confidence decays in terms of distance beyond the first nearest neighbor.

For the next step, we need some mathematical concepts: a k-simplex is a convex hull of k+1 independent points. For example, a 0-simplex is a point and a 1-simplex is a line, which has two faces: its two endpoints and a 2-simplex is a triangle with 3 1-simplex faces and 3 0-simplex faces. A simplicial complex is the result of gluing k-simplexes together along faces.

The open cover that is constructed can be translated into a simplicial complex: let each set in the cover be a 0-simplex; create a 1-simplex between to such sets if they have a non-empty intersection; create a 2-simplex between three such if the tripe intersection of all three is non-empty and so on.

If we would only allow for 0- and 1-simplices, we basically have a graph with vertices and edges and these edges have weights which can be interpreted as the probability of the edge existing.

Constructing a lower dimensional representation is considered an optimalisation problem: minimize the cross entropy between the manifold just constructed and a lower dimensional Euclidian manifold (here: a 2D plane). The cross entropy is defined as follows:

Where the first part represents an attractive force when w\_h is high and a repellend force when w\_h is low.

The distance on the manifold can be any type of distance. Here we used mahalanobis distance, which is a multidimensional generalization of measuring the amount of standard deviations you are away from a certain distribution mean.

Now, a Cech complex is used to convert the open cover in a simplicial complex.

Manifold = a topological space that locally resembles Euclidean space near each point -> each point has a neighborhood that is homeomorphic to an open subset ofEuclidean space of dimension n.

K-simplex = convex hull of k+1 independent points. Simplicial complex = gluing k-simplexes together along faces.

Open cover = family of sets whose union is the whole space

Cech complex = combinatorial way to covert an open cover in a simplicial complex: let each set in the cover be a 0-simplex; create a 1-simplex (line) between two such sets if they have a non-empty intersection; create a 2-simplex between three such sets if the triple intersection of all three is non-empty; and so on.

We want to learn about the underlying topological space we assume the data is drawn from. Place a ball of fixed radius around each data point. We assume the union of balls to be an open cover. Then construct the simplicial complex. Instead of using all levels of k-simplexes, one can also only use 0 and 1 simplices -> Vietoris-Rips complex = computationally easier.

In reality: lack of coverage of the manifold -> disconnected components; also because irregular at some points the cover covers “too much” -> high order simplices. (data is not uniformly distributed across the manifold).

Turn it around: assume the data is uniformly distributed on the manifold. If the data looks like it is not uniformly distributed, that must be because the notion of distance is varying across the manifold -> a unit ball about a point stretches to the k-th nearest neighbor of the point. We take these balls: we get open balls of raius one with a locally varying metric. 10 works for most datasets

Instead of being in the open set as a binary yes no, we assume a fuzzy value between zero and one. Prolbem: we get a lot of isolated points. Hence, we failed to capture the connectedness of the manifold: at any point on the manifold there is some sufficiently small neighborhood of the point that is connected -> in our case: no point should be completely isolated. Thus we take the fuzzy confidence decay in terms of distance beyond the first nearest neighbor.

Now we have incompatible weights (the distance between a and b might not be the same for a as for b). We take a + b – a \* b (probability that an edge exists -> probability that at least of the edges exists).

Finding a low dimensional representation:

Now the manifold is Euclidian space (in low dimension, 2D in our case) and distance is standard Euclidian distance. How to determine closeness of the match between the fuzzy topological structures? Weight of simplices = probability of the simplex existing (Bernoulli). We will minimize the cross entropy:

First term: attractive force between the points e spans whenever there is a large weight associated to the high dimensional case. Will be minimizes when is large as possible (when the distance between them is small).

Second term: repulsive force between the ends of e whenever is small. Will be minimized by making as small as possible.

Steps:

1. Constructing a fuzzy topological representation
2. Optimizing the low dimensional representation to have as close a fuzzy topological representation as possible as measured by cross entropy

We used mahalanobis distance: is a measure of the distance between a point P and a distribution D -> multidimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D.

1. A manifold is a topological space wherein each point has a neighborhood that is homeomorphic to an open subset of Euclidian space. A Riemannian manifold is a real, smooth manifold with a positive-definite inner product on the tangent space at each point. [↑](#footnote-ref-1)