Computational topology: Image classification using Vietoris-Rips complex

Due on Friday, May 13, 2016

Neža Mramor Kosta, Gregor Jerše

Andrej Dolenc, Peter Us, Rok Ivanšek

Contents

Obtaining and preprocessing data

```python
class Scaler(BaseEstimator, TransformerMixin):
    """ Class used for scaling the dataset x.
        Removes the mean and than scales to values from [-1, 1].
        Implements inverse_transform method.

        Args:
        axis: if 1, standardize samples (rows)
              if 0, standardize features (columns)
    """

    def __init__(self, axis=1):
        assert axis in [0,1]
        self.axis = axis

    def fit(self, x):
        shape = (len(x), 1) if self.axis == 1 else (1, len(x[0]))
        self.means = np.mean(x, self.axis).reshape(shape)
        self.maxs = np.max(np.abs(x - self.means), self.axis).reshape(shape)

    def transform(self, x):
        return np.nan_to_num((x - self.means)/self.maxs)

    def fit_transform(self, x):
        self.fit(x)
        return self.transform(x)

    def inverse_transform(self, x):
        return (x*self.maxs) + self.means
```

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

The classification model

Definition 1. *[Vietoris-Rips complex] Let X be a a set of m-dimensional points $X \in \mathbb{R}^m$ and let d be a metric. Pick a parameter $r > 0$. Construct a simplicial complex as follows:*

- *Add a 0-simplex for each point in X.*
- *For $x_1, x_2 \in X$ add a 1-simplex between x_1, x_2 if $d(x_1, x_2) \leq r$.*
- *For $x_1, x_2, x_3 \in X$ add a 2-simplex with vertices x_1, x_2, x_3 if $d(x_1, x_2), d(x_1, x_3), d(x_2, x_3) \leq r$.*
- *...*

The classification model continued on next page...

- *For $x_1, x_2, ..., x_m \in X$, add a $(m-1)$-simplex with vertices $x_1, x_2, ..., x_m$ if $d(x_i, x_j) \leq r$ for $0 \leq i, j \leq m$; that is, if all the points are within a distance of r from each other.*

The simplicial complex is called the Vietoris-Rips complex and is denoted $V_r(X)$.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Relation to single linkage clustering algorithm

Our intuition tells us, that the model we build using the Vietoris-Rips complex to classify the images, produces the same results as the well known single linkage clustering algorithm. In this section we aim to prove or at least give a strong intuition that this is indeed the case.

Definition 2. *A connected component in a graph $G(V, E)$ (V is a set of vertices, E a set of edges) is a set of vertices C such that between every two points $a, b \in C$ there exists a path in the graph G between them.*

Both algorithms take a set X of n samples with m features as input. We can think of a sample x in the set X as a point in the m-dimensional space $x \in \mathbb{R}^m$. The algortihms then constructs a graph with points X as the vertices (V) and edges E. The k connected components in the constructed graph corespond to the classes of samples.

Single linkage algorithm. The algorithm starts with n connected components (no edges in the graph). n each step the algorithm chooses the two connected components that are closest to each according to some distance metric d (in our case the euclidean distance) and joins them into one by adding an edge between their closest two vertices.

Definition 3. *The distance D between two connected components A and B is defined as the distance of the pair of vertices (one from A and one from B) that are closest to each other. More formally*

$$D(A, B) = \min_{a \in A, b \in B} d(a, b).$$

The algorithm stops when there are only k connected components left.

Vietoris-Rips classification algorithm. The algorithm builds a (1-dimensional) Vietoris-Rips complex $V_r(X)$ with parameter r. We choose the biggest r such that the Vietoris-Rips complex $V_r(X)$ has k connected components.

To prove that the two algorithms indeed produce the same connected components we will first prove the next claim.

Claim 1. *Let $G_{sl}(V_{sl}, E_{sl})$ be a graph produced by the single linkage algorithm for finding k clusters and let d_{max} denote the distance between vertices in graph G_{sl} that were connected in the last iteration of the algorithm. Graph G_{sl} has connected components $A_1...A_k$. The graph $G_{vr}(V_{vr}, E_{vr})$ induced by the Vietoris-Rips complex $V_{d_{max}}(V_{sl})$ has the same connected components $A_1...A_k$.*

Proof. To prove that we have the exact same connected components in G_{vr} as in G_{sl} the following must be true for G_{vr}:

1. The vertices inside A_i are connected i.e. there exists a path between arbitrary two vertices $a, b \in A_i$, for every $i \in 1...k$.

2. The sets of vertices $A_1, ..., A_k$ are disconnected i.e. there isn't an edges $e(a, b)$ between a pair of two points (a, b) such that $a \in A_i, b \in A_j, i \neq j$.

From the way the algorithm for single linkage works (it is adding edges of increasing length), we can easily see that d_{max} is the largest edge in the graph G_{sl}. From this and the definition of Vietoris-Rips (Definition 1) it follows that the set of edges in G_{sl} is a subset of edges in G_{vr} i.e. $E_{sl} \subseteq E_{vr}$. This means that all the edges found in G_{sl} will also appear in G_{vr} therefore the vertices inside individual subsets A_i will indeed be connected.
We prove the second part by contradiction. By definition of Vietoris-Rips complex $V_{d_{max}}(V_{sl})$, no edge $e \in E_{vr}$ will be greater than d_{max}. Let's now assume that there exists an edge e with length d_e between two vertices a, b that are in separate subsets $a \in A_i$ and $b \in A_j$, $i \neq j$. Now we have two options. First one $d_e < d_{max}$. If this was the case than d_e would necessarily have to connect two points from the same subset A_i. This follows from the fact that single linkage algorithm always connects two of the closest connected components so any edge with a length smaller than d_{max} is an edge inside one of the subsets $A_1, ...A_k$. TODO: further explanation needed for the first case??? The second option is that $d_e > d_{max}$. This also immediately leads to a contradiction, this time with the Vietoris-Rips definition. $\square$

Using Claim 1 we see that the connected components in G_{vr} and G_{sl} are indeed the same. We need to take into account that the Vietrois-Rips algorithm takes the biggest such r so that the graph has k connected components, so $r > d_{max}$. But we can quickly see that the extra edges in the graph induced by $V_r(V_{sl})$ will not change the connected components. After all we allready have k connected components in G_{vr}. To join two together would mean a violation of a fundemental rule of the algorithm.

Testing

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

References

[1] Leslie Lamport, *LaTeX: a document preparation system*, Addison Wesley, Massachusetts, 2nd edition, 1994.