Maximum Likelihood Estimation of Intrinsic Dimension

December 21, 2017

1 Introduction

In class and tutorial we discussed several dimension reduction algorithms. Yet, we mostly assumed that the dimension which we wish to reduce the data to, denoted by d, is given. Except for the PCA algorithm, for which we formulated a well defined criterion, albeit heuristic, we did not discus *how to choose d*. There is a good reason for this - there is yet no clear cut answer for this question.

In this document we follow [1] which suggests an (empirical) estimator for the intrinsic dimension of the data. We stress that the approach shown here is a possible approach; this question is an open one.

2 Algorithm

We first suggest to read the paper, it is well written and easy to follow. However, this is not essential in order to implement the algorithm. The output of the algorithm is an (empirical) estimator for the intrinsic dimension of the data, $\hat{m}$; intuitively, as discussed in class and tutorial, though the data may be embedded in a high dimension, it is often the case that the actual manifold from which the data is sampled is from lower dimension.

Let $\mathcal{D} = \{x_i\}_{i=1}^n$ be n data points, with $x \in \mathbb{R}^D$. Define $T_k(x)$ as the Euclidean distance of the point x from its k-th nearest neighbor(NN), x_k, in the sample,

$$T_k(x) = ||x - x_k||_2,$$

where $||\cdot||_2$ is the L_2 norm. A possible way to calculate $T_k(x_i)$ is to first calculate the distance matrix, D, with $D_{ij} = ||x_i - x_j||_2$, and then search for the k-th smallest value in the i-th row is the value of $T_k(x_i)$. Since you will be asked to calculate $T_k(x_i)$ for several k values, this way might be beneficial; it allows to calculate for few values k, $T_k(x_i)$ with ease.

The algorithm is given in the following and requires two hyper-parameters k_1 and k_2. In the formulation of the algorithm, we use an improved estimator for $\hat{m}_k$, given $\hat{m}_k(x_i)$. We follow http://www.inference.org.uk/mackay/dimension/.

Algorithm 1 Maximum Likelihood Estimation of Intrinsic Dimension

Initialize:
Sampled data, $\mathcal{D} = \{x_i\}_{i=1}^n$ and hyper-parameters $k_1, k_2 \in \mathbb{N}$.
Calculate $\forall x_i \in \mathcal{D}, \forall k \in [k_1, k_2]$, $T_k(x_i)$
Calculate $\forall x_i \in \mathcal{D}, \forall k \in [k_1, k_2]$, $\hat{m_k}^{-1}(x_i) = \frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{T_k(x_i)}{T_j(x_i)}$
Calculate $\forall k \in [k_1, k_2]$, $\hat{m_k}^{-1} = \frac{1}{n} \sum_{i=1}^{n} \hat{m_k}^{-1}(x_i)$
Calculate $\hat{m} = \frac{1}{k_2 - k_1 + 1} \sum_{k=k_1}^{k_2} \hat{m_k}$
return $\hat{m}$

References

[1] Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.