SLT coding exercise #1

Locally Linear Embedding

https://gitlab.vis.ethz.ch/vwegmayr/slt-coding-exercises

Due on Monday, March 6th, 2017

Nicolas Känzig
12-916-615

Contents

The Model **3**

The Questions **4**
(a) Get the data 4
(b) Locally linear embedding 4
(c) Cluster structure 4
(d) Nearest Neighbors 4
(e) Linear manifold interpolation 4
(b) Locally linear embedding 5
(c) Cluster structure 5
(d) Nearest Neighbors 5
(e) Linear manifold interpolation 5

The Implementation **7**

Your Page **8**

The Model

The model section is intended to allow you to recapitulate the essential ingredients used in Locally Linear Embedding. Write down the *necessary* equations to specify Locally Linear Embedding and and shortly explain the variables that are involved. This section should only introduce the equations, their solution should be outlined in the implementation section.
Hard limit: One page

Reconstruction Error

$$\mathcal{E}(W) = \sum_i |\vec{X}_i - \sum_j W_{ij}\vec{X}_j|^2 \tag{1}$$

$\vec{X}_i$ = Data Points, real-valued with dimensionality D
W = Weight Matrix. W_{ij} summarizes the contribution of the jth data point to the ith reconstruction

Embedding Error

$$\Phi(Y) = \sum_i |\vec{Y}_i - \sum_j W_{ij}\vec{Y}_j|^2 \tag{2}$$

$\vec{Y}_i$ = Embedded representation of $\vec{X}_i$, real-valued with dimensionality d

Constraints:

$$\sum_j W_{ij} = 1 \tag{3}$$

$$W_{ij} = 0 \text{ if } \vec{X}_j \text{ and } \vec{X}_i \text{ do not belong to the same set} \tag{4}$$

The Questions

This is the core section of your report, which contains the tasks for this exercise and your respective solutions. Make sure you present your results in an illustrative way by making use of graphics, plots, tables, etc. so that a reader can understand the results with a single glance. Check that your graphics have enough resolution or are vector graphics. Consider the use of GIFs when appropriate.
Hard limit: Two pages

(a) Get the data

For this exercise we will work with the MNIST data set. In order to learn more about it and download it, go to http://yann.lecun.com/exdb/mnist/.

(b) Locally linear embedding

Implement the LLE algorithm and apply it to the MNIST data set. Provide descriptive visualizations for 2D & 3D embedding spaces. Is it possible to see clusters?

(c) Cluster structure

Investigate the cluster structure of the data. Can you observe block structures in the M matrix (use matrix plots)? Also plot the singular values of M. Do you notice something? Can you think of ways to determine the optimal embedding dimension?

(d) Nearest Neighbors

Investigate the influence of the choice of how many nearest neighbors you take into account. Additionally, try different metrics to find the nearest neighbors (we are dealing with images!).

(e) Linear manifold interpolation

Assume you pick some point in the embedding space. How can you map it back to the original (high dimensional) space? Investigate how well this works for points within and outside the manifold (does it depend on the dimensionality of the embedding space?) Try things like linearly interpolating between two embedding vectors and plot the sequence of images along that line. What happens if you do that in the original space?

The Questions continued on next page...

(b) Locally linear embedding

The LLE algorithm was applied on a subset of a thousand images from the provided training set. Fig. 1 shows the calculated embeddings into 2-dimensional and 3-dimensional spaces using k=10 neighbors for each data point. Clearly a clustured structure can be observed, however not all of them are well separable.

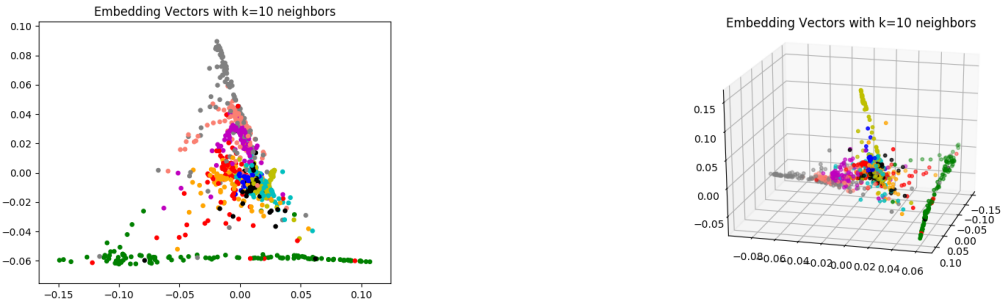

Figure 1: Directionality measurements

(c) Cluster structure

When plotting the M Matrix a strongly diagonal structure can be observed. As proven in Series 1 (Problem 1 - (4)), the embedding error is minimal when the embedded vectors are associated with the eigenvectors corresponding to the smallest d+1 eigenvalues of M. So for an optimal embedding dimension d, one has to ensure that the Matrix M provides at least d+1 small eigenvalues. The closer these values are to zero, the better.

(d) Nearest Neighbors

The influence of the number of the nearest neighbors that are added to a set was analyzed by comparing the embeddings calculated using different values. The embedded vectors for the values k = [3, 20, 40] are depicted in Figures 2-4. An Euclidean distance-metric has been used for finding the k nearest neighbors. It can be seen that when k is chosen too big the cluster structure gets lost.

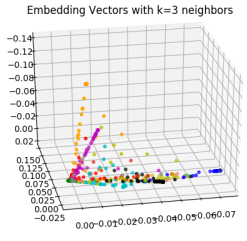

Figure 2: k=3

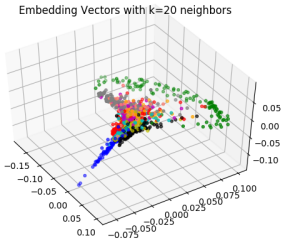

Figure 3: k=20

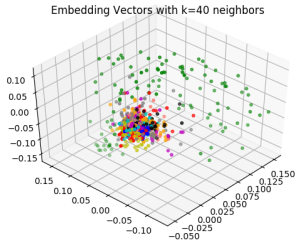

Figure 4: k=40

(e) Linear manifold interpolation

To reconstruct an image from the embedding space first one has to find the k nearest neighbors of the embedded image to be reconstructed. Second the weights that minimze the reconstruction error in the

embedding space must be computed. Lastly the image can be reconstructed by adding up the corresponding k neighbors in the original space using the calculated weigthts.
The higher the dimension of the embedded space the better the result, as less information gets lost during the dimensionality reduction.
Choosing a point that lies outside of the manifold leads to a less clear result, as the reconstruction error gets increased. If the point is chosen too far from a legit cluster, the reconstruction will lead to no result. Fig. 5 shows the reconstructions using two different values for k. When choosing k too big the result is much worse which matches our observations in part **(a)**.

(a) d=3 , k=10 (b) d=3 , k=40

Figure 5: Reconstructions

The Implementation

In the implementation section you give a concise insight to the practical aspects of this coding exercise. It mainly mentions the optimization methods used to solve the model equations. Did you encounter numerical or efficiency problems? If yes, how did you solve them? Provide the link to your git branch of this coding exercise.
Hard limit: One page

For implementing the LLE algorithm, methods provided by sklearn and numpy libraries have been used. The complexity of LLE increases quadratically (Partial eigenvalue decomposition) with the number of training data points. So to limit the computational effort, instead of using the whole provided training set containing 60k elements, a 1k subset was used for the embeddings.

Your Page

Your page gives you space to include ideas, observations and results which do not fall into the categories provided by us. You can also use it as an appendix to include things which did not have space in the other sections.
No page limit.

Your Answer
12-916-615/1_locally_linear_embedding