Lecture 9 Oct. 4 -2006

Data Visualization

STAT 442 / 890, CM 462

Lecture: Ali Ghodsi

1 LLE continued...

Previously, we minimized the following:

$$\min_{w} \sum_{i=1}^{n} ||x_i - \sum_{j=1}^{k} w_{ij} x_{N_i}||^2$$

such that $\sum_{i=1}^{k} w_{ij} = 1$. Now the other cost function for LLE must be minimized.

$$\min_{Y} \sum_{i=1}^{n} ||y_i - \sum_{j=1}^{k} w_{ij} y_j||^2$$

In this case there is no unique solution. We must add constraints in order to make the optimization well posed.

Note that translating the output y_i by a constant will not affect the cost function. This translational degree of freedom can be removed by constraining the outputs to be centered on the origin:

$$\sum_{i=1}^{n} y_i = 0$$

This is not enough to ensure a unique solution.

The output y_i can also be rotated without affecting the cost. To remove this rotational degree of freedom, the outputs can be required to have unit covariance. This is done through the constraint:

$$\frac{1}{n} Y Y^T = I_{dXd}$$

Where $Y = [y_1 \dots y_n]$.

This constraint also assures that the different coordinates in the embedding space will be uncorrelated. In addition it does not allow the solution to collapse to the trivial zero solution.

Crucially this optimization can be done in closed form. To see this, define the following vectors:

$$I_i^T = [0, 0, \dots, 0, 1, 0, \dots, 0]$$

Where the 1 is in the ith position.

$$W_i^T = [0, 0, \dots, 0, w_1, \dots w_k, \dots, 0]$$

Where the $w_1 \dots w_k$ are weights of the k nearest neighbours of x_i.

From these definitions we can rewrite the cost function as:

$$E(Y) = \sum_{i=1}^{n} ||Y I_i - Y W_i||^2$$

then,

$$E(Y) = ||YI - YW||^2$$

where I is the identity matrix and $W = [W_1\ W_2\ W_3\ \ldots\ W_n]$. Then factor the Y out and:

$$E(Y) = ||Y(I - W)||^2$$

Now recall that for any matrix A with column A_i the following identity is true:

$$||A||^2 = ||A_i^T A_i||^2 = Tr(A^T A)$$

There Tr represents the trace of a matrix. So $Tr(A) = \sum_{i=1}^{n} a_{ii}$.

$$E(Y) = Tr((I - W)^T Y^T Y(I - W))$$

But it is known that $Tr(A) = Tr(A^T)$.

$$E(Y) = Tr(Y(I - W)(I - W)^T Y^T)$$

Denote $(I - W)(I - W)^T$ by M.

$$E(Y) = Tr(YMY^T)$$

To add the constraint we once again make use of the Lagrange multiplier.

$$E(Y) = Tr(YMY^T) + Tr(\Lambda(\frac{1}{n}YY^T - I))$$

Then the derivative is taken and the result is set equal to zero.

$$\frac{dL}{dY} = MY^T + \frac{1}{n}\Lambda Y^T = 0$$

By placing all of the constants into the new $\Lambda^* = -\frac{1}{n}\Lambda$.

$$MY^T = \Lambda^* Y^T$$

Which leads to the conclusion that the columns of Y^T are the eigenvectors of M. Since we are looking to minimize the cost function, we are looking for the eigenvectors corresponding to the smallest eigenvalues. An important note here is that M has at least one zero eigenvalue. corresponding to an eigenvector $e^T = [1, 1, \ldots, 1]$ which is just a column of ones. Discarding this eigenvector enforces the constraint that the outputs have zero mean. Therefore, for a d-dimensional solution one must ignore the first eigenvector (this is e and corresponds to the zero eigenvalue). The remaining d eigenvectors (the eigenvectors from $2 \ldots d + 1$) form the d-dimensional embedding.

A few weaknesses of the LLE algorithm.

1. No out of sample extension.

2. No estimate of dimensionality. In other words, there is no way of knowing if reducing the data to d dimensions is the optimal reduction.

   Recall how PCA solves this problem. In PCA, one way to estimate this dimensionality is to examine the eigenvalue spectra of covariance matrix. The intrinsic dimensionality is then the number of eigenvalues that are comparable in magnitude to the largest eigenvalue. However, this is not true for LLE and the eigenvalues from the third step of LLE are not reliable indicators of intrinsic dimensionality.

3. LLE ensures that close points in high-dimension remain close in low-dimension as well. But This is not true for points that are far apart. Thus, LLE can map faraway inputs to nearby outputs in the embedding space.