Lecture 7,8 Sept. 27,29 -2006

Data Visualization

STAT 442 / 890, CM 462

Lecture: Ali Ghodsi

0.1 Centering

In the derivation of the kernel PCA we assumed that $\Phi(X)$ has zero mean.

$$\sum_{i=0}^{n} \Phi(x_i) = 0$$

The problem is that we must ensure that this condition is satisfied without ever knowing the exact value of $\Phi(X)$. The following normalization of the kernel allows us to satisfy this condition.

$$\tilde{K}(x,y) = K(x,y) - E_x[K(x,y)] - E_y[K(x,y)] + E_x[E_y[K(x,y)]]$$

In order to prove that the above statement satisfies the condition we define:

$$\tilde{\Phi}(X) = \Phi(X) - E_x[\Phi(X)]$$

Notice that we have subtracted the mean from $\Phi(X)$ which means that the transformed $\tilde{\Phi}(X)$ has mean zero. The corresponding kernel can be written as:

$$\tilde{K}(x,y) = \tilde{\Phi}(x)\Phi\tilde{(}y)$$

This kernel can be expanded follows:

$$\tilde{K}(x, y) = (\Phi(x) - E_x[\Phi(x)]).(\Phi(y) - E_y[\Phi(y)])$$

$$= K(x, y) - E_x[K(x, y)] - E_y[K(x, y)] + E_x[E_y[K(x, y)]]$$

To perform Kernel PCA, one needs to replace all dot products $x^T y$ by $\tilde{K}(x, y)$ in Algorithm 2 (Table **??**). Note that V is the set of eigenvectors of $K(X, X)$ corresponding to the top d eigenvalues, and Σ is diagonal matrix of square roots of the top d eigenvalues.

Unfortunately Kernel PCA does not inherit all the strength of PCA. More specifically reconstruction of training and test data points is not a trivial practice in Kernel PCA. Algorithm 2 (Table **??**) shows that data can be reconstructed in feature space $\hat{\Phi}(x)$. However finding the corresponding pattern x is difficult and sometimes even impossible [1].

1 Locally Linear Embedding

Locally linear embedding (LLE) is another approach which address the problem of nonlinear dimensionality reduction (See Figure 1 for an example) by computing low-dimensional, neighbourhood preserving embedding of high-dimensional data. A data set of dimensionality n, which is assumed to lie on or near a smooth nonlinear manifold of dimensionality $d < n$, is mapped into a single global coordinate system of lower dimensionality, d. The global nonlinear structure is recovered by locally linear fits.

Consider t n-dimensional real-valued vectors x_i sampled from some underlying manifold. We can assume each data point and its neighbours lie on, or are close to, a locally linear patch

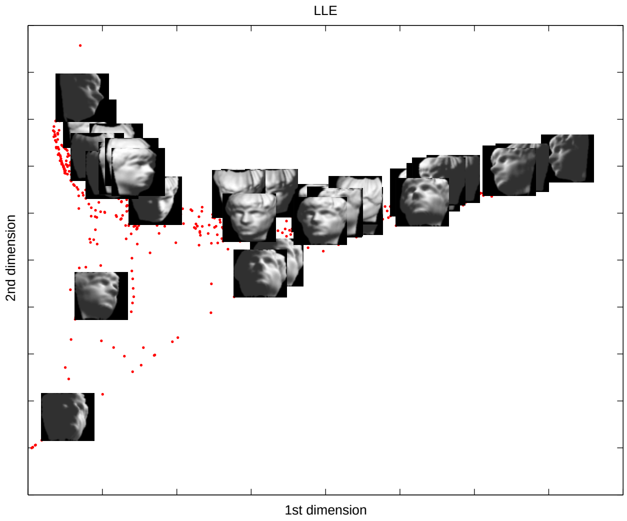

Figure 1: *LLE applied* $(k = 5)$ *to the same data set. A two-dimensional projection is shown, with a sample of the original input images.*

of the manifold. By a linear mapping, consisting of a translation, rotation, and rescaling, the high-dimensional coordinates of each neighbourhood can be mapped to global internal coordinates on the manifold. Thus, the nonlinear structure of the data can be identified through two linear steps: first, compute the locally linear patches, and second, compute the linear mapping to the coordinate system on the manifold.

The main goal here is to map the high-dimensional data points to the single global coordinate system of the manifold such that the relationships between neighbouring points are preserved. This proceeds in three steps:

1. Identify the neighbours of each data point x_i. This can be done by finding the k nearest neighbours, or by choosing all points within some fixed radius, ϵ.

2. Compute the weights that best linearly reconstruct x_i from its neighbours.

3. Find the low-dimensional embedding vector y_i which is best reconstructed by the weights determined in the previous step.

After finding the nearest neighbours in the first step, the second step must compute a local geometry for each locally linear patch. This geometry is characterized by linear coefficients that reconstruct each data point from its neighbours. The following cost function measures the reconstruction error.

$$\min_{w} \sum_{i=1}^{t} ||\mathbf{x}_i - \sum_{j=1}^{k} w_{ij}\mathbf{x}_{N_i(j)}||^2$$

where the neighbours of x_i are represented by $\mathbf{x}_{N_i(j)}$. i.e. $N_i(j)$ is the index of the jth neighbour of the ith point.

w_{ij} are computed to minimize this cost function subject to the constraint that the rows of the weight matrix sum to one, $\sum_{j=1}^{k} w_{ij} = 1$. The optimal weights W_{ij}, subject to this constraint, are found by solving a constrained least squares problem.

The i^{th} data point is reconstructed independently of all others, so the minimization can be done one term at a time. This is equivalent to solving a system of equations with n equations and K unknowns.

Ideally, the solution for the weights should be invariant under the linear mapping from the patch to the global coordinates on the manifold. This ensures that the reconstruction holds equally well in both high-dimensional and low-dimensional space. It can be shown that the weights are invariant to a transformation of the data point and its neighbours if and only if all rows of the weight matrix sum to 1.

$$\min_{w} \sum_{i=1}^{t} ||\mathbf{x}_i - \sum_{j=1}^{k} w_{ij}\mathbf{x}_{N_i(j)}||^2$$

such that for all i, $\sum_{j=1}^{k} w_{ij} = 1$ where $N_i(j)$ is the index of the jth neighbour of the ith point. For a given point x_i we have:

$$||\mathbf{x}_i - \sum_{j=1}^{k} w_{ij}\mathbf{x}_{N_i(j)}||^2$$

Define the matrix $N = \begin{bmatrix}\mathbf{x}_{N_i(1)} & \mathbf{x}_{N_i(2)} & \ldots & \mathbf{x}_{N_i(k)}\end{bmatrix}$ which is of dimension D by k where D is the original dimensionality of the data. Define also the matrix $W = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{pmatrix}$ which is a k by 1 matrix. The above statement can be rewritten as:

$$||\mathbf{x}_i - NW||^2$$

Another matrix $\chi = [x_i \; x_i \; x_i \; \ldots \; x_i]$ can be defined which is also of dimension D by k. The above statement can once again be rewritten as:

$$||\chi W - NW||^2$$

$$= ||(\chi - N)W||^2$$

$$= W^T(\chi - N)^T(\chi - N)W$$

The matrix $(\chi - N)^T(\chi - N)$ is called the Gram matrix and we can denote it by G.

The constraint $\sum_{j=1}^{k} w_{ij} = 1$ can also be written in matrix form as: $W^T e = 1$ where e is just a column of ones. The Lagrange Multipliers are then used to add the constraint to the function that needs to be minimised.

$$L(W) = W^T G W - \lambda(W^T e - 1)$$

Derivate with respect to W and obtain:

$$2GW - \lambda e = 0$$

$$GW = \frac{\lambda}{2} e$$

$$GW = ce$$

In practice we can set c arbitrarily to be equal to 1 and then solve for W. Once the solution for W has been found then W can be rescaled so that it sums to 1.

References

[1] B. Schölkopf, S. Mika, A. Smola, G. Rätsch, and K.-R. Müller. Kernel PCA pattern reconstruction *via* approximate pre-images. In L. Niklasson, M. Bodén, and T. Ziemke, editors, *Proceedings of the 8th International Conference on Artificial Neural Networks*, pages 147–152, 1998.