Locally Linear Embedding (Roweis & Saul, 2000)

Emin Orhan
eorhan@cns.nyu.edu

October 13, 2014

Problem statement: We have high-dimensional inputs $\mathbf{x}_i$. The "apparent" dimensionality of the input space is D. But we suspect that the data in fact live in a much smaller dimensional subspace. We assume that the "real" dimensionality of the data is rather d (with $d \ll D$). We want to find an embedding of the data from the full D-dimensional input space into a d-dimensional space such that the intrinsic geodesic structure of the data in the original space is preserved as much as possible.

Solution: The method proceeds in two steps. First, for each data point, we find a set of local weights that represent the point in a translation-, rotation- and scale-invariant way. This is done by solving the following optimization problem:

$$\text{minimize} \quad \sum_i \|\mathbf{x}_i - \sum_j W_{ij}\mathbf{x}_j\|^2 \tag{1}$$

$$\text{subject to} \quad W_{ij} = 0 \text{ if } j \notin \mathcal{N}(i) \text{ and } \sum_j W_{ij} = 1 \tag{2}$$

The first constraint says that we want to set the weight W_{ij} to zero if j is not a neighbor of i. The second constraint is required to make the solution invariant to translations of $\mathbf{x}$. The optimization problem above can be solved for each point separately (why?). We denote the point of interest by $\mathbf{x}$ (size: $D \times 1$), its K nearest neighbors by η (size: $D \times K$) and the weights by $\mathbf{w}$ (size: $K \times 1$). Now, the problem becomes:

$$\text{minimize} \quad (\mathbf{x} - \eta\mathbf{w})^T(\mathbf{x} - \eta\mathbf{w}) \tag{3}$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{1} = 1 \tag{4}$$

where we use $\mathbf{1}$ to denote a $K \times 1$ vector of 1's. Solving this problem is straightforward. We first write down the Lagrangian:

$$\mathcal{L} = (\mathbf{x} - \eta\mathbf{w})^T(\mathbf{x} - \eta\mathbf{w}) + \lambda(\mathbf{w}^T\mathbf{1} - 1) \tag{5}$$

and set its derivative with respect to $\mathbf{w}$ to zero:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = -2\eta^T(\mathbf{x} - \eta\mathbf{w}) + \lambda\mathbf{1} = 0 \tag{6}$$

$$\Rightarrow \quad \eta^T\eta\mathbf{w} = \eta^T\mathbf{x} - \frac{\lambda}{2}\mathbf{1} \tag{7}$$

$$\Rightarrow \quad \mathbf{w} = C^{-1}(\eta^T\mathbf{x} - \frac{\lambda}{2}\mathbf{1}) \tag{8}$$

where we denote $C = \eta^T\eta$. To find λ, we plug the solution $\mathbf{w}$ in the sum constraint and obtain:

$$\mathbf{1}^TC^{-1}(\eta^T\mathbf{x} - \frac{\lambda}{2}\mathbf{1}) = 1 \tag{9}$$

$$\Rightarrow \quad \frac{\lambda}{2} = \frac{\mathbf{1}^TC^{-1}\eta^T\mathbf{x} - 1}{\mathbf{1}^TC^{-1}\mathbf{1}} \tag{10}$$

It is easy to check that the solution $\mathbf{w}$ is invariant under translations, rotations and re-scalings of the inputs $\mathbf{x}$.

Mathematically, any transformation of the inputs that is of the form $\mathbf{x}' = s\mathbf{A}\mathbf{x} + \mathbf{b}$, where s is a scalar and $\mathbf{A}$ is an orthogonal matrix, leaves the optimal $\mathbf{w}$ unchanged. To see this, note that $\mathbf{x}' - \eta'\mathbf{w} = s\mathbf{A}\mathbf{x} + \mathbf{b} - (s\mathbf{A}\eta + \mathbf{b}\mathbf{1}^T)\mathbf{w} = s\mathbf{A}(\mathbf{x} - \eta\mathbf{w})$ where used the fact that $\mathbf{1}^T\mathbf{w} = 1$ due to the sum constraint. Thus, $(\mathbf{x}' - \eta'\mathbf{w})^T(\mathbf{x}' - \eta'\mathbf{w}) = s^2(\mathbf{x} - \eta\mathbf{w})^T(\mathbf{x} - \eta\mathbf{w})$ where we used the orthogonality of $\mathbf{A}$. We observe that the last expression is just the original objective scaled by the positive number s^2, hence the solution remains invariant to the transformation.

Now that we have found a translation-, rotation-, and scale-invariant representation of the inputs that reflects the intrinsic geometric relationships between them, we reason that the same representation must also hold in the lower-dimensional space as well. We thus use the same weights W_{ij} to find the coordinates $\mathbf{y}_i$ of the points in the lower-dimensional space. To do this, we solve the optimization problem:

$$\text{minimize} \quad \sum_i \|\mathbf{y}_i - \sum_j W_{ij}\mathbf{y}_j\|^2 \tag{11}$$

$$\text{subject to} \quad \sum_i \mathbf{y}_i = \mathbf{0} \text{ and } \sum_i \mathbf{y}_i\mathbf{y}_i^T = N\mathbf{I} \tag{12}$$

We will momentarily ignore the first constraint (i.e. $\sum_i \mathbf{y}_i = \mathbf{0}$), but will enforce it later once we have a solution to the optimization problem without the first constraint.

I find it easier to work with the matrix formulation of this optimization problem:

$$\text{minimize} \quad \text{Tr}\Big[(Y - YW)^T(Y - YW)\Big] \tag{13}$$

$$\text{subject to} \quad YY^T = N\mathbf{I} \tag{14}$$

The Lagrangian is given by:

$$\mathcal{L} = \text{Tr}\Big[(Y - YW)^T(Y - YW)\Big] + \Lambda(YY^T - N\mathbf{I}) \tag{15}$$

Here Y (size: $d \times N$) is a matrix whose columns are the coordinates of the data points in the lower-dimensional space, W (size: $N \times N$) is the weight matrix extended to all N points (note that this matrix will be sparse, because typically $K \ll N$) and Λ (size: $d \times d$) is a diagonal matrix of the Lagrange multipliers (note that there are d constraints in total). We then set the derivative of the Lagrangian with respect to Y to zero:

$$\frac{\partial \mathcal{L}}{\partial Y} = 2Y(I - W)(I - W)^T + 2\Lambda Y = 0 \tag{16}$$

$$\Rightarrow \quad Y(I - W)(I - W)^T = -\Lambda Y \tag{17}$$

Thus Y is a matrix of d eigenvectors of $(I - W)(I - W)^T$ and $-\Lambda$ contains the corresponding set of eigenvalues. But which eigenvectors should we pick? Note that from Equation 17, we have:

$$Y(I - W)(I - W)^T Y^T = -\Lambda \tag{18}$$

$$\text{Tr}\Big[Y(I - W)(I - W)^T Y^T\Big] = N\text{Tr}\Big[-\Lambda\Big] \tag{19}$$

$$\text{Tr}\Big[(I - W)^T Y^T Y(I - W)\Big] = N\text{Tr}\Big[-\Lambda\Big] \tag{20}$$

where in the last step we used the fact that the trace operator is invariant under cyclic permutations of products of matrices. But now the left-hand side in the last equation is just the objective function we want to minimize. So, we should choose the smallest eigenvalues and the corresponding eigenvectors. We discard the smallest eigenvalue and the corresponding eigenvector. This enforces the constraint that $\sum_i \mathbf{y}_i = Y\mathbf{1} = \mathbf{0}$ we had ignored earlier (why?).

References

[1] Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2323-2326.