A Survey of Dimensionality Reduction Techniques

RACHEL LEVANGER

This note set gives a short summary of a collection of commonly-used linear and non-linear dimensionality reduction techniques.

Contents

1 Linear Dimensionality Reduction (LDR) Techniques

1.1 Principal Component Analysis (PCA)

PCA is a tool used to describe a data set in terms of its intrinsic directions of highest covariance. Consider a set S of m points in $\mathbb{R}^n$. Then S can be represented as an $m \times n$ matrix X. Call $\mu_X = (\mu_1, ..., \mu_n)$ the center of the data set, or the average of each of the columns of X. The general sequence of steps to perform a PCA on X is as follows:

(1) Compute the covariance matrix Σ associated to the data set X, where $\Sigma_{ij} = \mathrm{cov}(X_{*i}, X_{*j}) = (1/n)\sum_{k=1}^{n}(X_{ki} - \mu_i)(X_{kj} - \mu_j)$. Notice that this is just the dot product of the mean-centered column vectors of the data matrix divided by the dimension of the space.

(2) Since Σ is real and symmetric, we can diagonalize Σ via the matrices Φ and Λ, where Φ is orthogonal and Λ is diagonal. The matrix Φ has columns that form the orthonormal basis of eigenvectors of Σ, Λ gives the corresponding eigenvalues, and the equation $\Sigma = \Phi\Lambda\Phi^T$ holds.

(3) The magnitude of the eigenvalues in Λ provide insight into the number of dimensions along which the data is organized. Furthermore, since the eigenvectors are orthogonal, the covariance along any two transformed basis vectors are zero (their dot product is zero).

(source: https://stat.duke.edu/ sayan/Sta613/2015/lec/IDAPILecture15.pdf)

rachel@math.rutgers.edu