Dimensionality Reduction

The dimensionality reduction problem is as follows. You have a bunch of observations $\mathbf{x}_{1:n}$. Each observation $\mathbf{x}_i$ is a high-dimensional vector, say $\mathbf{x}_i$ is in $\mathbb{R}^D$ with $D \gg 1$. What you would like to do is describe this dataset with a smaller number of dimensions without losing too much information. That is, you would like to project each one of the $\mathbf{x}_i$'s to d-dimensional vector $\mathbf{z}_i$ with $d \ll D$.

Why would you like to do dimensionality reduction? First, you can take any dataset, no matter how high-dimensional, and visualize it in two dimensions. This may help develop intuition about this dataset. Second, once you project the high-dimensional dataset to lower dimensions it is often easier to carry out unsupervised tasks like clustering or density estimation. Third, supervised tasks involving high-dimensional data also become easier if you reduce the dimensionality. For example, if you want to do regression between the high-dimensional $\mathbf{x}$ and a scalar quantity y it will probably pay off if you first reduced the dimensionality of $\mathbf{x}$ by projecting it to a lower-dimensional vector $\mathbf{z}$ and then did regression between $\mathbf{z}$ and y.

There are dozens of dimensionality reduction techniques. See [this] for an incomplete list. In this lecture we are going to develop the simplest and the more widely used dimensionality reduction technique: Principal Component Analysis (PCA). The details of PCA are presented in the video. If you want to read about it independently, I suggest [Chapter 12.1-2, Bishop (2006)].

By Ilias Bilionis (ibilion[at]purdue.edu)