Dimensionality Reduction and Unsupervised Learning

Greg Grudic

Outline

- Principle Component Analysis (PCA)
- Independent Component Analysis (ICA)
- Principle Curves
- Data with outputs
- Partial Least Squares (PLS)
- K Means
- Spectral Clustering
- Locally Linear Embedding (LLE)

Principle Component Analysis (PCA)

- Assume data \( D = \{ (x_1), \ldots, (x_n) \} \), \( x_i \in \mathbb{R}^d \)
- Find k projections in input space

\[
\begin{align*}
  z_1 &= \sum_{i=1}^{n} d_{1i} x_i \\
  z_2 &= \sum_{i=1}^{n} d_{2i} x_i \\
  &\vdots \\
  z_k &= \sum_{i=1}^{n} d_{ki} x_i
\end{align*}
\]

PCA II

- The k projections are called the k principle components
  - The principle components are uncorrelated
- The k projections are the eigenvectors of the data correlation matrix
Figure 3.8: Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects y onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.

Figure 14.20: The first linear principal component of a set of data. The line minimizes the total squared distance from each point to its orthogonal projection onto the line.

Figure 14.21: The best rank-two linear approximation to the half-sphere data. The right panel shows the projected points with coordinates given by $U_2 \Sigma_2$, the first two principal components of the data.

Figure 14.22: A sample of 130 handwritten threes shows a variety of writing styles.
Kernel PCA

- The data is projected into a kernel matrix
- The kernel matrix is centered and the top k eigenvectors are obtained
- This gives the following nonlinear projections

\[
\hat{f}(\lambda) = \tilde{x} + \lambda_1 v_1 + \lambda_2 v_2
\]

Here we have displayed the first two principal component directions, \(v_1\) and \(v_2\), as images.
Independent Component Analysis (ICA)

• Similar to PCA – find k projections
  \[ z_i = \sum_{i=1}^{d} a_{i1} x_1 + \sum_{i=2}^{d} a_{i2} x_2 + \cdots + \sum_{i=k}^{d} a_{ik} x_k \]

• However, the independent components are now assumed to be statistically independent rather than uncorrelated
  – How this independence is defined is an open research question (e.g. all moments have zero dependence)
What if your data has outputs?

- Data \( D = \{ (x_1, y_1), \ldots, (x_N, y_N) \} \)

- Can build models in ICA, PCA or Principle Curve Space:
  \[ \hat{y} = \hat{f}(z_1, \ldots, z_k) \]

- The model can be generated using any supervised learning algorithm

- However, the \( (z_1, \ldots, z_k) \) may not be good predictors

Partial Least Squares (PLS)

- Uses the outputs to obtain the principle components.

- For the components \( m = 1, \ldots, k \) PLS maximizes
  \[ \max_{H^1} \frac{\text{Corr}^2(y, X_\alpha) \text{Var}(X_\alpha)}{\sqrt{\sum_{\alpha=1}^m \text{Var}(X_\alpha)}} \]

- Compare to PCA
  \[ \max_{H^1} \frac{\text{Var}(X_\alpha)}{\sqrt{\sum_{\alpha=1}^m \text{Var}(X_\alpha)}} \]
Spectral Clustering

- Essentially K Means in the eigenvector space of the Kernel Matrix
  - Usually a Gaussian Kernel is used.

Locally Linear Embedding (LLE)