Top 10 Data Mining Algorithm (2008)

- C4.5
- K-means
- Support vector machines
- Apriori
- EM
- PageRank
- AdaBoost
- KNN
- Naïve Bayes
- CART

Nonlinear Dimensionality Reduction

Chen Yu
Indiana University

Dimensionality Reduction

- The goal:
  The meaningful low-dimensional structures hidden in their high-dimensional observations.
- Classical techniques
  - Principle Component Analysis—preserves the variance
  - Multidimensional Scaling—preserves inter-point distance
  - Isomap
  - Locally Linear Embedding

Why Dimensionality Reduction

- The curse of dimensionality
- Number of potential features can be huge
  - Image data: each pixel of an image
    - A 64X64 image = 4096 features
  - Text categorization: frequencies of phrases in a document or in a web page
    - More than ten thousand features
Why Dimensionality Reduction

• Two approaches to reduce number of features
  – Feature selection: select the salient features by some criteria
  – Feature extraction: obtain a reduced set of features by a transformation of all features

• Data visualization and exploratory data analysis also need to reduce dimension
  – Usually reduce to 2D or 3D

Dimensionality reduction

• Three (simple) examples of manifolds
  • All three are two-dim. data embedded in 3D
  • Linear, “S”-shape, “Swiss roll”
  • For all three, we would like to recover:
    • That the data is only two-dimensional
    • “Consistent” locations for the data in 2D

Manifold Learning

• A manifold is a topological space which is locally Euclidean
• An example of nonlinear manifold:

Manifold Learning

• Discover low dimensional representations (smooth manifold) for data in high dimension.

• Linear approaches (PCA, MDS)

• Non-linear approaches (ISOMAP, LLE, others)
PCA

Linear Approach - PCA

• PCA Finds subspace linear projections of input data.

Linear Approach - classical MDS

• MDS: Multidimensional scaling
  • Borg and Groenen, 1997

• MDS takes a matrix of pairwise distances and gives a mapping to R^p. It finds an embedding that preserves the interpoint distances, equivalent to PCA when those distance are Euclidean.
  • Low dimensional data for visualization

An example
Nonlinear Dimensionality Reduction

- Many data sets contain essential nonlinear structures that invisible to PCA and MDS
- Resorts to some nonlinear dimensionality reduction approaches.

Nonlinear Approaches- Isomap

Josh. Tenenbaum, Vin de Silva, John langford 2000

- Constructing neighbourhood graph G
- For each pair of points in G, Computing shortest path distances —— geodesic distances.
- Use Classical MDS with geodesic distances.
  Euclidean distance $\rightarrow$ Geodesic distance

Results

MDS

given: - a set of $n$ objects
- the dissimilarities $\delta_{ij}$ between them
find: points on the plane whose distances $d_{ij}$ are as close as possible to the $\delta_{ij}$

minimize: $\text{STRESS} = \frac{\sum_{i<j} (d_{ij} - \delta_{ij})^2}{\sum_{i<j} \delta_{ij}^2}$ [Kruskal 1964]
Intuition

- Built on top of MDS.
- Capturing in the geodesic manifold path of any two points by concatenating shortest paths in-between.
- Approximating these in-between shortest paths given only input-space distance.

Sample points with Swiss Roll

- Altogether there are 20,000 points in the “Swiss roll” data set. We sample 1000 out of 20,000.

Construct neighborhood graph G

K-nearest neighborhood (K=7), $D_{k}$ is 1000 by 1000 (Euclidean) distance matrix of two neighbors (figure A)

Compute all-points shortest path in G

Now $D_{g}$ is 1000 by 1000 geodesic distance matrix of two arbitrary points along the manifold (figure B)
Use MDS to embed graph in $\mathbb{R}^d$

Find a d-dimensional Euclidean space $Y$ (Figure c) to preserve the pairwise distances.

Algorithm Description

- Step 1
  Determining neighboring points within a fixed radius based on the input space distance $d_{ij}$
  These neighborhood relations are represented as a weighted graph $G$ over the data points.
- Step 2
  Estimating the geodesic distances $d_{ij}$ between all pairs of points on the manifold $M$ by computing their shortest path distances $d_{ij}$ in the graph $G$
- Step 3
  Constructing an embedding of the data in $d$-dimensional Euclidean space $Y$ that best preserves the manifold's geometry

Dynamic Programming

From S to F, 4 steps, two states in each step.

Isomap: Advantages

- Nonlinear
- Globally optimal
  - Still produces globally optimal low-dimensional Euclidean representation even though input space is highly folded, twisted, or curved.
- Guarantee to recover the true dimensionality.
Isomap: Disadvantages

- May not be stable, dependent on topology of data
- Guaranteed to recover geometric structure of nonlinear manifolds
  - As $N$ increases, pairwise distances provide better approximations to geodesics, but cost more computation
  - If $N$ is small, geodesic distances will be very inaccurate.

Applications

- Isomap and Nonparametric Models of Image Deformation
- Image Spaces and Video Trajectories: Using Isomap to Explore Video Sequences
- Mining the structural knowledge of high-dimensional medical data using isomap

Isomap Webpage: http://isomap.stanford.edu/
LLE

- Neighborhood preserving embeddings
- Mapping to a global coordinate system of low dimensionality
- No need to estimate pairwise distances between widely separated points
- Recovering global nonlinear structure from locally linear fits

Characteristics of a Manifold

Key: how to combine all local patches together?

$\mathbb{R}^n$ locally is a linear patch

$\mathbb{R}^2$ $x_1$ $x_2$

$LLE$: Intuition

- Assumption: manifold is approximately “linear” when viewed locally, that is, in a small neighborhood
  - Approximation error, $e(W)$, can be made small

$$
\min_{W} \| X_i - \sum_{j=1}^{K} W_{ij} X_j \|^2 \quad (1)
$$

- Local neighborhood is effected by the constraint $W_{ij}=0$ if $z_i$ is not a neighbor of $z_j$

- A good projection should preserve this local geometric property as much as possible

Fig. 1. The problem of nonlinear dimensionality reduction as illustrated [30] for three-dimensional data (B) sampled from a two-dimensional manifold (A). An unsupervised learning algorithm must discover the global internal coordinates of the manifold without signals that explicitly indicate how the data should be embedded in two dimensions. The color coding illustrates the neighborhood-preserving mapping discovered by LLE; black outlines in (B) and (C) show the neighborhood of a single point. Unlike LLE, projections of the data by principal component analysis (PCA) [28] or classical MDS (C) map faraway data points to nearby points in the plane, failing to identify the underlying structure of the manifold. Note that mixture models for local dimensionality reduction [20], which cluster the data and perform PCA within each cluster, do not address the problem considered here namely, how to map high-dimensional data into a single global coordinate system of lower dimensionality.
Fit Locally, Think Globally

We expect each data point and its neighbors to lie on or close to a locally linear patch of the manifold.

Each point can be written as a linear combination of its neighbors.
The weights chosen to minimize the reconstruction error.

\[
\text{min}_{w} \sum_{i=1}^{K} \left| X_i - \sum_{j=1}^{K} W_{ij} X_j \right|^2 \tag{1}
\]

Properties

- The weights that minimize the reconstruction errors are invariant to rotation, rescaling and translation of the data points.
- The weights characterize the intrinsic geometric properties of each neighborhood.
- The same weights that reconstruct the data points in D dimensions should reconstruct it in the manifold in d dimensions.

Local geometry is preserved

Dimension Reduction

\[
\Phi(Y) = \sum_{i=1}^{K} \left| \tilde{Y}_i - \sum_{j=1}^{K} W_{ij} \tilde{Y}_j \right|^2 \tag{2}
\]

The LLE algorithm

**LLE ALGORITHM**

1. Compute the neighbors of each data point, \( \tilde{X}_i \).
2. Compute the weights \( W_{ij} \) that best reconstruct each data point \( \tilde{X}_i \) from its neighbors, minimizing the cost in eq (1) by constrained linear fits.
3. Compute the vectors \( \tilde{Y}_i \) best reconstructed by the weights \( W_{ij} \), minimizing the quadratic form in eq (2) by (sp)s fema unie eigenvectors.

Figure 5: Summary of the LLE algorithm, mapping high dimensional data points, \( \tilde{X}_i \), to low dimensional embedding, vectors, \( \tilde{Y}_i \).
Summary
• LLE attempts to preserve local geometry of the data by mapping nearby points on the manifold to nearby points in the low dimensional space
  • Find k nearest neighbors in X space
    (a demo in NYC hall of science)
  • Solve for reconstruction weights W
  • Compute embedding coordinates Y using weights W.

Examples