Advanced Data Analysis and Machine Learning

Lecture: Dimensionality Reduction

Lasse Lensu

2015-10-04

Outline

LUT
Lappeenranta
University of Technology

Data dimensionality

LUT Lappeenranta University of Technology

- Various direct or indirect measurements as sources of information produce information on the measurement target.
- The data can be used to characterise the measurement target and model its behaviour.
- However, adding different kinds of measurements does not guarantee better characterisation or model because of the varying representativeness of a measurement, or noise.
- Data dimensionality increases with the number of measurement channels, and processing the data requires more computational effort.
- More samples are needed to estimate the parameters of a model characterising the measurement target or process.
- However, the measurements generating the optimal or good enough characterisation for a specific purpose can be selected.

Dimensionality reduction

LUT Lappeenranta University of Technology

- Data dimensionality can be reduced intelligently.
- Various linear and nonlinear methods exist for the purpose.
- Several methods are based on the idea that the data lies on or near a low-dimensional manifold (residing in the high-dimensional space).
- In addition to these methods, data preprocessing can be used to efficiently remove data characteristics problematic for efficient reduction of dimensions.
- Dimensionality reduction is a useful tool in data analysis and machine learning since it mitigates undesired properties of high number of dimensions.

Problem of dimensionality reduction

LUT Lappeenranta University of Technology

- The purpose of dimensionality reduction is to find a manifold to characterise a specific set of data optimally or well enough, and represent the data by using the manifold.
- Problem definition [2]:
  - Let us have a $n \times D$ matrix $\boldsymbol{X}$ consisting of n vectors of data x_i with dimensionality D. The dataset has intrinsic dimensionality d where $d < D$, and often $d \ll D$.
  - A dimensionality reduction technique transforms dataset $\boldsymbol{X}$ into a new dataset $\boldsymbol{Y}$ with dimensionality d, while retaining the geometry of the data as much as possible.
  - Dimensionality reduction is an ill-posed problem because generally the geometry of the manifold embedded in the high-dimensional space and the intrinsic dimensionality d of the dataset $\boldsymbol{X}$ are unknown $\Rightarrow$ it is necessary to make assumptions of the data to solve the problem.

Purpose of reducing dimensions

LUT Lappeenranta University of Technology

- The purpose of dimensionality reduction is to transform high-dimensional data into a representation of reduced dimensionality.
- In an ideal situation, the new representation corresponds to the intrinsic dimensionality of the data.
- The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [1].
- Where from do we get the number of intrinsic dimensions?

Dimensionality reduction methods

LUT Lappeenranta University of Technology

- Dimensionality reduction approaches can be diveded into the following categories [2]:
  - Linear methods
    such as principal component analysis (PCA)
  - Global nonlinear methods
    such as multidimensional scaling (MDS)
  - Local nonlinear methods
    such as locally linear embedding (LLE)
  - Variants of local nonlinear methods
    such as conformal eigenmaps
  - Global linear-model alignment methods
    such as locally linear coordination (LLC)

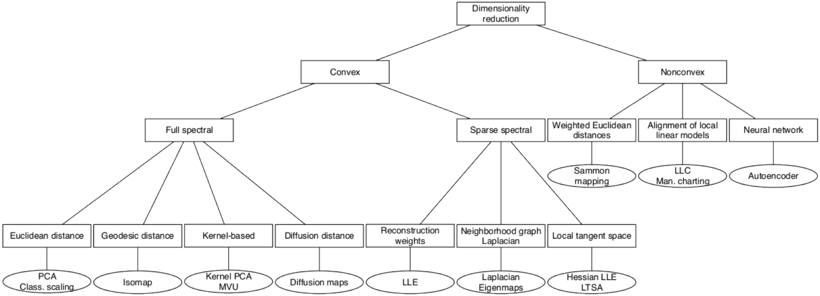
Dimensionality reduction concepts

LUT Lappeenranta University of Technology

- Convex methods optimise an objective function that does not contain any local optima, that is, the solution space is convex. A common form for the objective function is $\phi(\boldsymbol{Y}) = \frac{\boldsymbol{Y}^\intercal \boldsymbol{A} \boldsymbol{Y}}{\boldsymbol{Y}^\intercal \boldsymbol{B} \boldsymbol{Y}}$ (generalized Rayleigh quotient to get either exact or approximate eigenvalues).
- Non-convex methods optimise objective functions that do contain local optima.
- Full spectral methods perform an eigendecomposition of a full matrix capturing the covariances between dimensions, or the pairwise similarities between datapoints.
- Sparse spectral methods solve a sparse eigenproblem, and they commonly aim to retain the local structure in the data.

Method taxonomy

LUT
Lappeenranta
University of Technology



Dimensionality reduction (DR) method taxonomy [3].

Principal component analysis

LUT Lappeenranta University of Technology

- Constructs a representation of the data by finding a linear basis of reduced dimensionality in which the variance is maximal.
- PCA seeks for a linear mapping $\boldsymbol{M}$ which maximises trace$(\boldsymbol{M}^\mathsf{T}\boldsymbol{\Sigma}\boldsymbol{M})$ where $\Sigma_{ij} = \mathrm{cov}(x_i, x_j) = \mathrm{E}\left[(x_i - \mu_i)(x_j - \mu_j)\right]$.
- The principal components (PCs) are the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$. They are found by solving $\boldsymbol{\Sigma}\boldsymbol{M} = \lambda\boldsymbol{M}$, and the new representation $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{M}$.
- To reduce the number of dimensions, only the the first l eigenvectors (in the decreasing order of variance) are selected while minimising the total reconstruction error $\|\boldsymbol{X} - \boldsymbol{X}_L\|_2$.
- The main disadvantage of PCA is that the size of the covariance matrix is proportional to the data dimensionality, but alternative ways to determine the eigenvectors exist.

Multidimensional scaling

LUT Lappeenranta University of Technology

- Nonlinear DR method with various modifications.
- Produces a new representation by minimising a loss function taking into account the pairwise distances between the datapoints in both the low and high-dimensional spaces.
- The loss function can be defined a various ways: for example, the raw stress function $\phi(\boldsymbol{Y}) = \sum_{ij}(\|x_i - x_j\|^2 - \|y_i - y_j\|^2)$ where x is a high-dimensional datapoint, y is a low-dimensional datapoint, and $\|\cdot\|$ is the Euclidean distance ($y_i = x_i\boldsymbol{M}$ and $\|m_j\|^2 = 1\ \forall j$).
- The loss function can be minimised, for example, by performing the eigendecomposition of a pairwise dissimilarity matrix, or with the conjugate gradient method.
- Selection of the number of dimensions affects the difficulty of interpreting the results.

Locally linear embedding

- Local and nonlinear DR method.
- Constructs a low-dimensional representation of the original datapoints with the aim to preserve only the local properties of the manifold around each datapoint.
- The method describes each datapoint x_i as a linear combination W_i of its k nearest neighbours x_{i_j} by fitting a hyperplane (assumes the manifold to be locally linear) through the datapoint and its neighbours.
- LLE tries to retain the datapoint reconstruction weights in the low-dimensional representation as good as possible by minimising the cost function $\phi(\boldsymbol{Y}) = \sum_i \|y_i - \sum_{j=1}^k w_{ij} y_{i_j}\|^2$ (subject to $\|y^{(k)}\|^2 = 1 \; \forall k$).
- The points in the low-dimensional representation y_i minimising the cost function can be computed by finding the eigenvectors corresponding to the smallest d nonzero eigenvalues of the inproduct of $(\boldsymbol{I} - \boldsymbol{W})$.

Conformal eigenmaps

- Local nonlinear techniques for dimensionality reduction do not employ information on the geometry of the manifold that is contained in discarded eigenvectors (with small eigenvalues).
- A conformal transformation preserves the angles between neighbouring datapoints in dimensionality reduction.
- Conformal eigenmaps starts with a local nonlinear method for dimensionality reduction to reduce the high-dimensional data to a dataset of dimensionality d_t where $d < d_t < D$.
- Based on the previous representation and guided by a conformality measure, conformal eigenmaps constructs a d-dimensional representation that preserves the angles between the neighbouring datapoints as well as possible.

Locally linear coordination

LUT Lappeenranta University of Technology

- LLC computes a number of locally linear models and aligns the models globally.
- The method has two steps:
  1. Compute a mixture of local linear models (factor analysers) on the data by using the expectation maximisation (EM) algorithm. The mixture of models represents joint variations (correlations) in the high-dimensional data, and the idea is to seek for latent unobserved variables (factors) explaining the variations.
  2. Align the models by finding a linear transformation based on the data models that minimizes the LLE cost function (by solving a generalised eigenproblem).

Number of intrinsic dimensions

LUT
Lappeenranta
University of Technology

- The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [1].
- The intrinsic number of dimensions can be estimated by local or global estimators.
- Local estimators:
  - Rely on the idea that the number of datapoints within radius r from a data point increases proportional to r^d where d is the intrinsic dimensionality (around that datapoint).
  - Average over local dimensionality estimates.
- Global estimators:
  - Treat the data as a whole.
  - For examples, make use of eigenvalues, r-coverings (how many hyperspheres are needed to cover the whole dataset) or spanning trees.

Summary

LUT Lappeenranta University of Technology

- The purpose of dimensionality reduction is to find a manifold to characterise a specific set of data optimally or well enough, and represent the data by using the manifold.
- Various linear and nonlinear methods exist for the purpose, and the intrinsic dimensionality of a dataset can be estimated by using either a local or global estimator.
- Dimensionality reduction is a useful tool in data analysis and especially in machine learning since it mitigates undesired properties of high number of dimensions.

References

LUT Lappeenranta University of Technology

- K. Fukunaga.
  Introduction to Statistical Pattern Recognition.
  Academic Press, CA, USA, 1990.
- L.J.P. van der Maaten.
  An introduction to dimensionality reduction using matlab.
  Technical Report MICC 07-07. Maastricht University, Maastricht, The Netherlands, 2007.
- L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik.
  Dimensionality reduction: A comparative review.
  Tilburg University Technical Report, TiCC-TR 2009-005, 2009.