Field:

Theoretical and Applied Mathematics/Informatics

Session Topic:

Statistics for Large-Dimensional Data

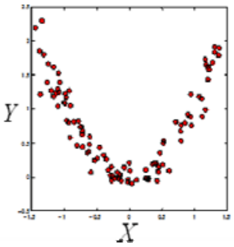
Speaker:

Kenji FUKUMIZU, The Institute of Statistical Mathematics

1. Introduction

In processing large-dimensional data, it is often useful to express them in a low-dimensional space. Such a technique is called dimension reduction. The purpose of dimension reduction includes easier interpretation of data, de-noising for extracting meaningful features, and preprocessing of further data analysis. Classical methods for dimension reduction are linear methods such as linear regression analysis and principal component analysis. These methods, however, consider only the linear relations among variables, and fail to capture more complex nonlinear structure (Fig.1).

There have been many approaches to extract nonlinear structure of data. One of the most popular ones is to use the powers of the variables or expansion by basis functions. For three dimensional vector (*x*, *y*, *z*), for example, expanding it into (*x*, *y*, *z*, x^2, y^2, z^2, *xy*, *yz*, *zx*, …) gives a representation of the nonlinear structure of the data. While this strategy may work effectively in some cases, there is computational difficulty for large-dimensional data. In fact, if we wish to use the moments up to the third order for 100 dimensional data, the power expression has more than 1,600,000 components, for which matrix operations required data analysis, such as inversion and eigendecomposition, are not feasible even with current high-power computers.



Corr(X, Y) = 0.17

Corr(X^2, Y) = 0.96

Figure 1.

2. Kernel method for nonlinear data analysis

The kernel method is a recent technique that extracts nonlinearity or higher-order structure of data in a more systematic and efficient way. The method uses positive definite kernels to map data into a special type of vector space, called reproducing kernel Hilbert space, to represent nonlinear features, in which the inner product of two such vectors is simply given by evaluation of the kernel. This efficient way of computing the inner product is often called “kernel trick”. Since most of the linear statistical methods use only the inner product of data points, many such methods can be easily and efficiently applied to the mapped feature vectors, which incorporate the nonlinear

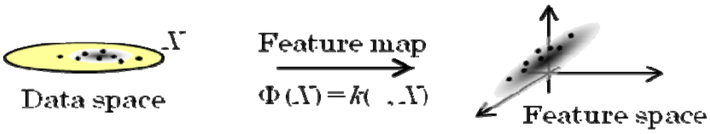


Figure 2. Idea of kernel method

structure of the data. The computation of the kernel methods is typically reduced to processing matrices of sample size; so the dimensionality of the data does not matter, while the number of data is expected not to be very large. This is a suitable property for recent data analysis, in which we often have small number of data of large dimensionality.

3. Kernel dimension reduction for regression

I discuss a kernel method for dimension reduction in regression situation, where the probabilistic relation from an m-dimensional vector $X = (X_1, \ldots, X_m)$ to another variable Y is estimated with data. The goal is to find a small number of linear combinations of X_i so that they explain the variable Y most effectively. While the problem is to extract linear combinations, it is still important to consider nonlinear dependence between X and Y.

The basic idea for finding the effective directions in X is to use the gradient of the regression function $g(x) = \mathrm{E}[Y \mid X = x]$, which explains the largest change of variable Y. As an extension of this idea, the regression functions of feature vector $\phi(Y)$, $\mathrm{E}[\phi(Y) \mid X = x]$, for various ϕ is advantageous for considering more general probabilistic structure among X and Y. The kernel method gives an efficient way of incorporating (almost all) nonlinear transform of Y in finding the effective directions; the involved computation is only the eigendecomposition of certain matrix given by the positive definite kernels and data.

To demonstrate the effectiveness of the method, I show some experimental results for practical data in my talk. As an example, for a speech signal classification task, in which speech signals of 26 alphabets are represented by 617 dimensional vectors, after taking 25 dimensional linear features given by the kernel method, a simple k-nearest neighbor classifier is applied. The classification accuracy is comparable with the best ones obtained by advanced nonlinear classifiers such as neural networks and C4.5. This means the method successfully extracts the effective directions for the classification task.

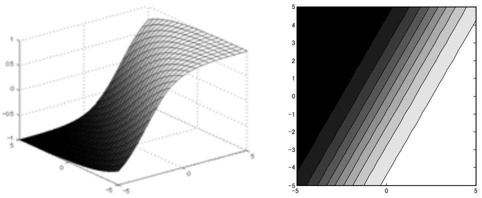


Figure 3. regression function and its contour

Conclusion

For processing large-dimensional data, dimension reduction techniques are often useful. Beyond the classical techniques, it is important to incorporate the nonlinear or higher-order structure in a computationally efficient manner. The kernel method is a recent strong methodology for this purpose. I explain the kernel method for dimension reduction in regression, and demonstrate its effectiveness with various real world data.

References

[1] Fukumizu, K. and C. Leng. Gradient-based kernel dimension reduction for supervised learning. 2011. arXiv:1109.0455v1.

[2] Fukumizu, K. Francis R. Bach and M. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics.* 37(4), pp.1871-1905 (2009)