Imperial College
London

12 June 2020

Guided Dimensionality Reduction for Anomaly Detection

M4R Oral Presentation

Tudor Trita Trita

Outline of the Presentation

1. Background to dimensionality reduction
2. Dimensionality reduction methods
3. Experiments
4. Closing Remarks

Background to dimensionality reduction

What is dimensionality reduction?

- It is the process of reducing the dimensionality of your data.
- If we assume that our data can be expressed as a matrix, we wish to reduce the number of columns (features) of our data.
- This motivates the concept of a dimension reducing function
- Let $R : \mathbb{R}^{n \times p} \to \mathbb{R}^{n \times q}, q < p$ and $X \in \mathbb{R}^{n \times p}$ be our data (feature) matrix. Then we call $R(X)$ a dimension reduction function.

Types of dimensionality reduction (1)

- We can divide dimension reduction techniques into two broad types; supervised and unsupervised.
- Unsupervised dimension reduction methods reduce the data using knowledge on the features X only. The most common example of this is principal components analysis (PCA).
- Supervised dimension reduction methods reduce the data making use of label (response) variables Y associated with the features X.
- It has been shown that supervised methods perform better than their unsupervised counterparts, but the requirement for labelled data limits their usage.

Types of dimensionality reduction (2)

- We can further subdivide the dim. reduction methods into linear and non-linear methods.
- Linear methods reduce the data by searching for linear directions (straight lines) in our data that capture as much information about the data as possible.
- Non-linear methods don't have this constraint and can search for a general manifold in our feature data.
- Non-linear methods outperform linear methods for capturing as much information in our data at the expense of both mathematical and computational complexity.

Why reduce the data in the first place?

1. The obvious reason is that we can reduce the storage space required to keep data.
2. If our data is highly collinear, i.e. the columns of the matrix X are highly correlated, we can apply dim. reduction methods to remove some of this multi-collinearity.
3. Visualisation of data in few dimensions (2 or 3).
4. Helps mitigate the curse of dimensionality effect.

Usages of dim. reduction techniques

- The most widely use of dim. reduction methods is as a pre-processing step that is part of a larger process.
- Data compression, for example image or video compression. A typical 2 hour HD movie could easily take 1 terabyte if it were fully uncompressed.
- Anomaly detection. Imagine a scenario where we need to monitor thousands of sensors in a machine. If we can express the data from these sensors in a lower-dimensional space, the detection of anomalies becomes much easier.

Sufficient dimension reduction

- In this project, we will be focusing on supervised dim. reduction techniques.
- We are interested in estimating the regression function $\mathbb{E}[Y|X]$ by a conditional expectation from a lower dimensional subspace $\mathbb{E}[Y|R(X)]$.
- We say that the $R(X)$ is a *sufficient dimension reduction* if we can retain all the information about Y after reducing the dimensionality of our data.

Dimension Reduction Subspaces

Definition (Dimension Reduction Subspace)

Let $\mathrm{x} \in \mathbb{R}^p$ be a vector, an observation from our feature space and suppose that we can express a sufficient dimension reduction $R(x) := B^T\mathrm{x}$, for some matrix $B \in \mathbb{R}^{p\times q}$. Denote $\mathcal{S}(\mathcal{B})$ to be the subspace of $\mathbb{R}^p$ spanned by the columns of B. The space $\mathcal{S}(\mathcal{B})$ is called a *dimension reduction subspace* (DRS).

Central Subspace

Definition (Central Subspace)

Let $\mathcal{S}_{y|\mathrm{x}}$ be the intersection of all DRS's. It can be shown that $\mathcal{S}_{y|\mathrm{x}}$ is a DRS under certain conditions. If we assume that $\mathcal{S}_{y|\mathrm{x}}$ is a DRS, then it is known as the *central subspace*, with associated *structural dimension* $d = \dim(\mathcal{S}_{y|\mathrm{x}})$.

All of the methods in the next section will attempt to discover the central subspace of data.

Dimensionality reduction methods

Partial Least Squares (PLS)

- Supervised extension of PCA. In PCA, we are interested in finding directions in X that are most correlated. In PLS, we require the directions in X to be correlated with the responses Y as well.
- Here, we want to decompose the matrices X and Y into

$$
\begin{aligned}
X &= TP^T + E_X, \\
Y &= UQ^T + E_Y,
\end{aligned}
$$

  where we try to maximise the covariance between the matrices T and U.

Positive Definite Kernel

- The remaining two methods to present use positive definite kernels as tools for extracting information from data.
- A positive definite (symmetric, real-valued) kernel is a function $k : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that $k(x, y) = k(y, x) \ \forall x, y \in \mathcal{H}$ and $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0, \forall x_i \in \mathcal{H}, c_i \in \mathbb{R}$.
- We can express a positive definite kernel in terms of its feature map. Let V be an inner product space, let $\Phi : \mathcal{X} \to V, \quad x \to \Phi(x)$. Then, the kernel on $\mathcal{X}$ defined by $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ is positive definite.

Reproducible Kernel Hilbert Spaces (RKHS)

An RKHS is a Hilbert space with the additional reproducing property. A definition is given below:

Definition (RKHS)

Let Ω be a set and $\mathcal{H}$ a Hilbert space. If $\mathcal{H}$ consists of functions on Ω s.t. $\forall x \in \Omega, \exists$ a function $k_x \in \mathcal{H}$ satisfying the reproducing property

$$\langle f, k_x \rangle_{\mathcal{H}} = f(x), \quad (\forall f \in \mathcal{H}), \tag{1}$$

then $\mathcal{H}$ is said to be a reproducible kernel Hilbert space and $k(\cdot, x) := k_x(\cdot)$ is called the reproducing kernel of $\mathcal{H}$.

It can be shown that there exists an unique RKHS for every positive kernel.

Kernel Dimension Reduction (1)

- In this project, we explore the gKDR method, as presented in the paper (Fukumizu, 2014).
- We can write the central space condition as finding a matrix B such that $p(Y|X) = \tilde{p}(Y|B^T X)$ holds, where $p, \tilde{p}$ are conditional pdfs.
- Assuming the partial derivatives exist, the expression

$$\frac{\partial}{\partial x}\mathbb{E}[Y|X = x] = B \int y \frac{\partial \tilde{p}(y|z)}{\partial z}\bigg|_{z=B^T x} \mathrm{d}y$$

holds and implies that the gradient is contained in the central subspace.

Imperial College
London

Kernel Dimension Reduction (2)

- We define the cross-covariance operator C_{XY} through

$$\langle g, C_{YX} f \rangle_{\mathcal{H}_\mathcal{Y}} = \mathbb{E}_{XY}[\langle f, \Phi_\mathcal{X}(X) \rangle_{\mathcal{H}_\mathcal{X}} \langle \Phi_\mathcal{Y}(Y), g \rangle_{\mathcal{H}_\mathcal{Y}}],$$

  which generalises covariance to work with kernels and RKHS.

- We can then find an estimator for

$$\mathbb{E}[g(Y)|X = \cdot] \approx (\widehat{C}_{XX}^{(n)} + \varepsilon_n I)^{-1} \widehat{C}_{XY}^{(n)} g.$$

- The last key property, $f \in \mathcal{H}_\mathcal{X}$, f continuously differentiable is

$$\frac{\partial f(x)}{\partial x} = \left\langle f, \frac{\partial}{\partial x} k(\cdot, x) \right\rangle_{\mathcal{H}_\mathcal{X}}$$

Kernel Dimension Reduction (3)

- After some algebra, and using the facts from the previous slide, we find the matrix whose eigenvectors are contained in the central subspace

$$\widehat{M}_n(x) = \nabla \mathrm{k}_X(x)^T \left(G_X + n\varepsilon_n I\right)^{-1} G_Y \left(G_X + n\varepsilon_n I\right)^{-1} \nabla \mathrm{k}_X(x).$$

- We then take the first d eigenvectors of the matrix

$$\tilde{M}_n := \frac{1}{n} \sum_{i=1}^{n} \widehat{M}_n\left(X_i\right)$$

as the columns of the estimator for B, which forms the gKDR estimator for the matrix whose columns span the central subspace.

Remarks on gKDR

- G_X, G_Y are gram matrices, whose elements are kernels
- Computing the gKDR estimator involves inverting the gram matrix, which is $N \times N$, where N is the number of observations in our data, hence making it an $O(N^3)$ algorithm in general. For some kernels, e.g. Gaussian RBF, we can use low-rank approximations to make this faster.
- Cross-validation or other model selection methods need to be used to optimise the choice of kernel and kernel hyperparameters as well as final dimensionality of our data d.

Dimensionality Reduction through HSIC

- We now introduce the third dimensionality reduction method explored in the project.
- Here, we want to estimate the central subspace by finding a transformation R which makes the transformed data $R(X)$ as **dependent** to Y as possible.
- Assuming we can express $R(X) = B^T X$, for some matrix B such that $BB^T = I$, the goal here is to find B which makes $B^T X$ as dependent to Y as possible.
- We use the Hilbert-Schmidt independence criterion (HSIC) to quantify the degree of dependence between two vectors.

A swift introduction to HSIC

- The HSIC is defined as the (Hilbert-Schmidt) norm of the kernel cross-covariance of X and Y:

  $$HSIC(X, Y) = \mathbb{E}_{x,y}\left[\|(\phi(x) - \mu_x) \otimes (\varphi(y) - \mu_y)\|_{HS}^2\right],$$

  where μ_z is defined by $\langle \mu_z, f\rangle_{\mathcal{H}_{\mathcal{Z}}} = \mathbb{E}_z[f(z)]$.
- We can obtain a (biased) estimator to the HSIC with

  $$\widehat{HSIC}(X, Y) = \frac{1}{(N-1)^2} tr(G_X H G_Y H), \quad H := I - \frac{1}{N} 1^T 1,$$

  where G_X, G_Y are the gram matrices for the kernels k_x, k_y.

Method for dim. reduction with HSIC

1. Given data (X, Y), choose kernels k_X, k_Y and an initial guess for the matrix B.
2. Find the matrix B that minimises $-tr(G_X(B)HG_Y H)$, where

$$(G_X(B))_{ij} = k_{\mathcal{X}}(B^T X_i, B^T X_j)$$

by gradient descent, subject to the constraint that $B^T B = I$.

3. Once the optimisation is complete, we take the optimising matrix B^* as our estimate for the matrix with columns that span the central subspace.

Remarks on HSIC method

- The method above maximises $HSIC(B^TX, Y)$, which is equivalent to making B^TX and Y as dependent as possible.
- The choice of kernel and kernel hyper-parameters greatly affect the performance of this method. Cross-validation or other model selection methods must be used.
- The optimisation step as well as the computation of the gram matrices may be very expensive.
- The optimisation process is equivalent to searching for B in the manifold of orthogonal matrices, also known as the Stiefel manifold.

Experiments

Experiment 1: Synthetic data for regression

- This experiment tests the three methods on 4 types of synthetic data.
- Model (A) is a simple linear combination of 3 columns of our data with additive Gaussian noise.
- Model (B) represents data that is highly collinear and has effectively 2 dimensions.
- Model (C) has data that is non-linear with additive Gaussian noise.
- Model (D) has non-linear data with multiplicative skewed noise drawn from a Beta(2, 4) distribution.

Imperial College London

Results from experiment 1

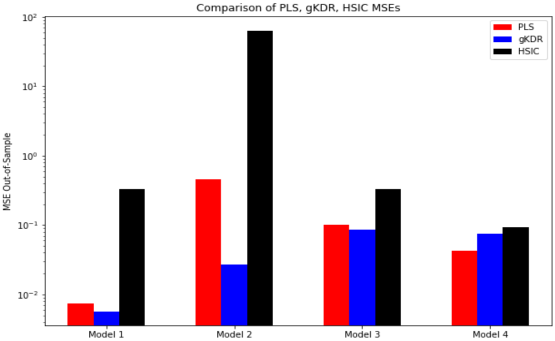

Figure 1: MSE for kNN after using dim. reduction

Experiment 2: Twitter dataset

- We have data on number of mentions of 10 publicly traded US companies every 5 minutes for 2 months between 26th February 2015 to 23rd April 2015. This is 15902 observations on 10 variables.
- We want to monitor for any anomalous data points in the volume of Apple mentions by looking at the volume for every other company for the previous 10 observations (55 minutes). This effectively makes our data have dimension 100.
- To make monitoring easier, we will apply dim. reduction to the data, trying to find a 3 dimensional representation, which we then perform anomaly detection on.

HSIC: 3-dimensional subspace

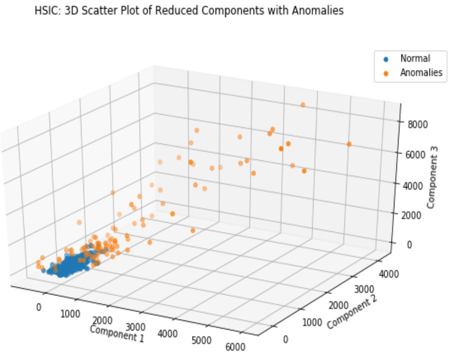

Figure 2: 3D representation of 100-dimensional data using HSIC

Imperial College
London

HSIC: 3-dimensional subspace

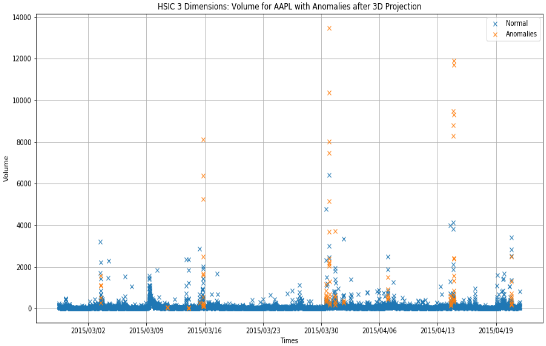

Figure 3: Plot of the anomalies on the graph of Apple volume

Remarks on Experiment 2

- Very similar results were found for both PLS and gKDR, namely, a cluster near the origin indicating normal points and anomalous points away from the main cluster.
- Figure 3 shows that days with high volume were (correctly) picked up as anomalous.
- Since labels for anomalous points were not available, it is difficult to quantify the level of error in the results.
- Both gKDR and HSIC needed to train on less than 2% of the data to achieve similar results as PLS trained on the entire data.
- The day with the highest volume was the 27th Feb. 2015, which was the day Apple got sued over some patent rights by Ericsson.

Closing Remarks

Closing Remarks

- In this project, we have shown that using HSIC with the novel optimisation procedure of searching through the Stiefel manifold performs well as a dimensionality reduction method.
- Although computing the HSIC estimator for the central subspace involves expensive matrix calculations, we showed that it needed a small amount of training data to yield good results.

Thank you for listening!