Anomaly Detection with Robust Deep Autoencoders

original article by C. Zhou and R. C. Peffenroth

Alessandro Trenta

Scuola Normale Superiore

Deep Autoencoders

- A Deep Autoencoder (DAE) is constituted by two main components: an encoder E and a Decoder D.
- The main objective of a DAE is to learn the identity map so that the reconstruction $\bar{X} = D(E(X))$ is as close as possible to the original input X.
- The encoder and decoder functions E, D can be any kind of mapping between the data space and the coding space. Usually they are Deep Neural Networks e.g. a feed forward network or even more complex models such as Long ShortTer Memory (LSTM).
- The objective is usually to find the minimum reconstruction error w.r.t. some parametrized encoding and decoding functions and a distance (in this case the L_2 norm)

$$\min_{\theta,\phi} \|X - D_\theta(E_\phi(X))\|_2 \tag{1}$$

Principal Component Analysis

- Assume to have a set of N samples of n dimensional data, so that $X \in \mathbb{R}^{N \times n}$ s.t. each column has 0 mean (we can just shift the data to fulfill this request).
- Principal Component Analysis (PCA) is defined as an orthogonal linear transformation such that the new coordinate system of $\mathbb{R}^n$ satisfies: the i-th component of the coordinate system has the i-th greatest data variance if we project all samples on that component.
- Ideally we are trying to fit a n-ellipsoid into the data. The length of an axis of the ellipsoid represents the variance of data along that axis.
- PCA is often used for dimensionality reduction or encoding: we can project the data on the first $k < n$ principal components.

Principal Component Analysis

Mathematically we can define:

$$w_1 = \underset{\|w\|_2=1}{\arg\max} \|Xw\|_2^2 = \arg\max_w \frac{w^T X^T X w}{w^T w} \tag{2}$$

for the first component. Then for the k-th component we first subtract the first $k-1$ principal component from X

$$\hat{X}_k = X - \sum_{i=1}^{k-1} X w_i w_i^T \tag{3}$$

and finally solving again the similar problem:

$$w_k = \underset{\|w\|_2=1}{\arg\max} \|\hat{X}_k w\|_2^2 = \arg\max_w \frac{w^T \hat{X}_k^T \hat{X}_k w}{w^T w} \tag{4}$$

Robust Principal Component Analysis

- Robust Principal Component Analysis (RPCA) is a generalization of PCA that aims to reduce the sensitivity of PCA to outliers.
- The idea is to find a low-dimensional representation of data cleaned from the sparse outliers that can disturb the PCA process.
- We therefore assume that data X can be represented as $X = L + S$: L has low rank and is the low-dimensional representation of X while S is a sparse matrix consisting of the outlier elements that cannot be captured by the representation.

Robust Principal Component Analysis

- The problem can be addressed as:

$$\min_{L,S} \rho(L) + \lambda \|S\|_0 \tag{5}$$

$$\text{s. t. } \|X - L - S\|_F^2 = 0 \tag{6}$$

where $\rho(\cdot)$ is the rank of a matrix and we used the zero norm.

- This optimization problem is NP-hard and tractable only for small metrices.
- Usually it is substituted by the following problem, which is convex and tractable also for large matrices:

$$\min_{L,S} \|L\|_* + \lambda \|S\|_1 \tag{7}$$

$$\text{s. t. } \|X - L - S\|_F^2 = 0 \tag{8}$$

where $\|\cdot\|_*$ is the nuclear norm i. e. the sum of singular values of a matrix.

https://github.com/AlexThirty/SaMLMfTSA

Thank you!

Subutai Ahmad, Alexander Lavin, Scott Purdy, and Zuha Agha.
Unsupervised real-time anomaly detection for streaming data.
Neurocomputing, 262:134–147, 2017.
Online Real-Time Learning Strategies for Data Streams.

Emmanuel J. Candes, Xiaodong Li, Yi Ma, and John Wright.
Robust principal component analysis?, 2009.

Chong Zhou and Randy C. Paffenroth.
Anomaly detection with robust deep autoencoders.
page 665–674, 2017.