Anomaly Detection

Santiago Velasco-Forero

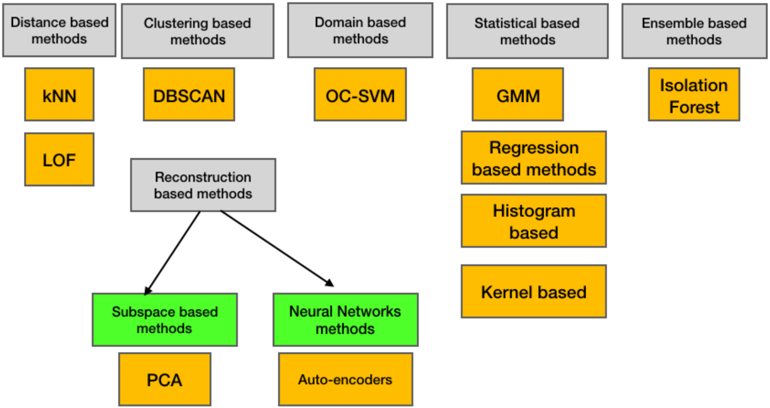
santiago.velasco@mines-paristech.fr ; http://cmm.ensmp.fr/∼velasco

CMM-Centre de Morphologie Mathématique, Mathématiques et Systèmes,
MINES-PARISTECH, FRANCE

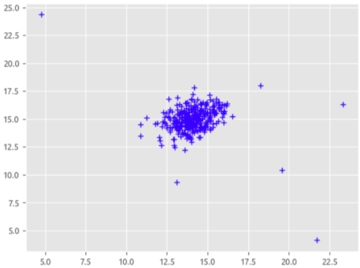
Deep Learning Course 2020

Plan

Taxonomy


Distance based methods
kNN
LOF
Clustering based methods
DBSCAN
Domain based methods
OC-SVM
Statistical based methods
GMM
Regression based methods
Histogram based
Kernel based
Ensemble based methods
Isolation Forest
Reconstruction based methods
Subspace based methods
PCA
Neural Networks methods
Auto-encoders

kNN distance for outlier detection



For an observation $\mathbf{x}$, its distance to its kth nearest neighbor could be viewed as the outlying score. It could be viewed as a way to measure the density. Many kNN detectors are supported:

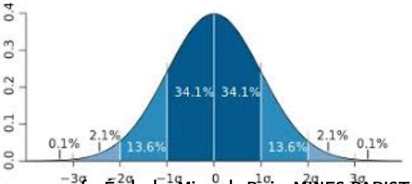
1. largest: use the distance to the kth neighbor as the outlier score.
2. Mean k-NN: use the average of all k neighbors as the outlier score.
3. Many variants...

Z-score Univariate Case

`Z-score` is an important concept in statistics. `Z score` is also called standard score, the procedure is called *standardization*. This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, `Z-score` tells how many standard deviations away a data point is from the mean. For a given univariate samples $X = \{x_1, \ldots, x_\mathcal{N}\}$, the `Z score` is defined by

$$\texttt{Z-score} = \frac{x_i - \bar{x}}{\sigma} \tag{1}$$

where σ and is the standard deviation and mean of the distribution of feature x, respectively, and x_i is the value of the feature x for the ith sample.



How to extend to multivariate case?

- One approach to this problem would be to assume that the D variables are *mutually independent*.
- They may then be standardized so that,, each is normally distributed with zero mean and unit variance.

How to extend to multivariate case?

- One approach to this problem would be to assume that the D variables are *mutually independent*.
- They may then be standardized so that,, each is normally distributed with zero mean and unit variance.
- Using sample estimates of the population parameters for the standarization, the sum of their squares is then asymptotically distributed like chi-square distribution.

How to extend to multivariate case?

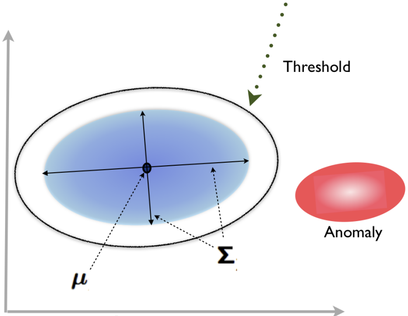
- One approach to this problem would be to assume that the D variables are *mutually independent*.
- They may then be standardized so that,, each is normally distributed with zero mean and unit variance.
- Using sample estimates of the population parameters for the standarization, the sum of their squares is then asymptotically distributed like chi-square distribution.
- In many application p variables are not independent. They are, in fact, very strongly correlated.

Mahalanobis distance

This circumstance suggests the use of Hotelling's T^2 as the appropriate statistic. This statistics is computed by the *sampled Mahalanobis distance*:

$$M(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x} - \boldsymbol{\mu})^T \quad (2)$$

where $\mathbf{x}$ is the row vector whose elements are the observed values of the D variables, $\boldsymbol{\mu}$ is the corresponding mean vector , and where $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of the D variables.



Covariance matrix estimation

The sample covariance matrix of the observations $\mathbf{x}_1, \ldots, \mathbf{x}_{\mathcal{N}} \in \mathbb{R}^D$ is defined by:

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\mathcal{N} - 1} \sum_{i=1}^{\mathcal{N}} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T \tag{3}$$

Where, $\bar{\mathbf{x}}$ denotes the empirical mean of the observations.

Difficulties on high-dimensional spaces

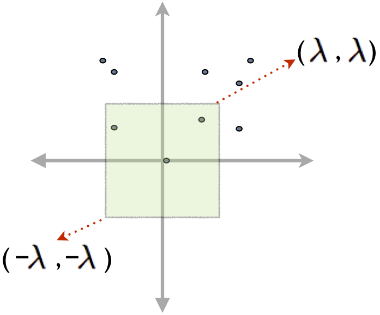
The use of this statistic gave rise to some difficulties.

1. The determinant of the covariance matrix was found to be very small so that the computation of the inverse was inaccurate.
2. The smallness of the determinant of the covariance matrix indicated that the variates were so strongly correlated that some of them could be dropped from consideration.
3. The estimation of $\hat{\boldsymbol{\Sigma}}$ depends of the relation between number of samples and the number of variables.
4. Distance measures in high dimensional spaces has strange behaviours (Curse of Dimensionality).
5. In general, intuition that we have in 3D are hard to extend to high dimensions.

Curse of Dimensionality

Connectivity in high dimensional spaces $C^D(\lambda)$ is the cube centered at the origin in $\mathbb{R}^D$ with side-length 2λ

$$C^D(\lambda) = \{(x_1, \ldots, x_D) | -\lambda \leq x_j \leq \lambda, \quad \forall j\}$$



As $D \to \infty$, $\mathbf{Vol}(C^D(\lambda)) =$

Curse of Dimensionality

As $D \to \infty$, $\textbf{Vol}\left(C^D(\lambda)\right) = (2\lambda \times 2\lambda \times \ldots \times 2\lambda) = (2\lambda)^D$

Curse of Dimensionality

As $D \to \infty$, **Vol** $(C^D(\lambda)) = (2\lambda \times 2\lambda \times \ldots \times 2\lambda) = (2\lambda)^D$

$$\lim_{d\to\infty} \mathbf{Vol}(C^D(\lambda)) = 0 \quad \text{if} \quad \lambda < 1/2,$$
$$\lim_{d\to\infty} \mathbf{Vol}(C^D(\lambda)) = \infty \quad \text{if} \quad \lambda > 1/2,$$
$$\lim_{d\to\infty} \mathbf{Vol}(C^D(\lambda)) = 1 \quad \text{if} \quad \lambda = 1/2$$

$$\mathbf{Diagonal}(C^D(\lambda)) =$$

Curse of Dimensionality

As $D \to \infty$, $\textbf{Vol}\ (C^D(\lambda)) = (2\lambda \times 2\lambda \times \ldots \times 2\lambda) = (2\lambda)^D$

$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = 0 \quad \text{if} \quad \lambda < 1/2,$$
$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = \infty \quad \text{if} \quad \lambda > 1/2,$$
$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = 1 \quad \text{if} \quad \lambda = 1/2$$

$$\textbf{Diagonal}(C^D(\lambda)) = 2\lambda\sqrt{D}$$

$\lambda = 1/2,$ we obtain a cool shape!:

Curse of Dimensionality

As $D \to \infty$, $\textbf{Vol}\ (C^D(\lambda)) = (2\lambda \times 2\lambda \times \ldots \times 2\lambda) = (2\lambda)^D$

$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = 0 \quad \text{if} \quad \lambda < 1/2,$$
$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = \infty \quad \text{if} \quad \lambda > 1/2,$$
$$\lim_{d\to\infty} \textbf{Vol}(C^D(\lambda)) = 1 \quad \text{if} \quad \lambda = 1/2$$

$$\textbf{Diagonal}(C^D(\lambda)) = 2\lambda\sqrt{D}$$
$$\lambda = 1/2, \text{ we obtain a cool shape!:}$$
$$\lim_{d\to\infty} \textbf{Vol}(C^D(1/2)) = 1$$
$$\lim_{d\to\infty} \textbf{Diagonal}(C^D(1/2)) = \infty$$

Using subspaces

- That is apply a linear dimensionality approach and then use a Mahalanobis distance or another method.
- (In this course) A new set of D' variables ($D' \leq D$) is obtained by means of a linear transformation of the original variables.
- Since the transformation is linear the multivariate normality of the new variables follows from that assumed for the original ones.
- $y = (\mathbf{x} - \boldsymbol{\mu})\mathbf{W}$ where $\mathbf{W}$ is a $D \times D'$ rectangular matrix.
  1. Principal component analysis (PCA) : Columns are the properly normalized eigenvectors corresponding to the first D' largest eigenvalues of the sample covariance matrix S.
  2. Negative Principal component analysis (NPCA) : Columns are the properly normalized eigenvectors corresponding to the first D' smallest eigenvalues of the sample covariance matrix S.

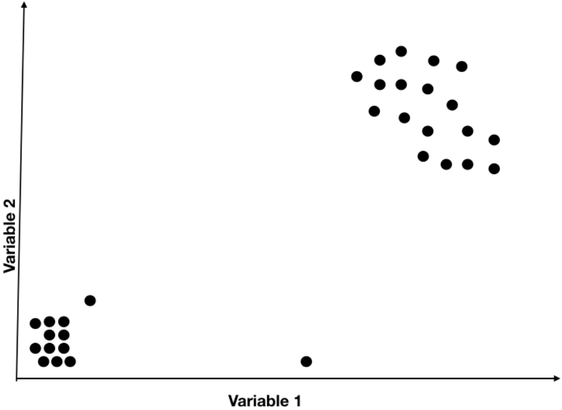
Singular Value Decomposition

The singular value decomposition of an $D \times \mathcal{N}$ real or complex matrix $\mathbf{X}$ is a factorization of the form $\mathbf{U}\Lambda\mathbf{V}^*$, where $\mathbf{U}$ is an $D \times D$ real or complex unitary matrix, Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and $\mathbf{V}$ is an $\mathcal{N} \times \mathcal{N}$ real unitary matrix. If $\mathbf{X}$ is real, $\mathbf{U}$ and $\mathbf{V}^T = \mathbf{V}^*$ are real orthogonal matrices. The diagonal entries $\lambda_i = \Lambda_{ii}$ of Λ are known as the singular values of $\mathbf{X}$. From the decomposition:

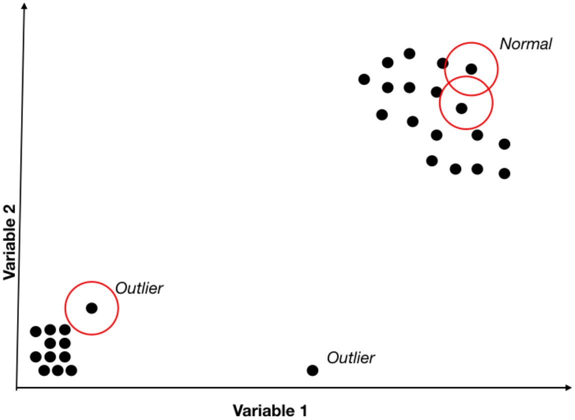
$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^T, \mathbf{X}^T\mathbf{X} = \mathbf{V}\Lambda^2\mathbf{V}^T \tag{4}$$

where Λ contains the positive real eigenvalues of decreasing magnitude and the i-th eigenvalue equals the square of the i-th singular value

Why local outliers?



Why local outliers?



Solutions based on absolute density cannot detect local objetcs

Local Outlier Factor

Consider relative density Let k-$\texttt{dist}(\mathbf{x})$ be the distance of $\mathbf{x}$ to the k-th nearest neighbor. Then the *reachability distance* denoted by `reach-dist` is

$$\texttt{reach-dist}(\mathbf{x}, \mathbf{y}) = \max(k\text{-}\texttt{dist}(\mathbf{y}), \texttt{dist}(\mathbf{x}, \mathbf{y}))$$

Note that `reach-dist` is not symmetric. The Local Reachability Density (LRD) is defined by:

$$\texttt{LRD}_k(\mathbf{x}) = \frac{1}{\frac{\sum_{z \in N_k} \texttt{reach-dist}(\mathbf{x},\mathbf{z})}{|N_k(\mathbf{x})|}}$$

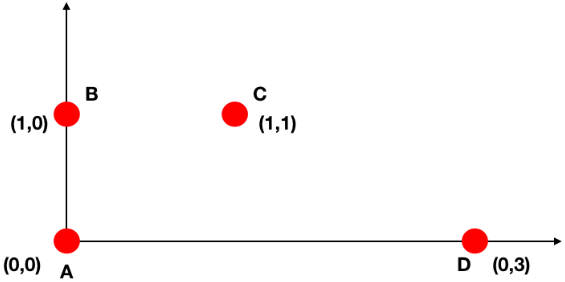
which is the inverse of the mean distance of the local reachability of $\mathbf{x}$ and its neighbors.

Local Outlier Factor

The local outlier factor (LOF) compares the local density with respect to the one from the k-nearest neighbors, *i.e.*

$$\begin{aligned}\mathtt{LOF}_k(\mathbf{x}) &= \frac{\sum_{z \in N_k(\mathbf{x})} \frac{\mathtt{lrd}_k(\mathbf{z})}{\mathtt{lrd}_k(\mathbf{x})}}{|N_k(\mathbf{x})|} \\ &= \frac{\sum_{z \in N_k(\mathbf{x})} \mathtt{lrd}_k(\mathbf{z})}{|N_k(\mathbf{x})| \mathtt{lrd}_k(\mathbf{x})}\end{aligned}$$

Example Four Points (1/3)



Example: Using the Manhattan distance (a.k.a. taxicab metric or L_1 norm, $\forall \mathbf{p}, \mathbf{q} \in \mathbb{R}^d, \mathtt{dist}(\mathbf{p}, \mathbf{q}) = ||\mathbf{p} - \mathbf{q}||_1 = \sum_{i=1}^{d} |\mathbf{p}_i - \mathbf{q}_i|$, calcule the local outlier factor, $\mathtt{LOF}_k(\mathbf{x})$ for $k = 2$.

Example Four Points (2/3)

$\mathtt{dist}(\cdot,\cdot)$	A	B	C	D
A	0	1	2	3
B		0	1	4
C			0	3
D				0

(a) $\mathtt{dist}(\cdot,\cdot)$

	k=1	k=2	k=3
A	1	2	3
B	1	1	4
C	1	2	3
D	3	3	4

(b) k-$\mathtt{dist}(\cdot)$ for $k = 1, 2, 3$

	k=2
A	{B,C}
B	{A,C}
C	{A,B}
D	{A,C}

(c) Set of 2-NN, $N_2(\cdot)$

	A	B	C	D
A		2	2	3
B	1		1	4
C	2	2		3
D	3	4	3	

(d) 2-$\mathtt{reach}$-$\mathtt{dist}(\cdot,\cdot)$

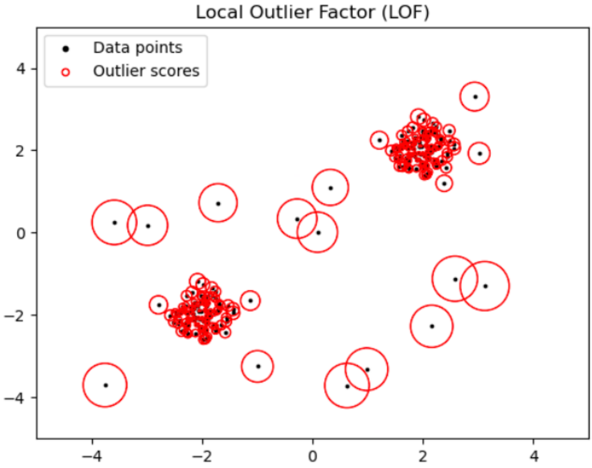
Example Four Points (3/3)

A	2/(1+2)	2/3
B	2/(2+2)	1/2
C	2/(2+1)	2/3
D	2/(3+3)	1/3

(e) $\mathtt{LRD}_2(\mathbf{x})$

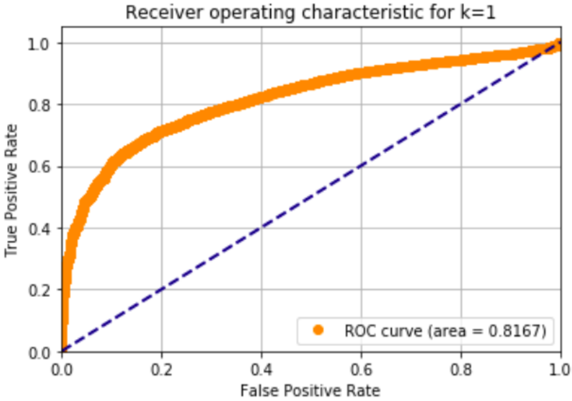
	k=1	**k=2**	**k=3**
A	1	2	3
B	1	1	4
C	1	2	3
D	3	3	4

(f) $\mathtt{LOF}_2$


Local Outlier Factor (LOF)
Data points
Outlier scores
4
2
0
−2
−4
−4
−2
0
2
4

ROC Curve

A receiver operating characteristic curve, or **ROC curve**, is a plot that illustrates the ability of a detector as its discrimination threshold is varied. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings



Finding the best threshold

- The optimal cut off would be where true positive is high and false negative is low