CS685: Data Mining
Anomaly or Outlier Detection

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs685/

1st semester, 2020-21

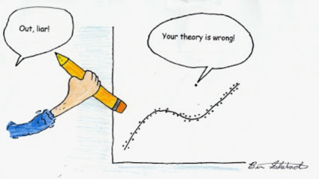
Mon 1030-1200 (online)

Anomaly Detection

- An anomaly or an outlier is an object that is so different from others that it is more likely to be generated from a *separate* process

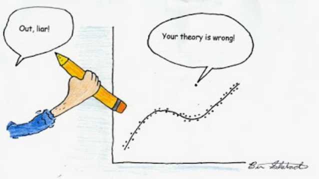
Anomaly Detection

- An anomaly or an outlier is an object that is so different from others that it is more likely to be generated from a *separate* process



Anomaly Detection

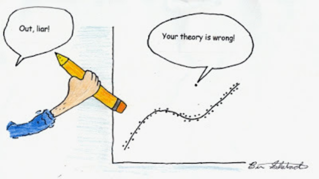
- An anomaly or an outlier is an object that is so different from others that it is more likely to be generated from a *separate* process



- Process of finding such anomalies is called anomaly detection
- Also called outlier detection, deviation detection, exception mining

Anomaly Detection

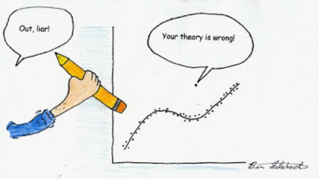
- An anomaly or an outlier is an object that is so different from others that it is more likely to be generated from a *separate* process



- Process of finding such anomalies is called anomaly detection
- Also called outlier detection, deviation detection, exception mining
- Important for
  - Intrusion detection
  - Fraud detection
  - Climate change

Anomaly Detection

- An anomaly or an outlier is an object that is so different from others that it is more likely to be generated from a *separate* process



- Process of finding such anomalies is called anomaly detection
- Also called outlier detection, deviation detection, exception mining
- Important for
  - Intrusion detection
  - Fraud detection
  - Climate change
- Anomalies may be caused by
  - Some other process(es)
  - Wide variations in data
  - Measurement errors

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these
- Contextual outlier
  - Data object is unusual only if context is taken into account

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these
- Contextual outlier
  - Data object is unusual only if context is taken into account
  - Suppose a person spends on average 5000 per month on a credit card
  - Suddenly in a month, she spends 50000
  - May be a contextual anomaly

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these
- Contextual outlier
  - Data object is unusual only if context is taken into account
  - Suppose a person spends on average 5000 per month on a credit card
  - Suddenly in a month, she spends 50000
  - May be a contextual anomaly
  - May not be unusual if spending patterns of all customers are analyzed
  - For example, if it is the festive season
  - Requires contextual analysis

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these
- Contextual outlier
  - Data object is unusual only if context is taken into account
  - Suppose a person spends on average 5000 per month on a credit card
  - Suddenly in a month, she spends 50000
  - May be a contextual anomaly
  - May not be unusual if spending patterns of all customers are analyzed
  - For example, if it is the festive season
  - Requires contextual analysis
- Collective outlier
  - Only a collection is unusual, but not the individual objects

Types of Outliers

- Global outlier
  - A single data object that deviates a lot from other objects
  - Also called point anomaly
  - Most methods find these
- Contextual outlier
  - Data object is unusual only if context is taken into account
  - Suppose a person spends on average 5000 per month on a credit card
  - Suddenly in a month, she spends 50000
  - May be a contextual anomaly
  - May not be unusual if spending patterns of all customers are analyzed
  - For example, if it is the festive season
  - Requires contextual analysis
- Collective outlier
  - Only a collection is unusual, but not the individual objects
  - A particular shipment lost is not an anomaly
  - However, most shipments lost to a particular address is
  - Hard to find

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg (obese)
  - Very uncommon to have someone who is both 3 ft and 90 Kg

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg (obese)
  - Very uncommon to have someone who is both 3 ft and 90 Kg
- Degree to which an object is an anomaly
  - Binary categorization (anomaly or not) may not always be appropriate
  - An unusual network scan may be due to ignorance
  - Not good to shut the user off
  - May use scores or probabilities to denote the *degree* of an anomaly

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg (obese)
  - Very uncommon to have someone who is both 3 ft and 90 Kg
- Degree to which an object is an anomaly
  - Binary categorization (anomaly or not) may not always be appropriate
  - An unusual network scan may be due to ignorance
  - Not good to shut the user off
  - May use scores or probabilities to denote the *degree* of an anomaly
- Evaluation
  - How to evaluate if an identified anomaly is really so

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg (obese)
  - Very uncommon to have someone who is both 3 ft and 90 Kg
- Degree to which an object is an anomaly
  - Binary categorization (anomaly or not) may not always be appropriate
  - An unusual network scan may be due to ignorance
  - Not good to shut the user off
  - May use scores or probabilities to denote the *degree* of an anomaly
- Evaluation
  - How to evaluate if an identified anomaly is really so
- Efficiency
  - Computational time
  - Intrusion detection systems require real time analysis

Issues in Anomaly Detection

- Set of attributes that define an anomaly
  - Not uncommon to have people who are 3 ft (children)
  - Not uncommon to have people who are 90 Kg (obese)
  - Very uncommon to have someone who is both 3 ft and 90 Kg
- Degree to which an object is an anomaly
  - Binary categorization (anomaly or not) may not always be appropriate
  - An unusual network scan may be due to ignorance
  - Not good to shut the user off
  - May use scores or probabilities to denote the *degree* of an anomaly
- Evaluation
  - How to evaluate if an identified anomaly is really so
- Efficiency
  - Computational time
  - Intrusion detection systems require real time analysis
- Importance of labeling anomalies
  - For email filtering, better to classify some spam as normal
  - For fraud detection, better to classify some normal as fraud

Types of Anomaly Detection

- *Supervised:* uses labels for normal and anomalous objects
- *Unsupervised:* no labels
- *Semi-supervised:* Small number of labels or labels only for some normal objects

Types of Anomaly Detection

- *Supervised:* uses labels for normal and anomalous objects
- *Unsupervised:* no labels
- *Semi-supervised:* Small number of labels or labels only for some normal objects
- Three main approaches
  - *Statistical*
  - *Proximity-based*
    - *Distance-based*
    - *Density-based*
  - *Clustering-based*

Statistical Methods

- Statistical methods use a model for normal objects
- Objects that are not likely to be generated from the model are deemed *anomalous*
- Also called *model-based methods*

Statistical Methods

- Statistical methods use a model for normal objects
- Objects that are not likely to be generated from the model are deemed *anomalous*
- Also called *model-based methods*
- Every object will have some probability of being generated from a model
- Uses probability thresholds to determine outliers

Statistical Methods

- Statistical methods use a model for normal objects
- Objects that are not likely to be generated from the model are deemed *anomalous*
- Also called *model-based methods*
- Every object will have some probability of being generated from a model
- Uses probability thresholds to determine outliers
- Can be parametric where the form of the model is assumed to be known (parameters may be learnt)
- Can be non-parametric where the model is constructed based on the data

General Framework

- Build a statistical model (for “normal” objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model

General Framework

- Build a statistical model (for “normal” objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
  - Object is not generated from the model, i.e., it is an outlier

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
    - Object is not generated from the model, i.e., it is an outlier
- Otherwise, *fail to reject*

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
  - Object is not generated from the model, i.e., it is an outlier
- Otherwise, *fail to reject*
- Two standard measures of statistical significance

General Framework

- Build a statistical model (for “normal” objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
  - Object is not generated from the model, i.e., it is an outlier
- Otherwise, *fail to reject*
- Two standard measures of statistical significance
- P-value: Probability that another object has a value as extreme (maximum or minimum) as the one under consideration

General Framework

- Build a statistical model (for “normal” objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
  - Object is not generated from the model, i.e., it is an outlier
- Otherwise, *fail to reject*
- Two standard measures of statistical significance
- P-value: Probability that another object has a value as extreme (maximum or minimum) as the one under consideration
- Z-score: Deviation from the mean, normalized by standard deviation

General Framework

- Build a statistical model (for "normal" objects)
- If necessary, learn its parameters
- Given an object, find its probability of being generated from the model
- If less than a threshold, reject it as *outlier*
- Null hypothesis: The object is generated from the model
- Alternate hypothesis: The object is *not* generated from the model
- For a particular level of significance, find the confidence interval
- If outside the confidence interval, *reject* the null hypothesis
    - Object is not generated from the model, i.e., it is an outlier
- Otherwise, *fail to reject*
- Two standard measures of statistical significance
- P-value: Probability that another object has a value as extreme (maximum or minimum) as the one under consideration
- Z-score: Deviation from the mean, normalized by standard deviation
- P-value is more appropriate

Univariate Normal Distribution

- Model $N(\mu, \sigma)$
- μ and σ^2 are estimated as *sample mean* and *sample variance*

$$\widehat{\mu} = \sum_{i=1}^{n} x_i / n$$

$$\widehat{\sigma^2} = \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 / (n-1)$$

Univariate Normal Distribution

- Model $N(\mu, \sigma)$
- μ and σ^2 are estimated as *sample mean* and *sample variance*

$$\widehat{\mu} = \sum_{i=1}^{n} x_i / n$$

$$\widehat{\sigma^2} = \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 / (n-1)$$

- A value c in $N(\mu, \sigma)$ is equivalent to $c' = (c - \mu)/\sigma$ in $N(0, 1)$
- For a particular value c', the standard normal distribution chart gives the probability of $\alpha = P(|x| \geq c')$
- Example: $\alpha = 0.3173$ for $c' = 1$, $\alpha = 0.0027$ for $c' = 3$

Univariate Normal Distribution

- Model $N(\mu, \sigma)$
- μ and σ^2 are estimated as *sample mean* and *sample variance*

$$\widehat{\mu} = \sum_{i=1}^{n} x_i / n$$

$$\widehat{\sigma^2} = \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 / (n-1)$$

- A value c in $N(\mu, \sigma)$ is equivalent to $c' = (c - \mu)/\sigma$ in $N(0, 1)$
- For a particular value c', the standard normal distribution chart gives the probability of $\alpha = P(|x| \geq c')$
- Example: $\alpha = 0.3173$ for $c' = 1$, $\alpha = 0.0027$ for $c' = 3$
- Hence, c is outlier with probability $1 - \alpha$ (*two-tailed*) or $1 - (\alpha/2)$ (*one-tailed*)

Univariate Normal Distribution

- Model $N(\mu, \sigma)$
- μ and σ^2 are estimated as *sample mean* and *sample variance*

$$\widehat{\mu} = \sum_{i=1}^{n} x_i / n$$

$$\widehat{\sigma^2} = \sum_{i=1}^{n} (x_i - \widehat{\mu})^2 / (n-1)$$

- A value c in $N(\mu, \sigma)$ is equivalent to $c' = (c - \mu)/\sigma$ in $N(0, 1)$
- For a particular value c', the standard normal distribution chart gives the probability of $\alpha = P(|x| \geq c')$
- Example: $\alpha = 0.3173$ for $c' = 1$, $\alpha = 0.0027$ for $c' = 3$
- Hence, c is outlier with probability $1 - \alpha$ (*two-tailed*) or $1 - (\alpha/2)$ (*one-tailed*)
- Z-score is simply $c' = (c - \mu)/\sigma$

Multivariate Normal Distribution

- Model has mean μ and covariance matrix Σ
- For d-dimensional data, μ is $d \times 1$ and Σ is $d \times d$

Multivariate Normal Distribution

- Model has mean μ and covariance matrix Σ
- For d-dimensional data, μ is $d \times 1$ and Σ is $d \times d$
- Uses (square of) Mahalanobis distance from an object x to the mean

$$M(x, \mu) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

Multivariate Normal Distribution

- Model has mean μ and covariance matrix Σ
- For d-dimensional data, μ is $d \times 1$ and Σ is $d \times d$
- Uses (square of) Mahalanobis distance from an object x to the mean

$$M(x, \mu) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

- Mahalanobis distance becomes a univariate variable
- It takes into account the covariances

Multivariate Data Distributions

- Uses chi-square statistic

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold
- A sequence of values from a particular alphabet set
- Expected probability of each symbol is from a multinomial distribution $p_1, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold
- A sequence of values from a particular alphabet set
- Expected probability of each symbol is from a multinomial distribution $p_1, \dots, p_k$ where $\sum_{i=1}^{k} p_i = 1$
- For a subsequence of length ℓ, expected probabilities of symbols are

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold
- A sequence of values from a particular alphabet set
- Expected probability of each symbol is from a multinomial distribution $p_1, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$
- For a subsequence of length ℓ, expected probabilities of symbols are $p_i \times \ell$
- Using observed probabilities, chi-square value can be computed

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold
- A sequence of values from a particular alphabet set
- Expected probability of each symbol is from a multinomial distribution $p_1, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$
- For a subsequence of length ℓ, expected probabilities of symbols are $p_i \times \ell$
- Using observed probabilities, chi-square value can be computed
- Subsequences of different lengths follow the *same* chi-square distribution

Multivariate Data Distributions

- Uses chi-square statistic
- Follows chi-square distribution
- Characterized by degrees of freedom
- Can use p-value to test for a level of significance
- Higher the chi-square value, more unusual it is
- May just use a threshold
- A sequence of values from a particular alphabet set
- Expected probability of each symbol is from a multinomial distribution $p_1, \ldots, p_k$ where $\sum_{i=1}^{k} p_i = 1$
- For a subsequence of length ℓ, expected probabilities of symbols are $p_i \times \ell$
- Using observed probabilities, chi-square value can be computed
- Subsequences of different lengths follow the *same* chi-square distribution
- Degrees of freedom $= k - 1$
- May simply identify subsequences with large chi-square values

Mixture Models

- Uses more than one model $f_1(\theta_1), \ldots, f_n(\theta_n)$
- Final probability of an object is

$$P(x|\theta_1, \ldots, \theta_n) = w_1 f_1(x; \theta_1) + \cdots + w_n f_n(x; \theta_n)$$

- w_i's are the weights of the models (generally the same)
- Set cutoffs based on these final probabilities

Mixture Models

- Uses more than one model $f_1(\theta_1), \ldots, f_n(\theta_n)$
- Final probability of an object is

$$P(x|\theta_1, \ldots, \theta_n) = w_1 f_1(x; \theta_1) + \cdots + w_n f_n(x; \theta_n)$$

- w_i's are the weights of the models (generally the same)
- Set cutoffs based on these final probabilities
- Parameters of the models are learnt using the EM algorithm
- Models are generally assumed to be Gaussian

Mixture Models

- Uses more than one model $f_1(\theta_1), \ldots, f_n(\theta_n)$
- Final probability of an object is

$$P(x|\theta_1, \ldots, \theta_n) = w_1 f_1(x; \theta_1) + \cdots + w_n f_n(x; \theta_n)$$

- w_i's are the weights of the models (generally the same)
- Set cutoffs based on these final probabilities
- Parameters of the models are learnt using the EM algorithm
- Models are generally assumed to be Gaussian
- If model parameters are not known, for anomaly detection, a simpler approach can be used
- Assume all objects to be in "normal" class to start with
- Learn the parameters

Anomaly Detection using Two Models

- Assume two classes, "normal" M and "anomalous" A
- Suppose λ is the fraction of expected number of outliers
- Probability of an object is then

$$P(x) = (1 - \lambda).M(x) + \lambda.A(x)$$

Anomaly Detection using Two Models

- Assume two classes, "normal" M and "anomalous" A
- Suppose λ is the fraction of expected number of outliers
- Probability of an object is then

$$P(x) = (1 - \lambda).M(x) + \lambda.A(x)$$

- Measure the overall likelihood $L(D)$ of the data at any time
- All objects are in class M to start with
- If $L(D)$ changes (increases) by more than a threshold τ, then identify x as an outlier
- Transfer x from M to A
- Continue till no change

Anomaly Detection using Two Models

- Assume two classes, "normal" M and "anomalous" A
- Suppose λ is the fraction of expected number of outliers
- Probability of an object is then

$$P(x) = (1 - \lambda).M(x) + \lambda.A(x)$$

- Measure the overall likelihood $L(D)$ of the data at any time
- All objects are in class M to start with
- If $L(D)$ changes (increases) by more than a threshold τ, then identify x as an outlier
- Transfer x from M to A
- Continue till no change
- Order dependent

Anomaly Detection using Two Models

- Assume two classes, "normal" M and "anomalous" A
- Suppose λ is the fraction of expected number of outliers
- Probability of an object is then

$$P(x) = (1 - \lambda).M(x) + \lambda.A(x)$$

- Measure the overall likelihood $L(D)$ of the data at any time
- All objects are in class M to start with
- If $L(D)$ changes (increases) by more than a threshold τ, then identify x as an outlier
- Transfer x from M to A
- Continue till no change
- Order dependent
- Tends to classify low probability objects from M as outliers

Non-Parametric Methods

- May use histograms
- Probability is height of bin
- Bins with probability less than threshold are outliers

Non-Parametric Methods

- May use histograms
- Probability is height of bin
- Bins with probability less than threshold are outliers
- Choosing boundaries of bins is an important issue

Non-Parametric Methods

- May use histograms
- Probability is height of bin
- Bins with probability less than threshold are outliers
- Choosing boundaries of bins is an important issue
- Too narrow bins identify many normal objects as outliers
- Too wide bins identify many outliers as normal objects

Non-Parametric Methods

- May use histograms
- Probability is height of bin
- Bins with probability less than threshold are outliers
- Choosing boundaries of bins is an important issue
- Too narrow bins identify many normal objects as outliers
- Too wide bins identify many outliers as normal objects
- P-value may be estimated from histograms

Non-Parametric Methods

- May use histograms
- Probability is height of bin
- Bins with probability less than threshold are outliers
- Choosing boundaries of bins is an important issue
- Too narrow bins identify many normal objects as outliers
- Too wide bins identify many outliers as normal objects
- P-value may be estimated from histograms
- May also use kernels to estimate probability density functions
- Uses the idea of influence functions to model

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier
- Very sensitive to k

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier
- Very sensitive to k
- If k is small, set of nearby outliers remain undetected

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier
- Very sensitive to k
- If k is small, set of nearby outliers remain undetected
- If k is large, small clusters get identified as outliers

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier
- Very sensitive to k
- If k is small, set of nearby outliers remain undetected
- If k is large, small clusters get identified as outliers
- Can use a range r instead
- A point is an outlier if there are $< \pi$ points in its r-neighborhood

Distance-Based Methods

- For each point, find distance to k^{th} neighbor
- If this k^{th} distance $> \rho$, then it is an outlier
- Very sensitive to k
- If k is small, set of nearby outliers remain undetected
- If k is large, small clusters get identified as outliers
- Can use a range r instead
- A point is an outlier if there are $< \pi$ points in its r-neighborhood
- Sensitive to π and r

Density-Based Methods

- Use a range r to normalize density
- Density of a point is the number of neighbors in its r-neighborhood
- A point is an outlier if density $< \delta$

Density-Based Methods

- Use a range r to normalize density
- Density of a point is the number of neighbors in its r-neighborhood
- A point is an outlier if density $< \delta$
- Sensitive to δ and r

Density-Based Methods

- Use a range r to normalize density
- Density of a point is the number of neighbors in its r-neighborhood
- A point is an outlier if density $< \delta$
- Sensitive to δ and r
- Density can also be measured as the inverse of the distance to k neighbors
- Sensitive to k

Density-Based Methods

- Use a range r to normalize density
- Density of a point is the number of neighbors in its r-neighborhood
- A point is an outlier if density $< \delta$
- Sensitive to δ and r
- Density can also be measured as the inverse of the distance to k neighbors
- Sensitive to k
- Fails when dataset has regions of varying density

Relative Density

- A point is an outlier if it is much sparser than its neighbors
- Relative density with respect to neighbors

Relative Density

- A point is an outlier if it is much sparser than its neighbors
- Relative density with respect to neighbors
- *Local density* of x is estimated as the distance at which x can be reached *from* its neighbors
- Distance $dist_k(x)$ of a point x is largest distance from points in the k-neighborhood $N_k(x)$
- Reachability distance of x from y is

$$rd(x, y) = \max\{dist_k(y), dist(x, y)\}$$

- It is the distance of x from y but y should at least reach k neighbors

Relative Density

- A point is an outlier if it is much sparser than its neighbors
- Relative density with respect to neighbors
- *Local density* of x is estimated as the distance at which x can be reached *from* its neighbors
- Distance $dist_k(x)$ of a point x is largest distance from points in the k-neighborhood $N_k(x)$
- Reachability distance of x from y is

$$rd(x, y) = \max\{dist_k(y), dist(x, y)\}$$

- It is the distance of x from y but y should at least reach k neighbors
- Local reachability distance of x is

$$lrd_k(x) = 1 \Bigg/ \left(\frac{\sum_{y \in N_k(x)} rd(x, y)}{|N_k(x)|} \right)$$

- It is *inverse* of average reachability distance from k-neighbors of x

Local Outlier Factor

- Local relative density or local outlier factor is average of *ratio* of local reachability distances of neighbors

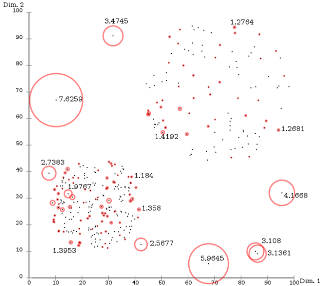
$$lof_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{|N_k(x)|}$$

Local Outlier Factor

- Local relative density or local outlier factor is average of *ratio* of local reachability distances of neighbors

$$lof_k(x) = \frac{\sum_{y \in N_k(x)} \frac{lrd_k(y)}{lrd_k(x)}}{|N_k(x)|}$$

- Higher the outlier score, the more chance than it is an outlier
  - Around 1 is normal
  - Much greater than 1 indicates sparseness, i.e., *outlier*
  - Much less than 1 indicates inlier, i.e., denser
- Choose top-t or greater than a threshold



Clustering-Based Methods

- Some clustering algorithms directly identify outliers

Clustering-Based Methods

- Some clustering algorithms directly identify outliers
- Can use distance of a point to the cluster centre as *outlier score*
- Can also use fitness to the cluster model

Clustering-Based Methods

- Some clustering algorithms directly identify outliers
- Can use distance of a point to the cluster centre as *outlier score*
- Can also use fitness to the cluster model
- Small set of points forming a sparse cluster by itself may also be outliers

Clustering-Based Methods

- Some clustering algorithms directly identify outliers
- Can use distance of a point to the cluster centre as *outlier score*
- Can also use fitness to the cluster model
- Small set of points forming a sparse cluster by itself may also be outliers
- May also use classification methods where normal and anomalous classes are known

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset
  - May be hard to identify the context exactly

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset
  - May be hard to identify the context exactly
- Collective outliers

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset
  - May be hard to identify the context exactly
- Collective outliers
  - Identify small clusters and detect all as outliers

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset
  - May be hard to identify the context exactly
- Collective outliers
  - Identify small clusters and detect all as outliers
  - Build graph of data points and identify anomalous (connected) subgraphs
  - Can be very time consuming

Contextual and Collective Outliers

- Contextual outliers
  - Context must be identified
  - Analysis is done by treating only the contextual data as the dataset
  - May be hard to identify the context exactly
- Collective outliers
  - Identify small clusters and detect all as outliers
  - Build graph of data points and identify anomalous (connected) subgraphs
  - Can be very time consuming
  - For a long sequence, can use chi-square statistic to identify anomalous subsequences
  - Overlapping subsequences having large chi-squares may be collective outliers

Angle-Based Methods

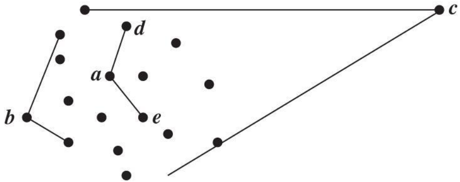
- For high dimensions, proximity (e.g., distance or density) tends to become uniform

Angle-Based Methods

- For high dimensions, proximity (e.g., distance or density) tends to become uniform
- When looked from a distance, all angles are similar
- When within the crowd, angles vary widely

Angle-Based Methods

- For high dimensions, proximity (e.g., distance or density) tends to become uniform
- When looked from a distance, all angles are similar
- When within the crowd, angles vary widely
- For each point, measure its angle with every other pair of points
- Compute the variance of these angles
- Arrange the points according to this variance
- The one with the least variance is the most likely outlier
- Pick top-k or below a threshold



Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall

Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall
- Impose an *equi-depth* grid on the entire high dimensional data
- Each dimension is partitioned into ϕ ranges, where each range contains a fraction f of the data ($f = 1/\phi$)

Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall
- Impose an *equi-depth* grid on the entire high dimensional data
- Each dimension is partitioned into ϕ ranges, where each range contains a fraction f of the data ($f = 1/\phi$)
- Equi-depth partitions help capture the locality of data

Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall
- Impose an *equi-depth* grid on the entire high dimensional data
- Each dimension is partitioned into ϕ ranges, where each range contains a fraction f of the data ($f = 1/\phi$)
- Equi-depth partitions help capture the locality of data
- A k-dimensional cube formed using k ranges is expected to contain f^k fraction of the total points
- If the dataset has a total of n points, the standard deviation of number of points in a k-dimensional region is $\sqrt{f^k.(1-f^k).n}$

Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall
- Impose an *equi-depth* grid on the entire high dimensional data
- Each dimension is partitioned into ϕ ranges, where each range contains a fraction f of the data ($f = 1/\phi$)
- Equi-depth partitions help capture the locality of data
- A k-dimensional cube formed using k ranges is expected to contain f^k fraction of the total points
- If the dataset has a total of n points, the standard deviation of number of points in a k-dimensional region is $\sqrt{f^k.(1-f^k).n}$
- For a particular k-dimensional cube C having n_C points, its sparsity coefficient is defined as

$$S_C = \frac{n_C - f^k.n}{\sqrt{f^k.(1-f^k).n}}$$

Outliers in Subspaces

- Outliers may exist *only* in subspaces, but not overall
- Impose an *equi-depth* grid on the entire high dimensional data
- Each dimension is partitioned into ϕ ranges, where each range contains a fraction f of the data ($f = 1/\phi$)
- Equi-depth partitions help capture the locality of data
- A k-dimensional cube formed using k ranges is expected to contain f^k fraction of the total points
- If the dataset has a total of n points, the standard deviation of number of points in a k-dimensional region is $\sqrt{f^k.(1-f^k).n}$
- For a particular k-dimensional cube C having n_C points, its sparsity coefficient is defined as

$$S_C = \frac{n_C - f^k.n}{\sqrt{f^k.(1-f^k).n}}$$

- If $S_C < 0$, then there are less points than expected
- Lower the S_C, sparser the cell and, therefore, more likely that it contains outliers