Outlier Analysis

Ngô Xuân Bách
Bài giảng của DSLab
Viện nghiên cứu cao cấp về Toán (VIASM)

Hanoi, June 2018

Vietnam Institute for Advanced Study in Mathematics

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study

References

- Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.
- Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2011.
- D.M. Hawkins. *Identification of Outliers*. Monographs on Statistics and Applied Probability, 1980.
- Sanjay Chawla, Varun Chandola. Anomaly Detection: A Tutorial.
- Proceedings of International Conferences
  - ACM International Conference on Knowledge Discovery and Data Mining (KDD).
  - IEEE International Conference on Data Mining (ICDM).

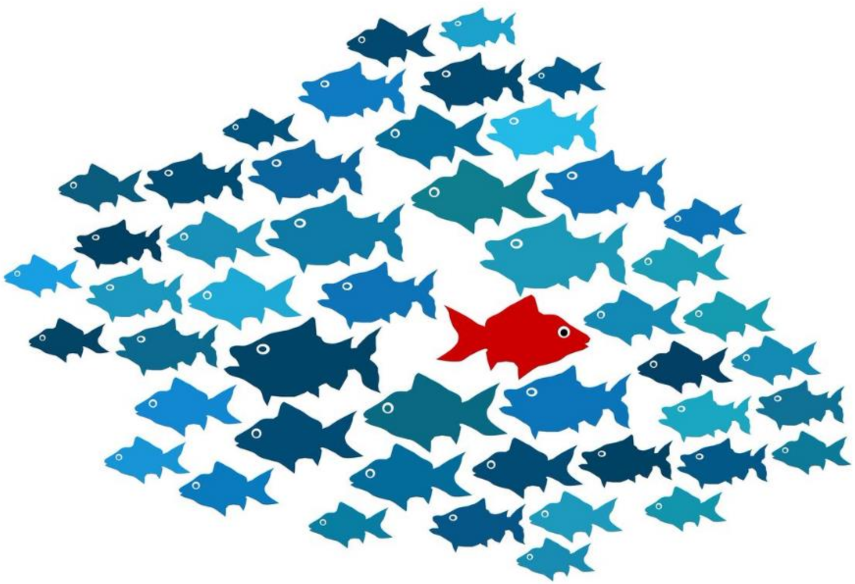
Content

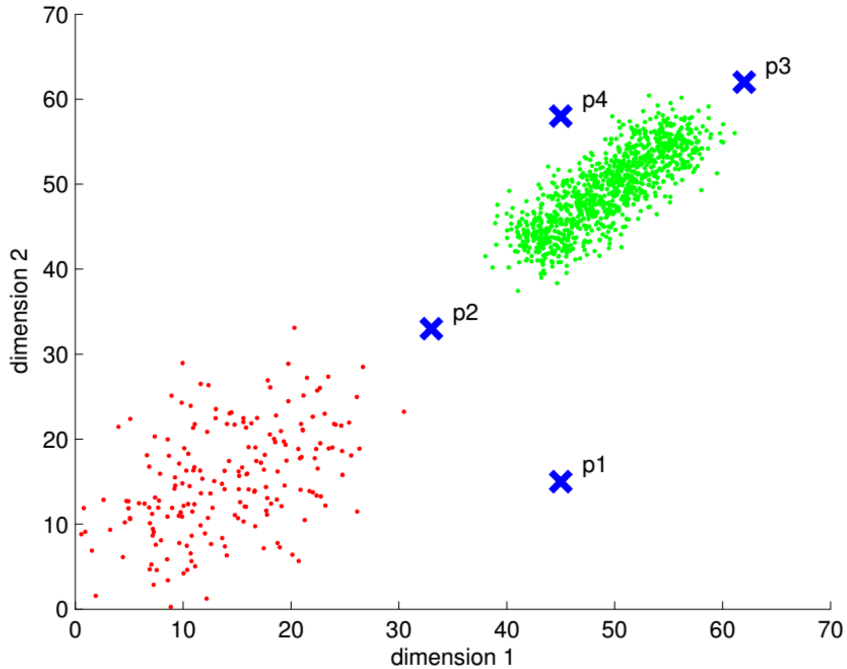
- Part I: Introduction
  - Outliers and Outlier Analysis
  - Some Applications
  - Key Models for Outlier Analysis
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study

Outliers

- An outlier is a data point that is very different from most of the remaining data
- Hawkins: *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."* *(Howkins, 1980)*

Outliers

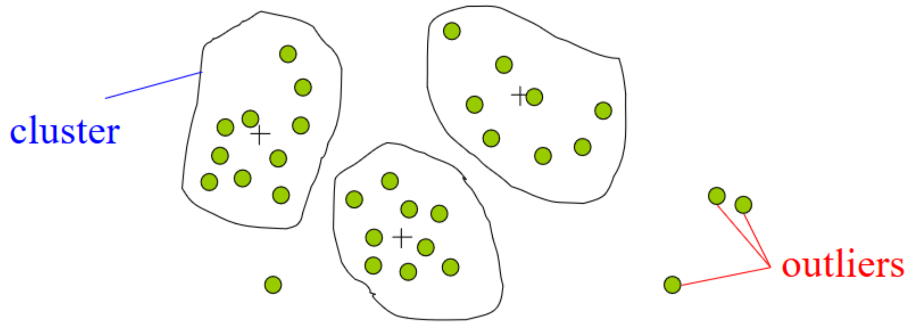




Vietnam Institute for Advanced Study in Mathematics

Outlier Detection

- Outlier detection is the process of detecting and subsequently excluding outliers from a given set of data
- Outlier detection vs. Clustering
  - Clustering: find groups of data points that are similar
  - Outlier detection: detect individual data points that are different from the remaining data



Terminologies

- Outliers = Anomalies = Abnormalities = Deviants
- Outlier Analysis = Outlier Detection = Anomaly Detection

Challenges

- Outliers are different and depend on problems
  - Depend on types of data
- Difficult to collect labeled data
  - Most outlier analysis methods are unsupervised
  - Difficult to validate

Some Applications

- Data cleaning
  - Outlier detection methods are useful for removing noise data
- Quality control and fault detection
- Fraud detection
  - Credit card fraud or insurance transactions
- Web log analytics
  - The anomalies in user behaviors may be determined with the use of Web log analytics
- Network intrusion detection
- Earth science applications
  - Detect unusual changes in the climate, or important events, such as the detection of hurricanes

Key Models for Outlier Analysis

- Most outlier detection methods create a model of normal patterns
  - Outliers are data points that do not naturally fit within this normal model
- The "outlierness" of a data point is quantified by a numeric value, known as the outlier score
- Output of outlier detection methods
  - Real-valued outlier score: quantifies the tendency for a data point to be considered an outlier
  - Binary label: whether or not a data point is an outlier

Key Models for Outlier Analysis

- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Outlier Detection
- Information-Theoretic Models

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

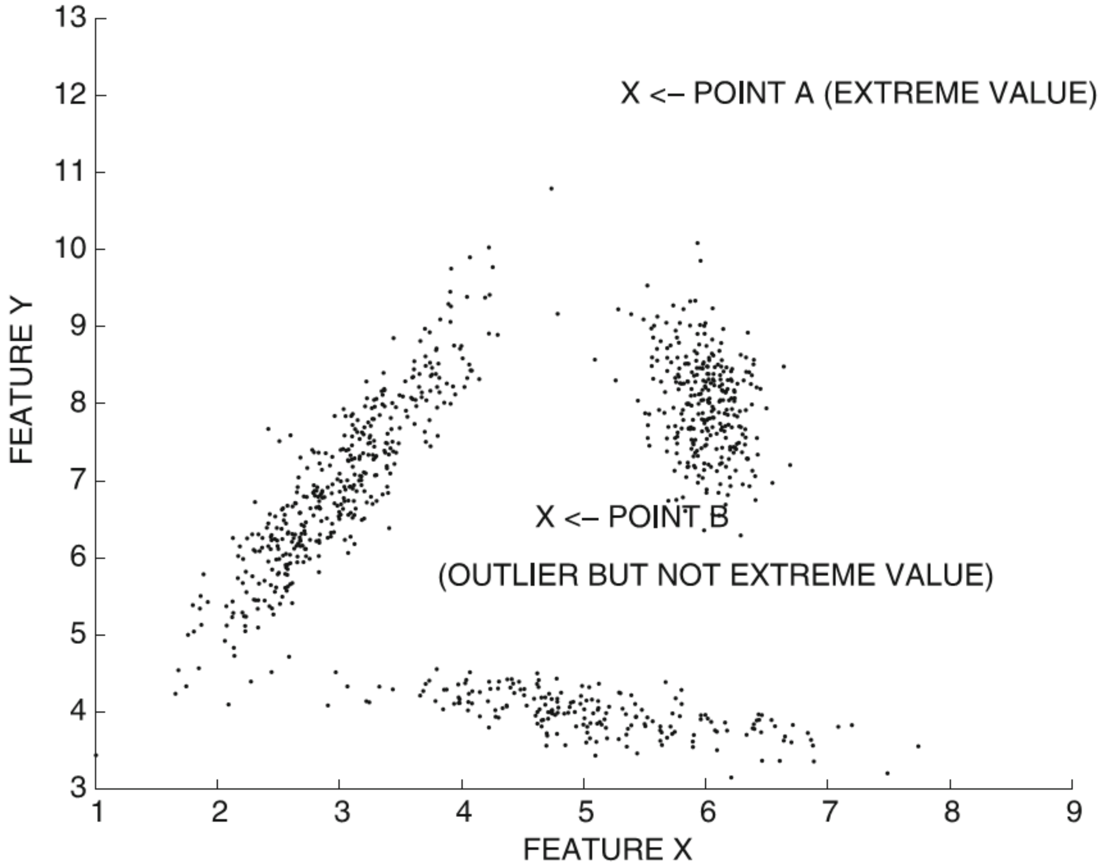
Content

Extreme Value Analysis

- Data point lying at one end of a probability distribution
  - Tails of the probability distribution
- Extreme values are specialized types of outliers
  - All extreme values are outliers, but the reverse may not be true
- Example of univariate extreme values
  - {1,3,3,3,50,97,97,97,100}
  - 1 and 100: extreme values ➡ outliers
  - 50 is the mean of the data set ➡ not an extreme value
  - 50 is the most isolated point ➡ outlier from a generative perspective

Extreme Value Analysis

- Example of multivariate extreme values



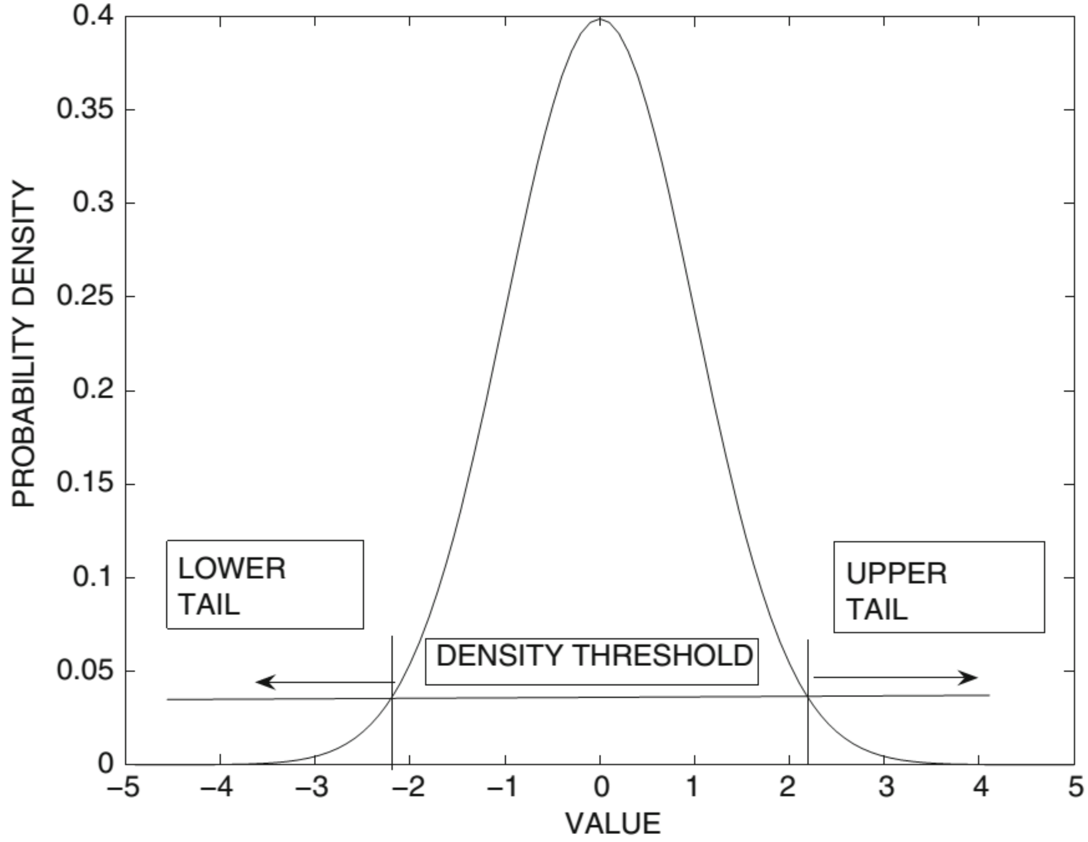
Vietnam Institute for Advanced Study in Mathematics

Univariate Extreme Value Analysis

- How is the "tail" of a distribution defined?
  - The upper tail: all extreme values larger than a particular threshold
  - The lower tail: all extreme values lower than a particular threshold
- Consider the density distribution $f_X(x)$
  - The tail may be defined as the two extreme regions of the distribution for which $f_X(x) \leq \theta$, for some user defined threshold θ

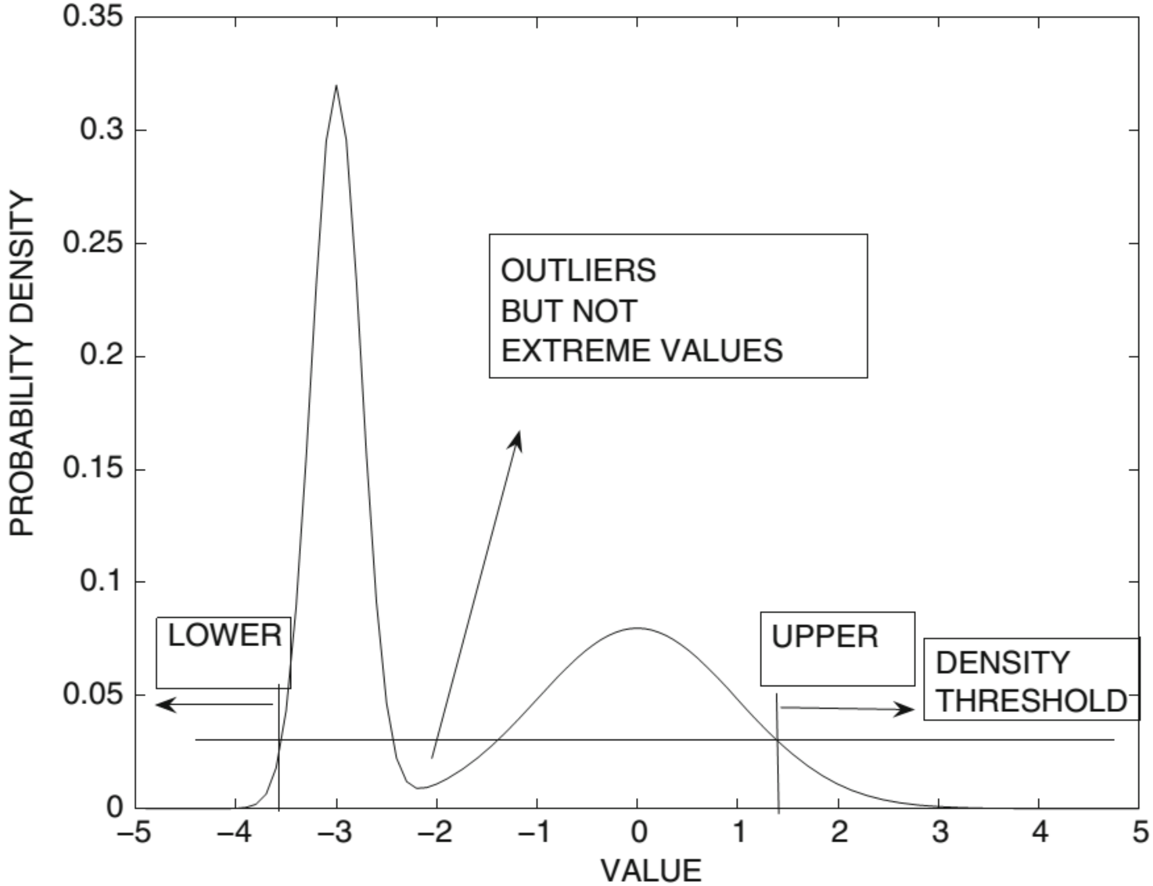
Univariate Extreme Value Analysis

- Symmetric distribution



Univariate Extreme Value Analysis

- Asymmetric distribution



Univariate Extreme Value Analysis

- The most commonly used model for quantifying the tail probability is the normal distribution

$$f_X(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2 \cdot \sigma^2}}$$

  - μ: mean
  - σ: standard deviation
  - In some application scenarios, μ and σ may be known through prior domain knowledge
  - When a large number of data samples is available, μ and σ may be estimated very accurately

Univariate Extreme Value Analysis

- Compute the Z-value for a random variable

$$z_i = \frac{(x_i - \mu)}{\sigma}$$

  - Large positive values of z_i correspond to the upper tail
  - Large negative values correspond to the lower tail
- We have

$$E[z_i] = \frac{E[x_i - \mu]}{\sigma} = \frac{E[x_i] - \mu}{\delta} = 0$$

$$var(z_i) = E\left[z_i^2\right] - E[z_i]^2 = \frac{E[(x_i - \mu)^2]}{\delta^2} = 1$$

Univariate Extreme Value Analysis

- Therefore

$$f_X(z_i) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot e^{\frac{-z_i^2}{2}}$$

- A rule of thumb
  - If $|z_i| > 3$, x_i is considered extreme value
  - The cumulative area inside the tail can be shown to be less than 0.01% for the normal distribution

Multivariate Extreme Values

- Tails are defined for univariate distributions
  - Extreme regions with probability density less than a particular threshold
  - An analogous concept can also be defined for multivariate distributions
- A multivariate Gaussian model is used
  - The corresponding parameters are estimated in a data-driven manner

Multivariate Extreme Values

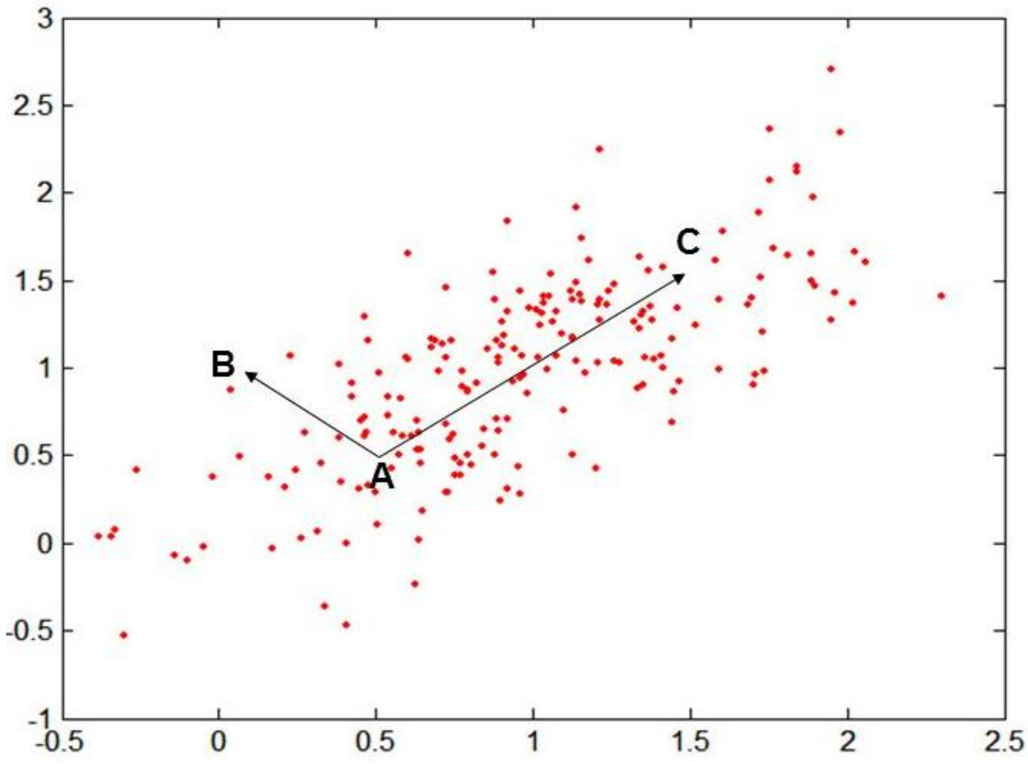
- The probability distribution $f(\bar{X})$ for a d-dimensional data point $\bar{X}$

$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \Sigma^{-1} (\overline{X} - \overline{\mu})^T}$$

  - $\bar{\mu}$: d-dimensional mean vector of the data set
  - Σ: be its $d \times d$ covariance matrix
  - $\Sigma[\mathrm{i}, \mathrm{j}]$: covariance between the dimensions i and j
  - $|\Sigma|$: determinant of the covariance matrix
- $Maha(\bar{X}, \bar{\mu}, \Sigma)$ represents the Mahalanobis distance

$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma|} \cdot (2 \cdot \pi)^{(d/2)}} \cdot e^{-\frac{1}{2} \cdot Maha(\overline{X}, \overline{\mu}, \Sigma)^2}$$

Mahalanobis Distance



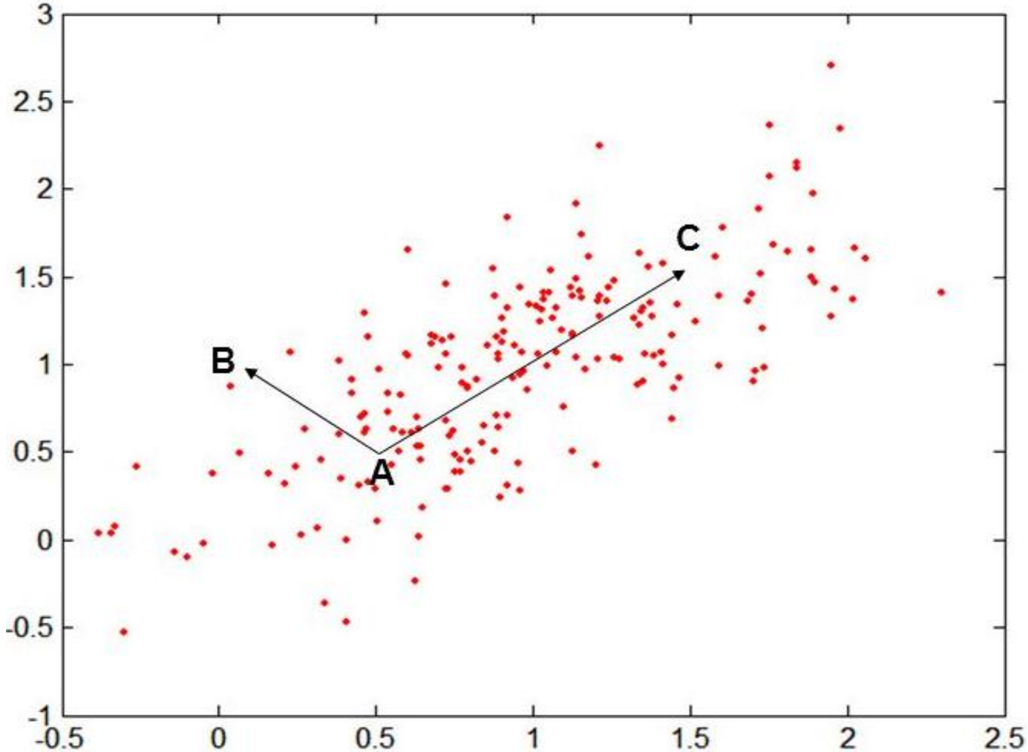
Covariance matrix

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A = (0.5 , 0.5)
B = (0 , 1)
C = (1.5 , 1.5)

$$Maha(A, B) = ?$$
$$Maha(A, C) = ?$$

Mahalanobis Distance



Covariance matrix

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A = (0.5 , 0.5)

B = (0 , 1)

C = (1.5 , 1.5)

Inverse matrix

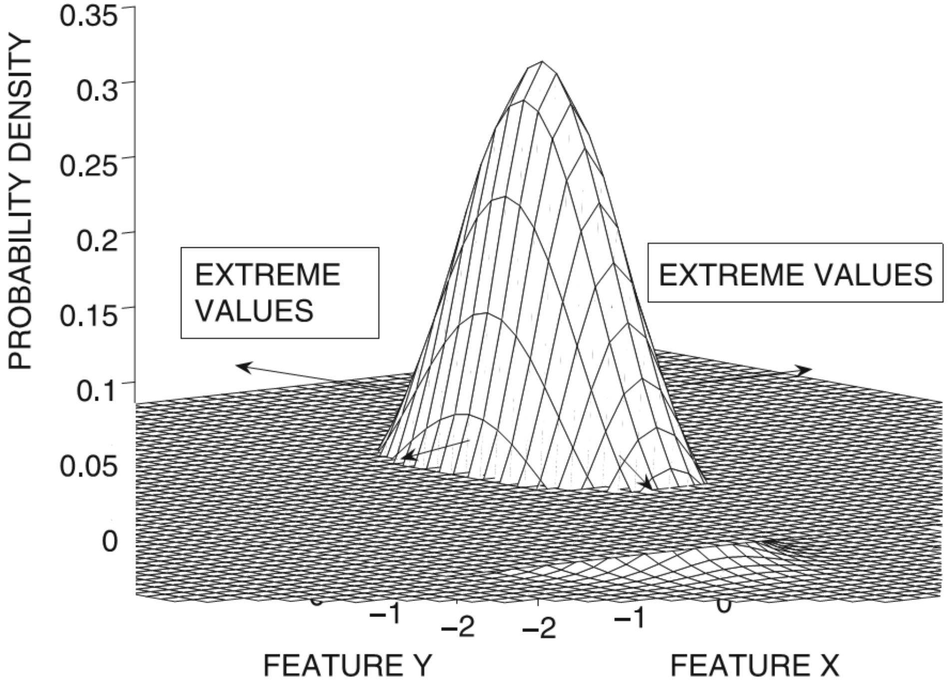
$$\Sigma^{-1} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}$$

$$Maha(A, B) = \sqrt{(A - B)\Sigma^{-1}(A - B)^T} = \sqrt{5}$$

$$Maha(A, C) = \sqrt{(A - C)\Sigma^{-1}(A - C)^T} = \sqrt{4}$$

Multivariate Extreme Values

- For $f(\bar{X})$ less than a particular threshold
  - $Maha(.)$ needs to be larger than a threshold
  - $Maha(.)$ can be used as an extreme-value score



Depth-Based Method

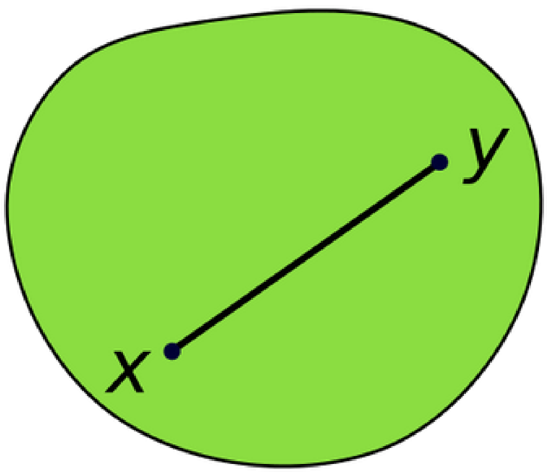
- Convex set
  - If S is a convex set in n-dimensional space, and $u_1, \ldots, u_r$ are r n-dimensional vectors in S $(r > 1)$, then every convex combination of $u_1, \ldots, u_r$ are also in S

$$\sum_{k=1}^{r} \lambda_k u_k \in S$$

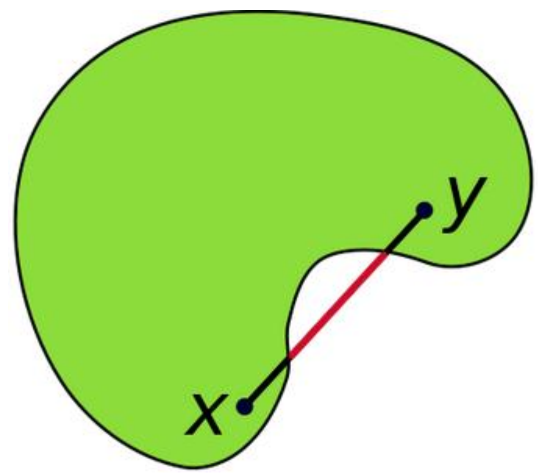
for any nonnegative numbers $\lambda_1, \ldots, \lambda_r$ such that $\lambda_1 + \cdots + \lambda_r = 1$

Depth-Based Method

- Examples of convex and non-convex sets



Convex set



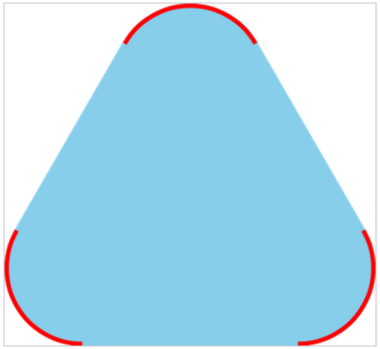
Non-convex set

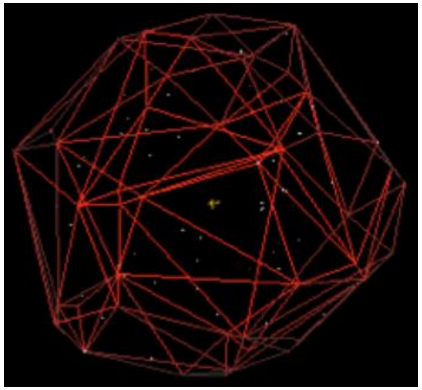
Depth-Based Method

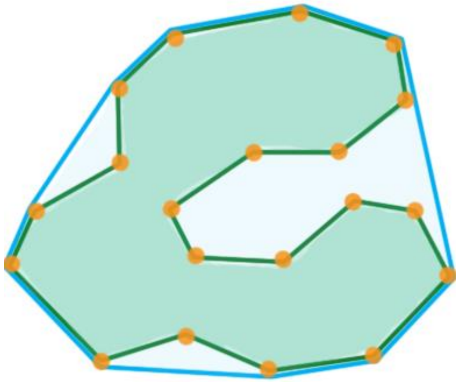
- Convex hull
  - Convex hull of a set X of points the is the smallest convex set that contains X
  - Convex hull may be defined as the intersection of all convex sets containing X or as the set of all convex combinations of points in X

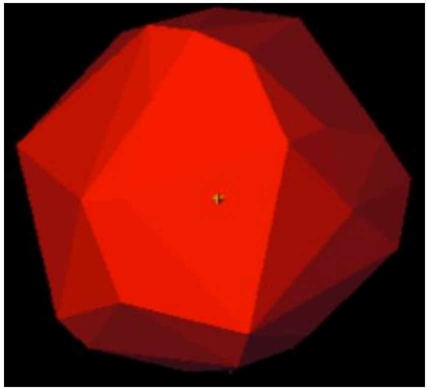
Depth-Based Method

- Examples of convex hulls









Depth-Based Method

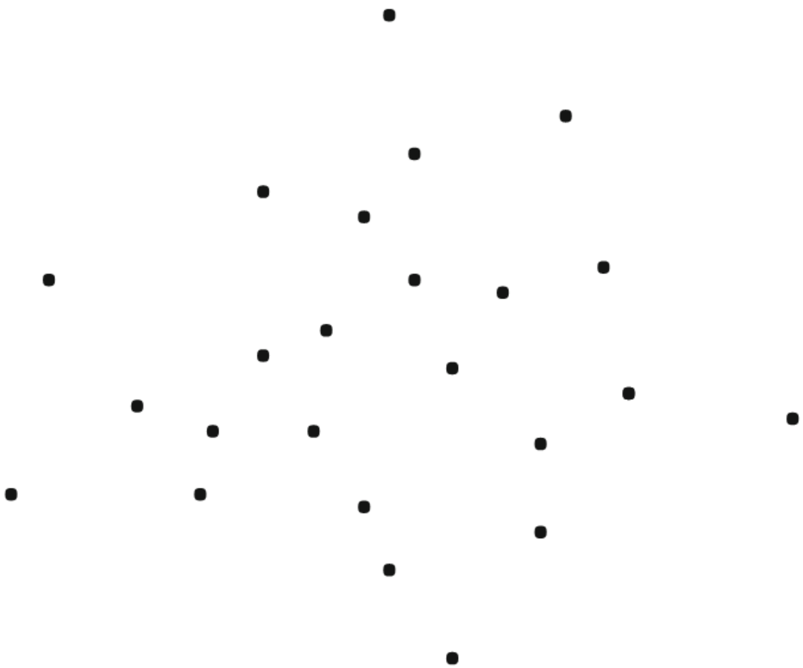
Algorithm $FindDepthOutliers$(Data Set: $\mathcal{D}$, Score Threshold: r)
begin
  $k = 1$;
  repeat
    Find set S of corners of convex hull of $\mathcal{D}$;
    Assign depth k to points in S;
    $\mathcal{D} = \mathcal{D} - S$;
    $k = k + 1$;
  until(D is empty);
  Report points with depth at most r as outliers;
end

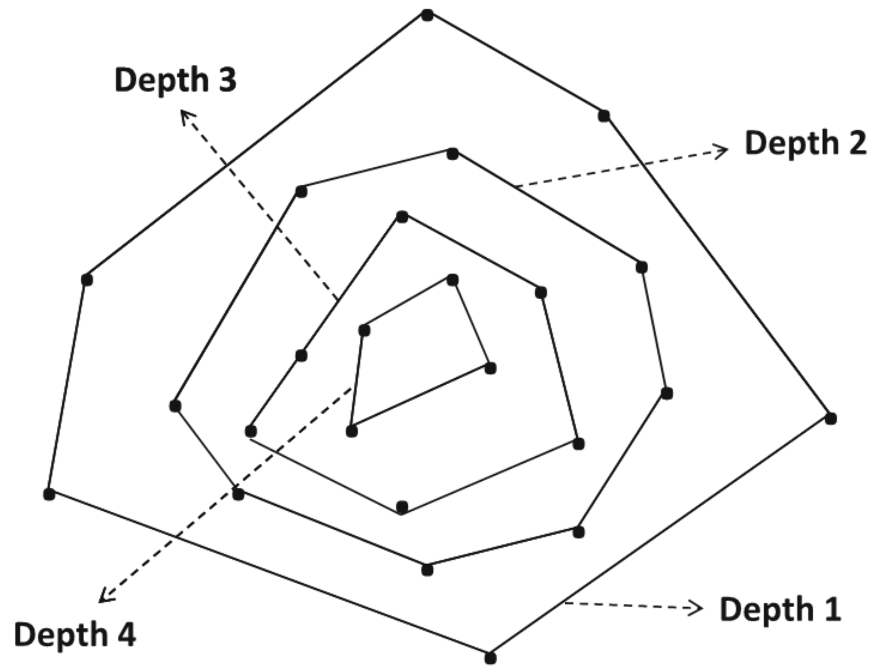
The index of the iteration k provides an outlier score

- Smaller values indicate a greater tendency to be an outlier

Depth-Based Method

- Example





Depth-Based Method

- Less effective (quality and computation) compared to the multivariate method
- Qualitative perspective
  - Do not normalize for the characteristics of the statistical data distribution
  - All data points at the corners of a convex hull are treated equally
  - The scores of many data points are indistinguishable
- Computational complexity
  - Increases significantly with dimensionality

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Probabilistic Models

- Based on a generalization of the multivariate extreme values analysis
  - By generalizing this model to multiple mixture components
- Assume that data were generated from a mixture of k distributions with the probability distributions $\mathcal{G}_1, \ldots, \mathcal{G}_k$
  1. Select a mixture component with prior probability $\alpha_i, i \in \{1 \ldots k\}$
  2. Assume that the r^{th} one is selected
  3. Generate a data point from $\mathcal{G}_r$
- This generative model is denoted by $\mathcal{M}$
  - $\mathcal{M}$ generates data set $\mathcal{D}$ (used to estimate the parameters)
  - Outliers are data points in $\mathcal{D}$ that are highly unlikely to be generated by $\mathcal{M}$

Probabilistic Models

- Based on a generalization of the multivariate extreme values analysis
  - By generalizing this model to multiple mixture components
- Assume that data were generated from a mixture of k distributions with the probability distributions $\mathcal{G}_1, \ldots, \mathcal{G}_k$
  1. Select a mixture component with prior probability $\alpha_i, i \in \{1 \ldots k\}$
  2. Assume that the r^{th} one is selected
  3. Generate a data point from $\mathcal{G}_r$
- This generative model is denoted by $\mathcal{M}$
  - $\mathcal{M}$ generates data set $\mathcal{D}$ (used to estimate the parameters)
  - Outliers are data points in $\mathcal{D}$ that are highly unlikely to be generated by $\mathcal{M}$

Reflects Hawkins's definition of outliers

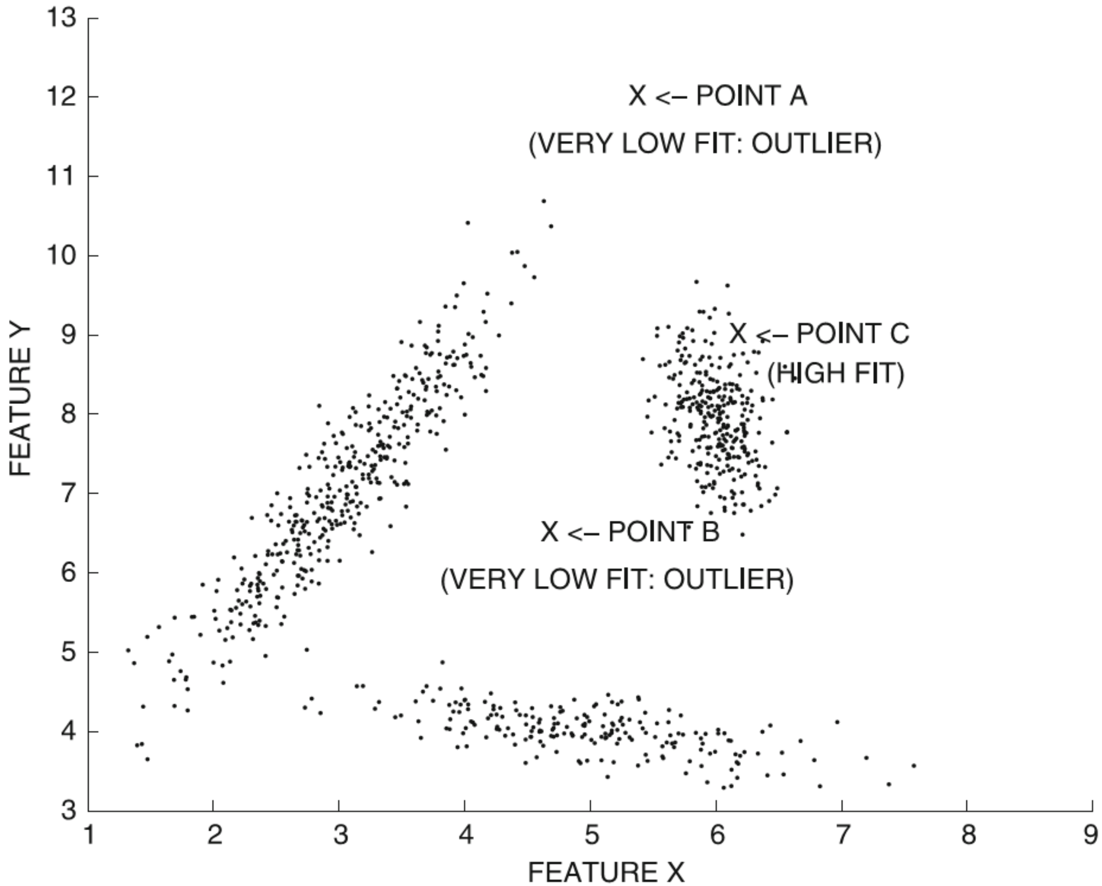
Probabilistic Models

- Parameter estimation
  - α_i and the parameters of distributions $\mathcal{G}_r$
  - Maximum likelihood estimation
- The probability (density function) of the data point $\overline{X_j}$ being generated by $\mathcal{M}$

$$f^{point}(\overline{X_j}|\mathcal{M}) = \sum_{i=1}^{k} \alpha_i \cdot f^i(\overline{X_j})$$

  - $f^i(\cdot)$: density function of $\mathcal{G}_i$ is given by
  - Density value $f^{point}(\overline{X_j}|\mathcal{M})$ provides an estimate of the outlier score of the data point $\overline{X_j}$

Probabilistic Models



- A and B have very low fit to the mixture model ➡ outliers
- C has high fit to the mixture model ➡ not outlier

Probabilistic Models

Maximum Likelihood estimation

- Data set $\mathcal{D}$ containing n data points $\overline{X_1}, \ldots, \overline{X_n}$
  - Probability density of the data set generated by M

$$f^{data}(\mathcal{D}|\mathcal{M}) = \prod_{j=1}^{n} f^{point}(\overline{X_j}|\mathcal{M})$$

  - The log-likelihood fit

$$\mathcal{L}(\mathcal{D}|\mathcal{M}) = \log(\prod_{j=1}^{n} f^{point}(\overline{X_j}|\mathcal{M})) = \sum_{j=1}^{n} \log(\sum_{i=1}^{\kappa} \alpha_i \cdot f^i(\overline{X_j}))$$

Log-likelihood fit can be optimized using EM algorithm

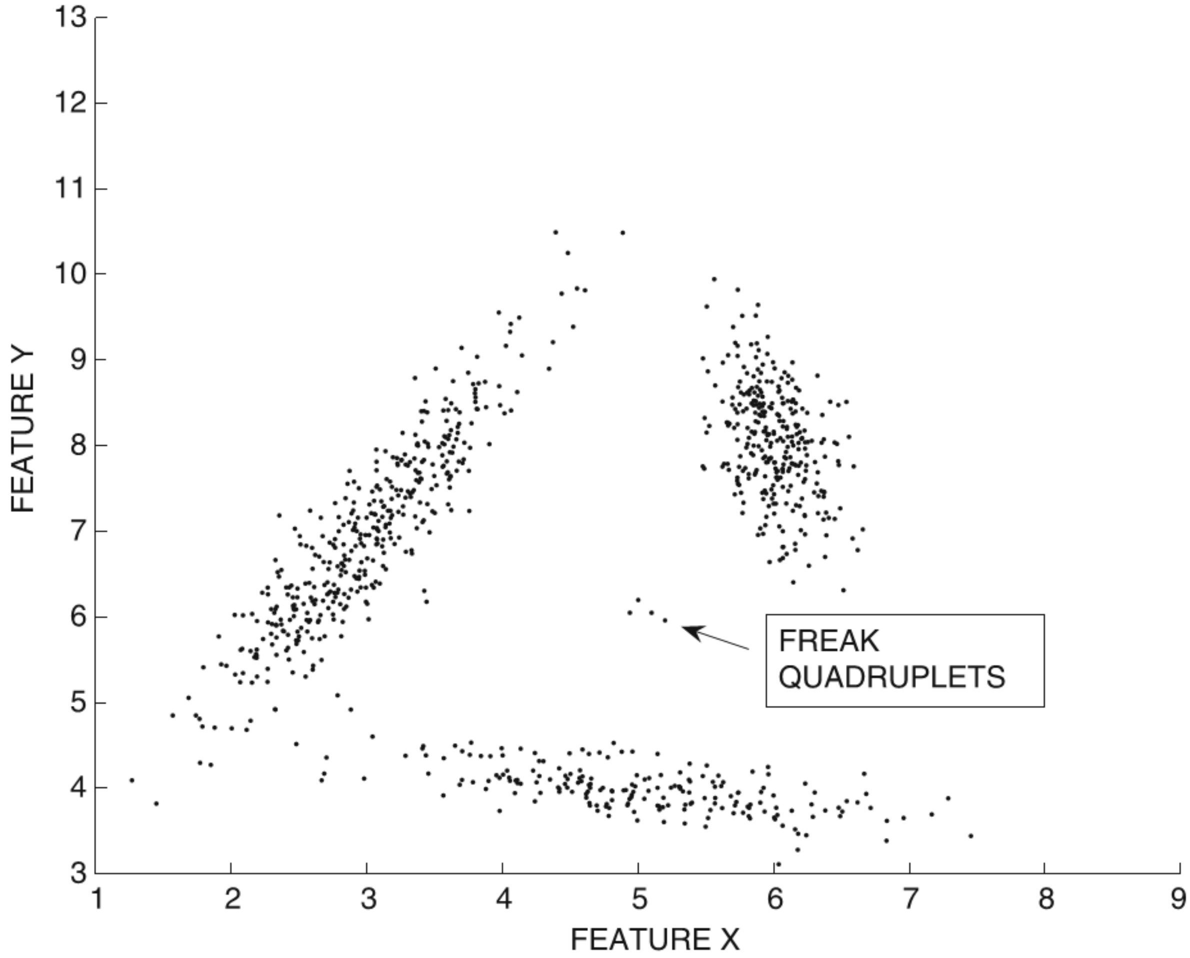
Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Clustering for Outlier Detection

- Outlier detection vs. Clustering
  - Clustering: find groups of data points that are similar
  - Outlier detection: detect individual data points that are different from the remaining data
  - A simplistic view: every data point is either a member of a cluster or an outlier
- In general clustering is not an appropriate approach because it is not optimized for outlier detection
- Clustering models have some advantages
  - Outliers often tend to occur in small clusters
    - anomaly in the generating process may be repeated a few times
  - A small group of related outliers may be created

Clustering for Outlier Detection



Clustering for Outlier Detection

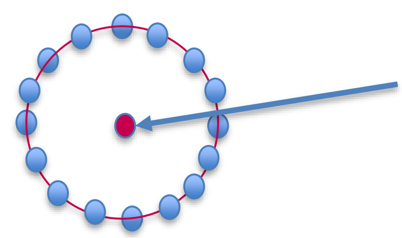
- Defining the outlier score of a data point
  - The distance of the data point to its closest cluster centroid
- The Mahalanobis distance

$$Maha(\overline{X}, \overline{\mu_r}, \Sigma_r) = \sqrt{(\overline{X} - \overline{\mu_r})\Sigma_r^{-1}(\overline{X} - \overline{\mu_r})^T}.$$

  - $\bar{X}$: a data point
  - $\overline{\mu_r}$: d-dimensional mean vector of the r^{th} cluster
  - Σ_r: $d \times d$ covariance matrix

Clustering for Outlier Detection

- The major problem with clustering algorithms
  - Sometimes not able to properly distinguish between ambient noise and truly isolated anomaly
  - The distance to the closest cluster centroid does not accurately reflect the instance-specific isolation of the underlying data point



Outlier but very close to the cluster centroid

Content

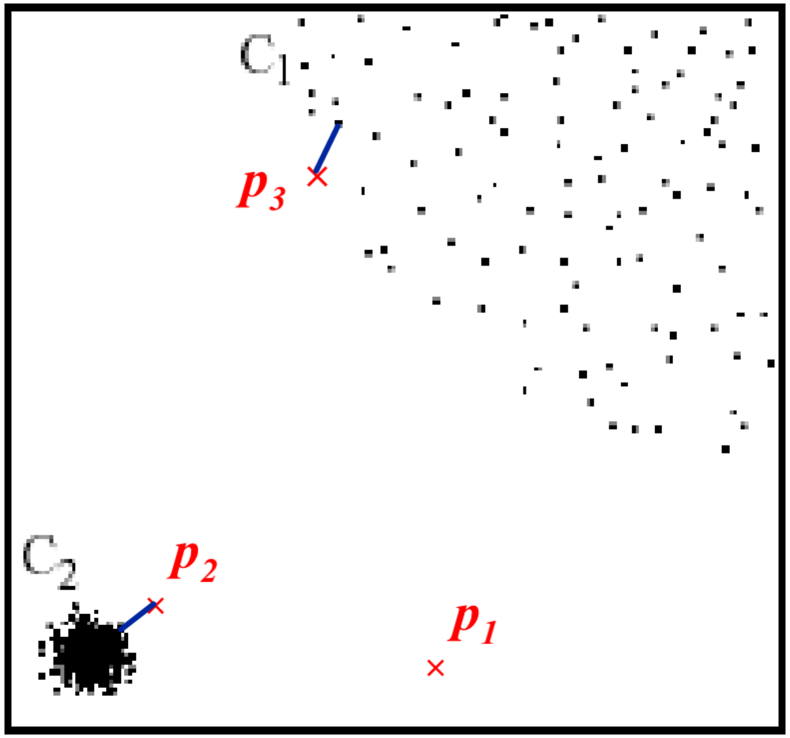
- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Distance-Based Outlier Detection

- Outliers: Data points far away from the "crowded regions" (or clusters)
- Principle of distance-based methods
  - Outlier score: distance to the k^{th} nearest neighbor
  - Other variation: the average distance to the k-nearest neighbors
- Distance-based methods can distinguish between ambient noise and truly isolated anomalies
  - Ambient noise will typically have a lower k-nearest neighbor distance than a truly isolated anomaly

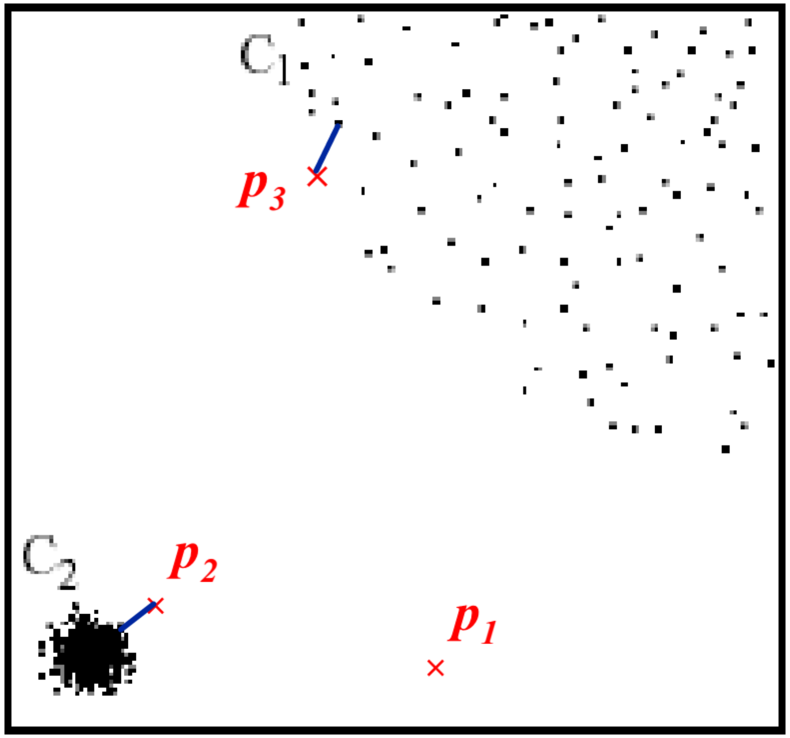
Example

- Clustering based
  - p_1 and p_3 are outliers (far from cluster centroid)
- Distance based
  - p_1 is an outlier (distances to k^{th} NN of p_2 and p_3 are the same)
- Correct
  - p_1 and p_2 are outliers



Example

- Clustering based
  - p_1 and p_3 are outliers (far from cluster centroid)
- Distance based
  - p_1 is an outlier (distances to k^{th} NN of p_2 and p_3 are the same)
- Correct
  - p_1 and p_2 are outliers



Need to differentiate dense and spare clusters

Local Outlier Factor (LOF)

- For data point $\bar{X}$
  - $V^k(\bar{X})$ be the distance to its k-nearest neighbor
  - $L_k(\bar{X})$ be the set of points within the k-nearest neighbor distance of $\bar{X}$
  - Reachability distance of $\bar{X}$ with respect $\bar{Y}$

$$R_k(\overline{X}, \overline{Y}) = \max\{Dist(\overline{X}, \overline{Y}), V^k(\overline{Y})\}$$

  - $\bar{Y}$ in dense region ➡ $Dist(\bar{X}, \bar{Y})$
  - $\bar{Y}$ in spare region and $Dist(\bar{X}, \bar{Y})$ small ➡ $V^k(\bar{Y})$

Local Outlier Factor (LOF)

- Average reachability distance $\bar{X}$ with respect to its neighborhood $L_k(\bar{X})$

$$AR_k(\overline{X}) = \text{MEAN}_{\overline{Y} \in L_k(\overline{X})} R_k(\overline{X}, \overline{Y})$$

  - ❑ $\bar{X}$ in dense region: $AR_k(\bar{X})$ is small

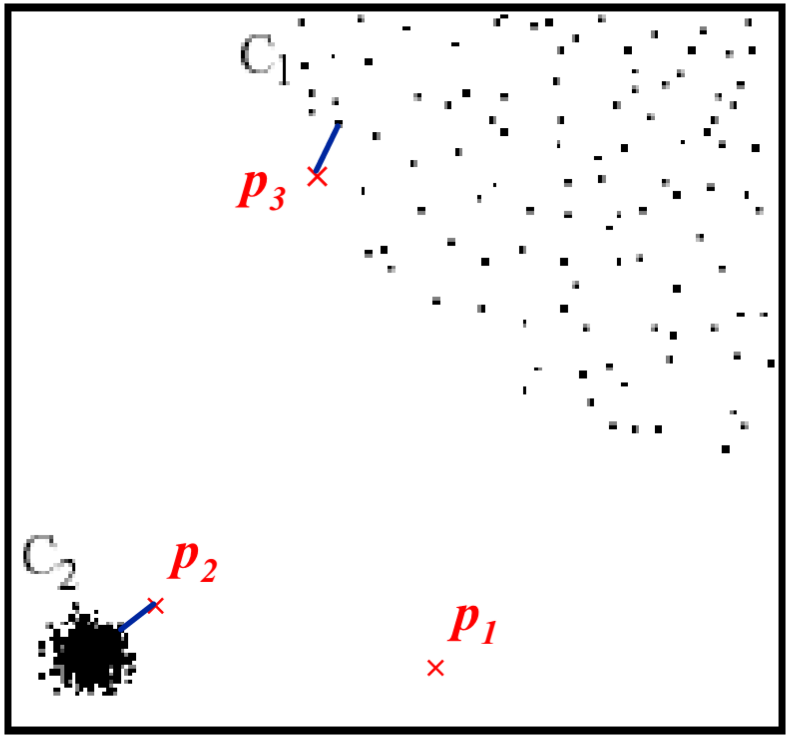
- Local Outlier Factor of $\bar{X}$

$$LOF_k(\overline{X}) = \text{MEAN}_{\overline{Y} \in L_k(\overline{X})} \frac{AR_k(\overline{X})}{AR_k(\overline{Y})}$$

  - ❑ $\bar{Y}$ in dense region: $AR_k(\bar{Y})$ is small, increase $LOF_k(\bar{X})$

Example

- Clustering based
  - p_1 and p_3 are outliers (far from cluster centroid)
- Distance based
  - p_1 is an outlier (distances to k^{th} NN of p_2 and p_3 are the same)
- LOF
  - p_1 and p_2 are outliers



Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Density-Based Outlier Detection

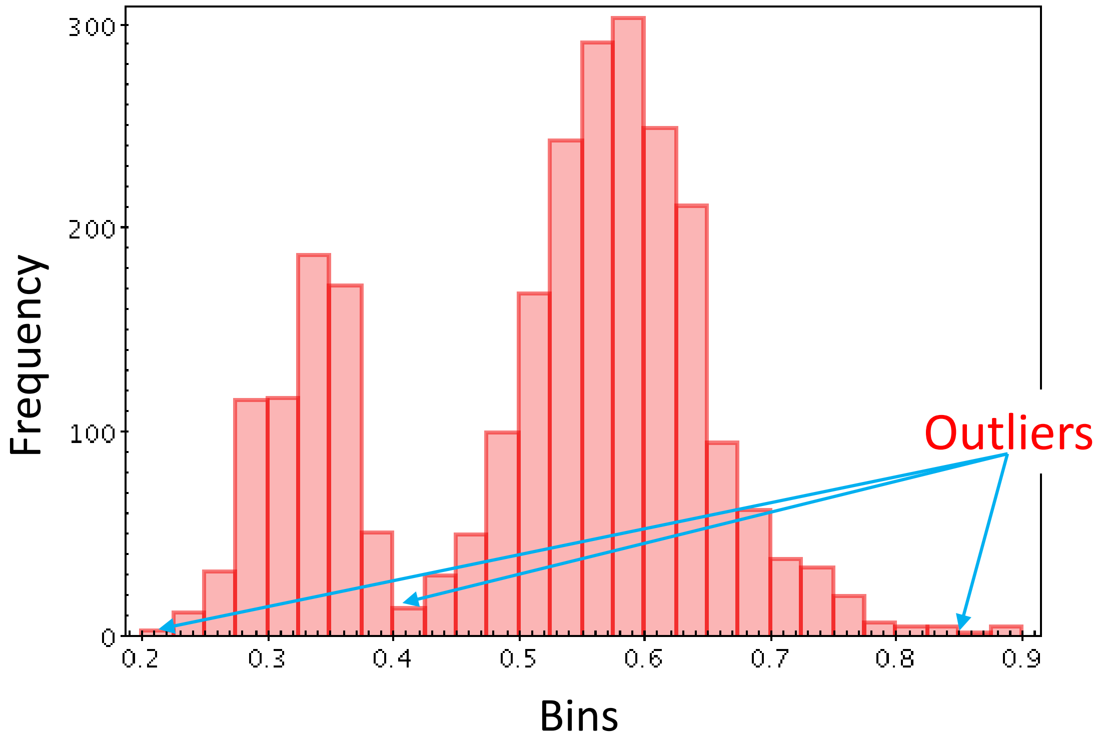
- Based on similar principles as density-based clustering
  - Determine sparse regions in the underlying data in order to report outliers
- Main methods
  - Histogram and Grid-based techniques

Histogram and Grid-based Techniques

- Univariate data
  - Histograms are simple and easy, therefore used quite frequently in many application domains
  - The data is discretized into bins
  - Data points in bins with very low frequency are outliers
  - The number of other data points in the bin for data point $\bar{X}$ is the outlier score for $\bar{X}$

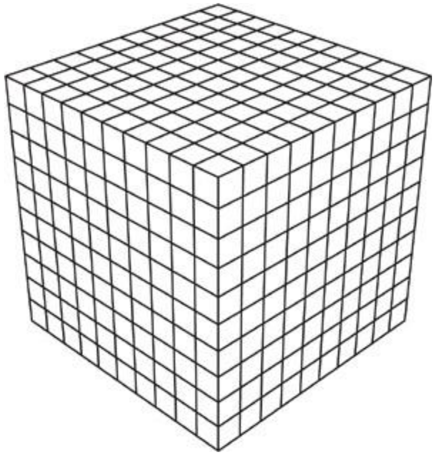
Histogram and Grid-based Techniques

- Univariate data



Histogram and Grid-based Techniques

- Multivariate data
  - A grid-structure is used
  - Each dimension is partitioned into p equi-width ranges
  - Number of points in a grid region is outlier score
  - Data points that have density less than τ in any particular grid region are reported as outliers



Histogram and Grid-based Techniques

Major challenges

- Hard to determine the optimal histogram width
  - Too wide, or too narrow, will not model the frequency distribution well
  - Grid-structures have similar issues
- Too local in nature
  - Do not take the global characteristics of the data into account
- Do not work very well in high dimensionality
  - The sparsity of the grid structure with increasing dimensionality
  - A d-dimensional space will contain at least 2^d grid-cells
  - Number of data points expected to populate each cell reduces exponentially with increasing dimensionality

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Information-Theoretic Models

- Outliers are data points that do not naturally fit the remaining data distribution
  - If we can somehow compress the data, the outliers will increase the minimum code length required to describe it
- Example

ABABABABABABABABABABABABABABABABAB

ABABACABABABABABABABABABABABABABAB

  - The first string: “AB 17 times”
  - The second string can no longer be described as concisely
  - Symbol C increases the *minimum description length*
  - C is considered as an outlier

Information-Theoretic Models

- Outliers are data points that do not naturally fit the remaining data distribution
  - If we can somehow compress the data, the outliers will increase the minimum code length required to describe it
- Example

ABABABABABABABABABABABABABABABABABAB

ABABACABABABABABABABABABABABABABABAB

  - The first string: “AB 17 times”
  - The second string can no longer be described as concisely
  - Symbol C increases the *minimum description length*
  - C is considered as an outlier

Outliers increase the model complexity

Information-Theoretic Models

- Every conventional model can be converted into an information-theoretic version
- Probabilistic models
  - Model complexity: number of mixture components
  - Information-theoretic version examine the size of the model required
- Clustering-based models
  - Model complexity: number of clusters
  - Information-theoretic version reports required model size as the outlier score
- Density-based models
  - Model complexity: number bins

Information-Theoretic Models

- When do we use conventional models?
  - Cases where the summary models can be explicitly constructed
  - Because: outlier scores are directly optimized
- When do we use information-theoretic models?
  - Cases where an accurate summary model of the data is hard to explicitly construct
  - Solution: Kolmogorov complexity can be used to estimate the compressed space requirements of the data set indirectly

Content

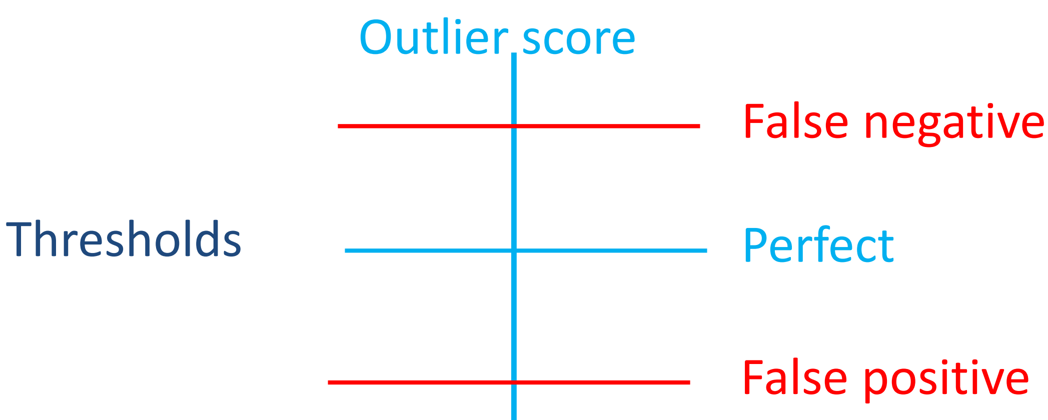
- Part I: Introduction
- Part II: Models for Outlier Analysis
  - Extreme Value Analysis
  - Probabilistic Models
  - Clustering for Outlier Detection
  - Distance-Based Outlier Detection
  - Density-Based Outlier Detection
  - Information-Theoretic Models
  - Outlier Validity
- Part III: Advanced Topics
- Part IV: A Case Study

Methodological Challenges

- Outlier analysis is an unsupervised problem
  - Hard to validate because of the lack of external criteria
- Whether internal criteria can be defined for outlier validation?
  - Almost never because of the small sample solution space
  - A model only needs to be correct on a few outlier data points to be considered a good model

ROC Curve

- Outlier detection algorithms are typically evaluated with external measures
  - Known outlier labels are used as ground-truth
- False-positives and False-negatives
  - A threshold is used to generate a binary label



- Receiver Operating Characteristic (ROC) curve is used for trade-off problem

ROC Curve

- True positive rate (TPR or recall)

$$TPR(t) = Recall(t) = 100 * \frac{|\mathcal{S}(t) \cap \mathcal{G}|}{|\mathcal{G}|}$$

  - t: a given threshold
  - $\mathcal{S}(t)$: the declared outlier set
  - $\mathcal{G}$: the true set (ground-truth set)

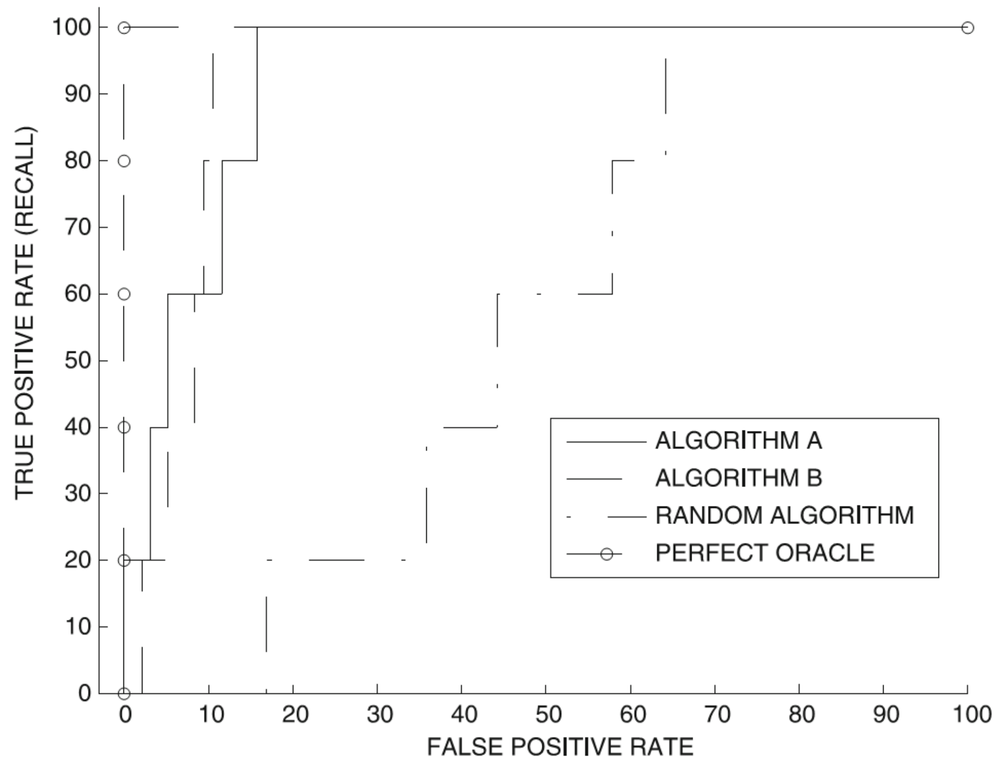
- False positive rate (FPR)

$$FPR(t) = 100 * \frac{|\mathcal{S}(t) - \mathcal{G}|}{|\mathcal{D} - \mathcal{G}|}$$

  - $\mathcal{D}$: the whole dataset

ROC Curve

- The ROC curve is defined by plotting the FPR(t) on the X-axis, and TPR(t) on the Y-axis for varying values of t



Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
  - Outlier Detection with Categorical Data
  - High-Dimensional Outlier Detection
  - Outlier Ensembles
- Part IV: A Case Study

Outlier Detection with Categorical Data

- Categorical variables
  - Take on one of a limited, and usually fixed, number of possible values
  - Represent types of data which may be divided into groups
  - Examples: sex, age group, and educational level
- We can modify presented algorithms for outlier analysis to work with categorical data
  - Probabilistic Models
  - Clustering and Distance-Based Methods

Probabilistic Models

- We use a generative mixture model ($\mathcal{M}$) of categorical data (instead of numerical data)
- k components of the mixture model are denoted by $\mathcal{G}_1, \ldots, \mathcal{G}_k$
- The generative process to generate each point in the d-dimensional data set $\mathcal{D}$
  1. Select a mixture component with prior probability $\alpha_i, i \in \{1 \ldots k\}$
  2. Assume that the r^{th} one is selected
  3. Generate a data point from $\mathcal{G}_r$

Probabilistic Models

- The value of the generative probability $g^{m,\Theta}(\bar{X})$ of a data point from cluster m

$$g^{m,\Theta}(\overline{X}) = \prod_{r=1}^{d} p_{rj_r m}$$

  - $\bar{X}$ contains the attribute value indices $j_1, \ldots, j_d$
  - p_{ijm}: probability in which the j^{th} value of the i^{th} attribute is generated by cluster m
- Sum of the probabilities over all components

$$P(\bar{X}|\mathcal{M}) = \sum_{r=1}^{k} \alpha_r \cdot g^{r,\Theta}(\bar{X})$$

  - is used as the outlier score

Clustering and Distance-Based Methods

- Centroid of a categorical data set
  - Convert categorical data to numerical or binary data
    - Calculations are conducted on numerical/binary data
  - Use probability histogram of values on each attribute

Data	(Color, Shape)
1	(Blue, Square)
2	(Red, Circle)
3	(Green, Cube)
4	(Blue, Cube)
5	(Green, Square)
6	(Red, Circle)
7	(Blue, Square)
8	(Green, Cube)
9	(Blue, Circle)
10	(Green, Cube)

Attribute	Histogram	Mode
Color	Blue $= 0.4$ Green $= 0.4$ Red $= 0.2$	Blue *or* Green
Shape	Cube $= 0.4$ Square $= 0.3$ Circle $= 0.3$	Cube

Clustering and Distance-Based Methods

- Calculating similarity
  - Convert categorical data to numerical data or to binary data
    - Calculations are conducted on numerical/binary data
  - Match-based similarity using probability histogram

Clustering and Distance-Based Methods

- Calculating similarity

Data	(Color, Shape)
1	(Blue, Square)
2	(Red, Circle)
3	(Green, Cube)
4	(Blue, Cube)
5	(Green, Square)
6	(Red, Circle)
7	(Blue, Square)
8	(Green, Cube)
9	(Blue, Circle)
10	(Green, Cube)

Attribute	Histogram	Mode
Color	Blue= 0.4 Green = 0.4 Red = 0.2	Blue *or* Green
Shape	Cube = 0.4 Square = 0.3 Circle = 0.3	Cube

(Blue, Square) vs (Blue, Cube): 0.4 + 0
(Blue, Square) vs (Blue, Square): 0.4 + 0.3
(Blue, Square) vs (Red, Cube): 0 + 0

Outliers in Transaction Data

- Frequent patterns are much less likely to occur in outlier transactions
  - Compute sum of all the supports of frequent patterns occurring in a particular transaction
  - Normalize by dividing with the number of frequent patterns
  - This provides an outlier score for the pattern

Outliers in Transaction Data

- Frequent Pattern Outlier Factor

$$FPOF(T_i) = \frac{\sum_{X \in FPS(\mathcal{D}, s_m), X \subseteq T_i} s(X, \mathcal{D})}{|FPS(\mathcal{D}, s_m)|}$$

  - $\mathcal{D}$: transaction database containing transactions $T_1, \dots, T_N$
  - $s(X, \mathcal{D})$: support of itemset X in $\mathcal{D}$
  - $FPS(\mathcal{D}, s_m)$: set of frequent patterns in the database $\mathcal{D}$ at minimum support level s_m
  - $FPOF(T_i)$: frequent pattern outlier factor of transaction T_i

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
  - Outlier Detection with Categorical Data
  - High-Dimensional Outlier Detection
  - Outlier Ensembles
- Part IV: A Case Study

High-Dimensional Outlier Detection

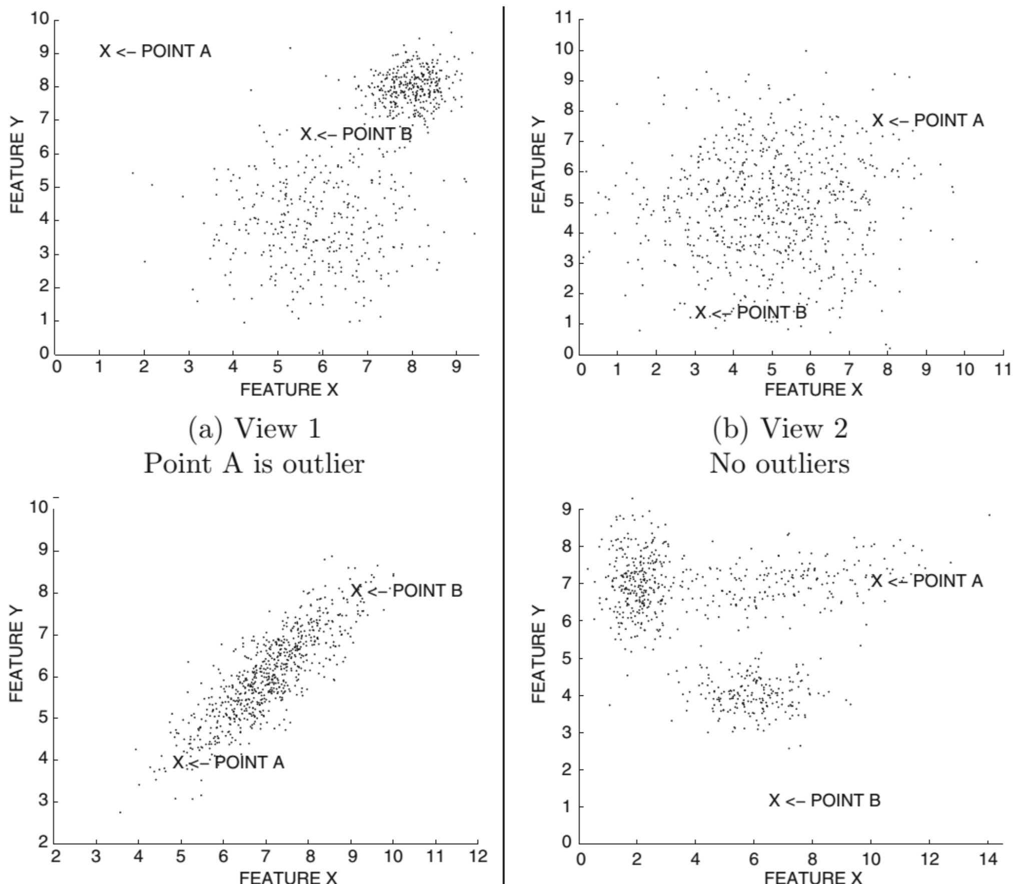
- Challenging
  - Varying importance of the different attributes
  - Complexity
- The idea: causality of an anomaly can be typically perceived in only a small subset of the dimensions
  - The remaining dimensions are irrelevant and only add noise
  - Different subsets of dimensions may be relevant to different anomalies
  - Full-dimensional analysis often does not properly expose the outliers in high-dimensional data

High-Dimensional Outlier Detection

- Challenging
  - Varying importance of the different attributes
  - Complexity
- The idea: causality of an anomaly can be typically perceived in only a small subset of the dimensions
  - The remaining dimensions are irrelevant and only add noise
  - Different subsets of dimensions may be relevant to different anomalies
  - Full-dimensional analysis often does not properly expose the outliers in high-dimensional data

An outlier is defined by associating it with subspaces specific to that outlier

High-Dimensional Outlier Detection



(a) View 1
Point A is outlier

(b) View 2
No outliers

(c) View 3
No outliers

(d) View 4
Point B is outlier

High-Dimensional Outlier Detection

- Methods: search the data points and dimensions in an integrated way to reveal the most relevant outliers
- Grid-based rare subspace exploration
  - Rare subspaces of the data are explored after discretizing the data into a grid-like structure
- Random subspace sampling
  - Subspaces of the data are sampled to discover the most relevant outliers

Grid-based Rare Subspace Exploration

- Data discretization
  - Select k attributes, each attribute is divided into p ranges to creates a grid cell of dimensionality k
  - Expected fraction of data points in a grid cell: $f^k = (^1/_p)^k$
- Sparsity coefficient of a k-dimensional cube $\mathcal{R}$
  - The presence of any point in $\mathcal{R}$ is a Bernoulli random variable with probability f^k
  - The expected number and standard deviation of the (n) points in $\mathcal{R}$ are $\mathrm{n} \cdot f^k$ and $\sqrt{\mathrm{n} \cdot f^k(1 - f^k)}$ (Multinomial)

$$S(\mathcal{R}) = \frac{n_{\mathcal{R}} - n \cdot f^k}{\sqrt{n \cdot f^k \cdot (1 - f^k)}}$$

  - $n_{\mathcal{R}}$: number of data points in $\mathcal{R}$

Random Subspace Sampling

- Idea of random subspace sampling
  - Explore many possible subspaces and examine if at least one of them contains outliers
- Feature bagging (for the t^{th} iteration)
  1. Randomly chose the size of the feature subset N_t from a uniform distribution between $\lfloor d/2 \rfloor$ and $(d-1)$
  2. Randomly pick, without replacement, N_t features to create a data set D_t
  3. Apply an outlier detection algorithm O_t on the data set D_t to create score vectors S_t
  - Combine outlier scores S_t from different iterations

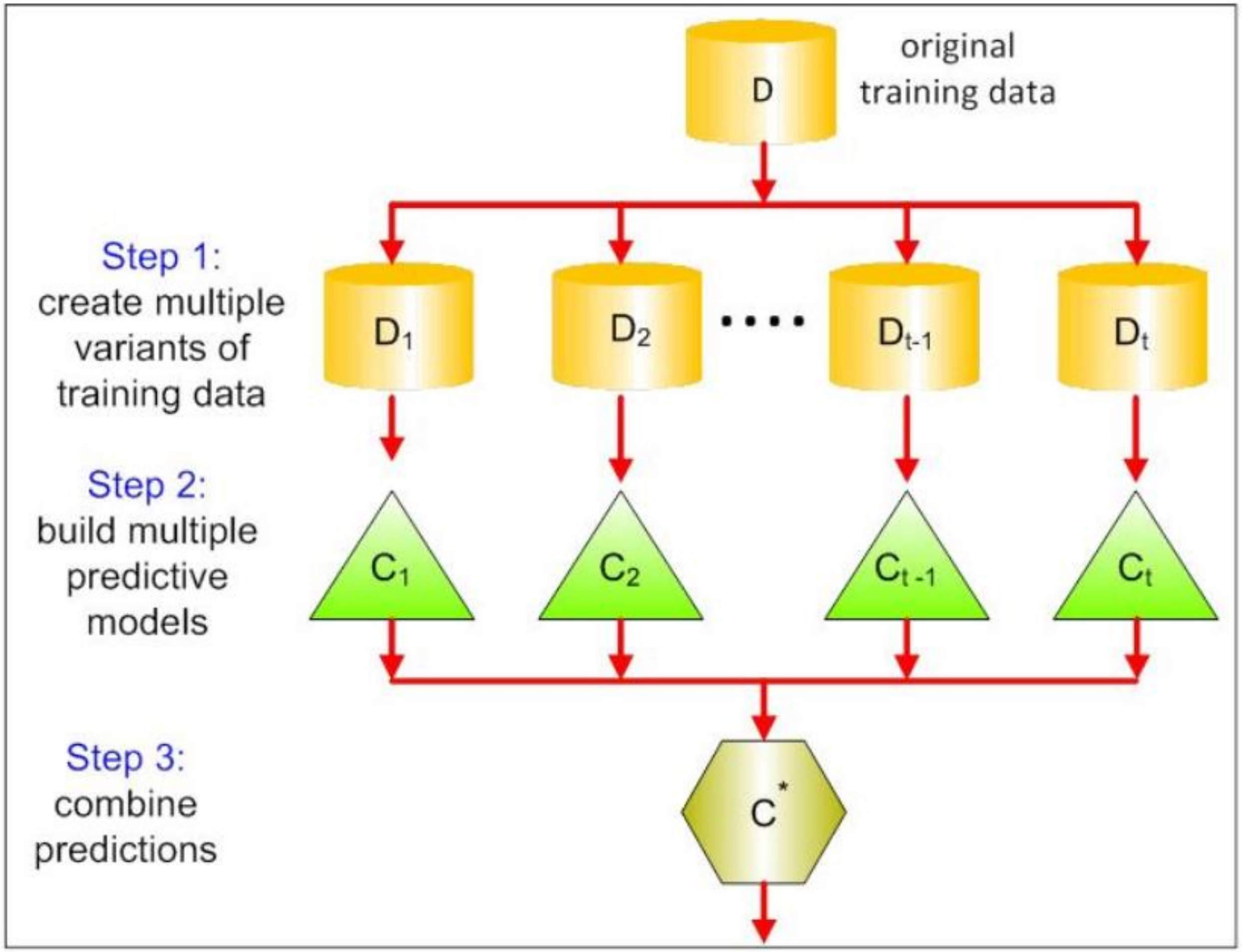
Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
  - Outlier Detection with Categorical Data
  - High-Dimensional Outlier Detection
  - Outlier Ensembles
- Part IV: A Case Study

Outlier Ensembles

- Ensemble methods
  - Combine outputs of multiple models to create the final one
  - Take advantages of different algorithms in different situations
  - Used popular in classification, clustering, outlier detection
- Outlier Ensembles
  - Combine outlier scores of multiple outlier analysis models
  - How to create multiple models?
  - How to combine outputs?

Ensemble Methods



Multiple Models

- Use different version of the dataset
  - Random subspace sampling
- Use different algorithms
  - Probabilistic, Clustering-based, distance-based, and so on
- Use different parameters
- Use all of the above

Output Combination

- Use a maximum function
  - The score is the maximum of the outlier scores from the different components
- Use a average function
  - The score is the average of the outlier scores from the different components
- Use majority voting
  - Do not use with outlier scores
  - A data point is considered as an outlier if more than a half of models say that

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study

Content

- Part I: Introduction
- Part II: Models for Outlier Analysis
- Part III: Advanced Topics
- Part IV: A Case Study
  - Fraud Detection

Fraud Detection

- Question: Identify fraudulent activities or users from observed transaction data
- Data
  - Transactions of different users
  - Meta information about users
- Applications
  - Insurance (auto, health)
    - Claimant, Provider, Payer
  - Credit Cards
    - Customer, Supplier, Bank
  - Telecommunications
    - Customer, Provider

Fraud Detection

- Challenges:
  - Track and model human behavior
  - Anomalies caused by human intentionally
  - Massive data sizes
  - Difficult to distinguish between frauds and noises
  - Need domain knowledge

Generic Fraud Detection Methods

- Activity monitoring
  - Build profiles for individuals (customers, users, etc.) based on historic data
  - Compare current behavior with historical profile for significant deviations
- Clustering based
  - Cluster historical profiles of customers
  - Identify small clusters or outlying profiles as anomalies

Generic Fraud Detection Methods

- Strengths
  - Anomaly detection is fast (good for real time)
  - Results are easy to explain
- Weaknesses
  - Need to create and maintain a large number of profiles
  - Adequate historical data might not be available
  - Too many false positives

Other Methods

- Classification based
  - Advantages: there are many effective supervised methods which can build a good model
  - Disadvantages: difficult to collect enough positive samples
- Density based
  - Use histogram and grid-based techniques
  - Advantages: simple and easy to implement; can be effective in some situations
  - Disadvantages: cannot detect frauds in difficult situations

Outlier Analysis
Q&A

Ngô Xuân Bách

Vietnam Institute for Advanced Study in Mathematics