Outlier Analysis

Lijun Zhang
zlj@nju.edu.cn
http://cs.nju.edu.cn/zlj

Outline

† Introduction

† Extreme Value Analysis

† Probabilistic Models

† Clustering for Outlier Detection

† Distance-Based Outlier Detection

† Density-Based Methods

† Information-Theoretic Models

† Outlier Validity

† Summary

Introduction (1)

- A Quote

> "You are unique, and if that is not fulfilled, then something has been lost."—Martha Graham

- An Informal Definition

> *"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."*

- A Complementary Concept to Clustering
  - Clustering attempts to determine groups of data points that are similar
  - Outliers are individual data points that are different from the remaining data

Introduction (2)

- Applications
  - Data cleaning
    - Remove noise in data
  - Credit card fraud
    - Unusual patterns of credit card activity
  - Network intrusion detection
    - Unusual records/changes in network traffic

Introduction (3)

- The Key Idea
    - 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 outlier score
- Outputs of Outlier Detection Algorithms
    - Real-valued outlier score
    - Binary label

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Extreme Value Analysis (1)

- Statistical Tails

http://www.regentsprep.org/regents/math/algtrig/ats2/normallesson.htm

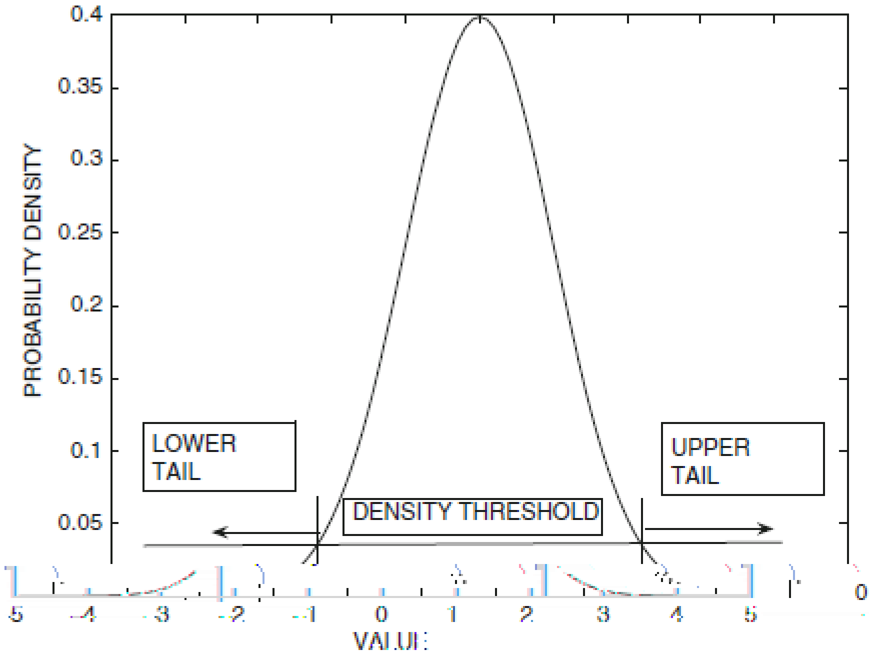
- All extreme values are outliers
- Outliers may not be extreme values
  - {1,3,3,3,50,97,97,97,100}
  - 1 and 100 are extreme values
  - 50 is an outlier but not extreme value

Extreme Value Analysis (2)

- All extreme values are outlies
- Outlies may not be extreme values

Univariate Extreme Value Analysis (1)

- Statistical Tail Confidence Tests
  - Suppose the density distribution is $f_{\tilde{N}}(x)$
  - Tails are extreme regions s.t. $f_{\tilde{N}}(x) \leq \theta$
- Symmetric Distribution
  - Two symmetric tails
  - The areas inside tails represent the cumulative probability



(a) Symmetric distribution

Univariate Extreme Value Analysis (2)

† Statistical Tail Confidence Tests

”

The Procedure (1)

- A model distribution is selected
  - Normal Distribution with mean μ and standard deviation σ

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

- Parameter Selection
  - Prior domain knowledge
  - Estimate from data

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

The Procedure (2)

- Z-value of a random variable

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

  - Large positive values of z_i correspond to the upper tail
  - Large negative values of z_i correspond to the lower tail
  - z_i follows the normal distribution

- Extreme values
  - $|z_i| \geq 3$

[redacted]

Multivariate Extreme Values (1)

- Unimodal probability distributions with a single peak
  - Suppose the density distribution is $f_{\tilde{N}}(x)$
  - Tails are extreme regions s.t. $f_{\tilde{N}}(x) \leq \theta$
- Multivariate Gaussian Distribution

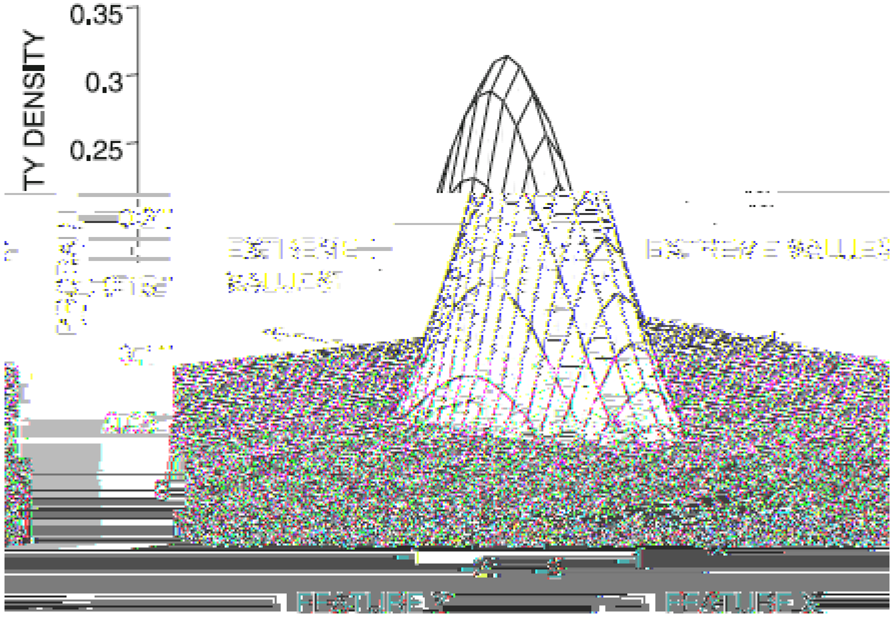
$$f(\overline{X}) = \frac{1}{\sqrt{|\Sigma| \cdot (2 \cdot \pi)^{(d/2)}}} \cdot e^{-\frac{1}{2} \cdot (\overline{X} - \overline{\mu}) \Sigma^{-1} (\overline{X} - \overline{\mu})^T}$$

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

where $Maha(\bar{X}, \bar{\mu}, \Sigma)$ is the Mahalanobis distance between $\bar{X}$ and $\bar{\mu}$

Multivariate Extreme Values (2)

- Extreme-value Score of $\bar{X}$
  - $Maha(\bar{X}, \bar{\mu}, \Sigma)$
  - Larger values imply more extreme behavior



(b) Multivariate extreme values

Multivariate Extreme Values (2)

- Extreme-value Score of $\bar{X}$
  - $Maha(\bar{X}, \bar{\mu}, \Sigma)$
  - Larger values imply more extreme behavior
- Extreme-value Probability of $\bar{X}$
  - Let $\mathcal{R}$ be the region
    - ì L{ ;$ / = D(=;$ á Sā) - R / (=:$Dá Sā)}-
  - Cumulative probability of $\mathcal{R}$
  - Cumulative Probability of χ^6 distribution for which the value is larger than $Maha(\bar{X}, \bar{\mu}, \Sigma)$

Why χ^2 distribution?

† The Mahalanobis distance

„ Let Σ be the covariance matrix

/ = D(;$á ;ã) - √(L(;$ F)ã ? 5 (;$ F)ã C)

„ Projection+Normalization

9 Let - L 7 & 7 C L Ã × Ü @ 5 ê Ü 6 Ü Ü C

9 Then, - ? 5 L 7 ? 5 & 7 C L Ã × Ü @ 5 ê Ü ? 6 Ü Ü C

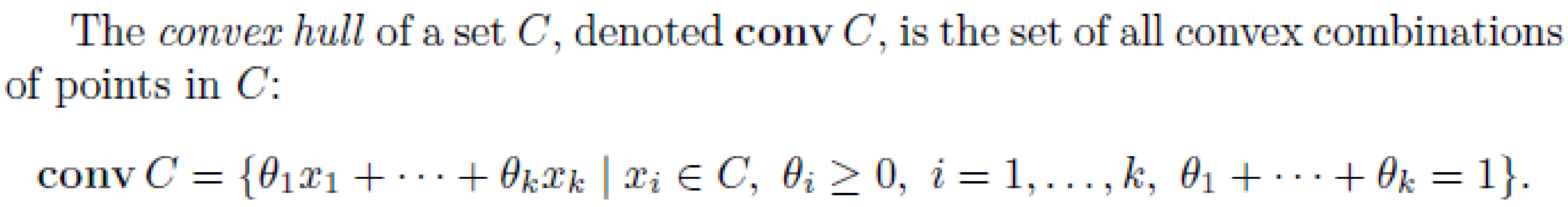
/ = D(;$á ;ã) - √(L(;$ F)ã (Í × Ü @ 5 ê Ü ? 6 Ü Ü C) (;$ F)ã C) L √(Í × Ü @ 5 (Ü (;$ F)ã C / ê Ü) 6)

Adaptive to the Shape

- B is an extreme value

Depth-Based Methods

† Convex Hull



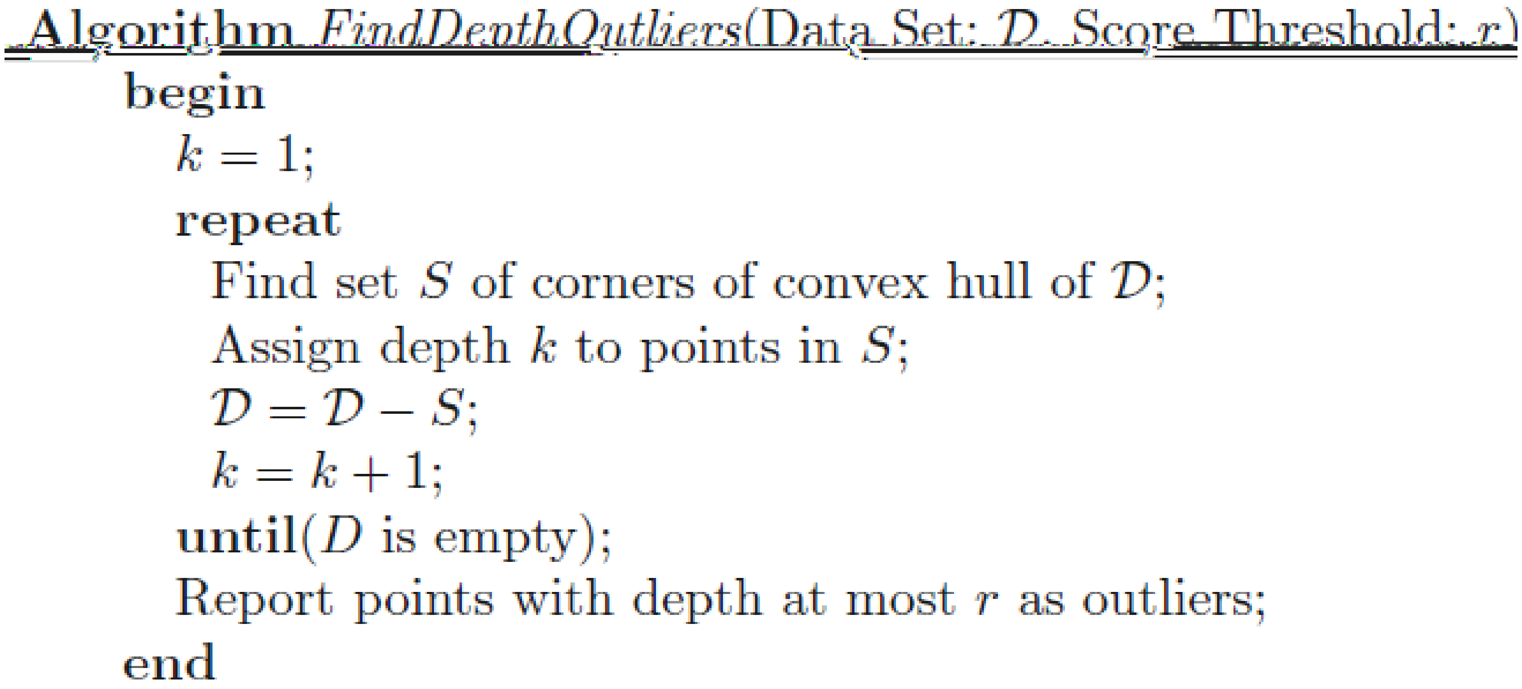
The *convex hull* of a set C, denoted **conv** C, is the set of all convex combinations of points in C:

$$\mathbf{conv}\, C = \{\theta_1 x_1 + \cdots + \theta_k x_k \mid x_i \in C,\ \theta_i \geq 0,\ i = 1, \ldots, k,\ \theta_1 + \cdots + \theta_k = 1\}.$$

„ Corners

The Procedure

- The index k is the outlier score
  - Smaller values indicate a grate tendency



```
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
```

An Example

- Peeling Layers of an Onion

Limitations

- No Normalization
- Many data points are indistinguishable
- The computational complexity increases significantly with dimensionality

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Probabilistic Models

- Related to Probabilistic Model-Based Clustering
- The Key Idea
  - Assume data is generated from a mixture-based generative model
  - Learn the parameter of the model from data
    - EM algorithm
  - Evaluate the probability of each data point being generated by the model
    - Points with low values are outliers

Mixture-based Generative Model

- Data was generated from a mixture of k distributions with probability distribution $\mathcal{G}_1, \dots, \mathcal{G}_k$
- $\mathcal{G}_i$ represents a cluster/mixture component
- Each point $\bar{X}$ is generated as follows
  - Select a mixture component with probability $\alpha_i = P(\mathcal{G}_i),\ i = 1, \dots, k$
  - Assume the r-th component is selected
  - Generate a data point from G_r

Learning Parameter from Data

† The probability that $\overline{X}_{Ý}$ generated by the mixture model $\mathcal{M}$ is given by

Þ Þ Þ
Bã â Ü á ç (?%Ý ç) L Í 2 (á Ü á %) L Í 2 (á Ü 2 :%Ý á Ü; L Í Ù Ü „ Ü B:%Ý;
Ü @ 5 Ü @ 5 Ü @ 5

† The probability of the data set $\mathcal{D} = \{\overline{X}_5, \ldots, \overline{X}_{á}\}$ generated by $\mathcal{M}$

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

† Learning parameters that maximize

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

Identify Outliers

- Outlier Score is defined as

Bã â Ü(á ç% .Ý ç) L Í 2 :

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Clustering for Outlier Detection

- Outlier Analysis v.s. Clustering
  - Clustering is about finding “crowds” of data points
  - Outlier analysis is about finding data points that are far away from these crowds
- Every data point is
  - Either a member of a cluster
  - Or an outlier
- Some clustering algorithms also detect outliers
  - DBSCAN, DENCLUE

The Procedure (1)

- A Simple Way
  1. Cluster the data
  2. Define the outlier score as the distance of the data point to its cluster centroid

The Procedure (2)

- A Better Approach
  1. Cluster the data
  2. Define the outlier score as the local Mahalanobis distance
     - Suppose :$belongs to cluster N

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

     - ¯āå is the mean vector of the Nth cluster
     - -å is the covariance matrix of the Nth cluster
- Multivariate Extreme Value Analysis
  - Global Mahalanobis distance

A Post-processing Step

- Remove Small-Size Clusters

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Distance-Based Outlier Detection

- An Instance-Specific Definition
  - The distance-based outlier score of an object O is its distance to its k-th nearest neighbor

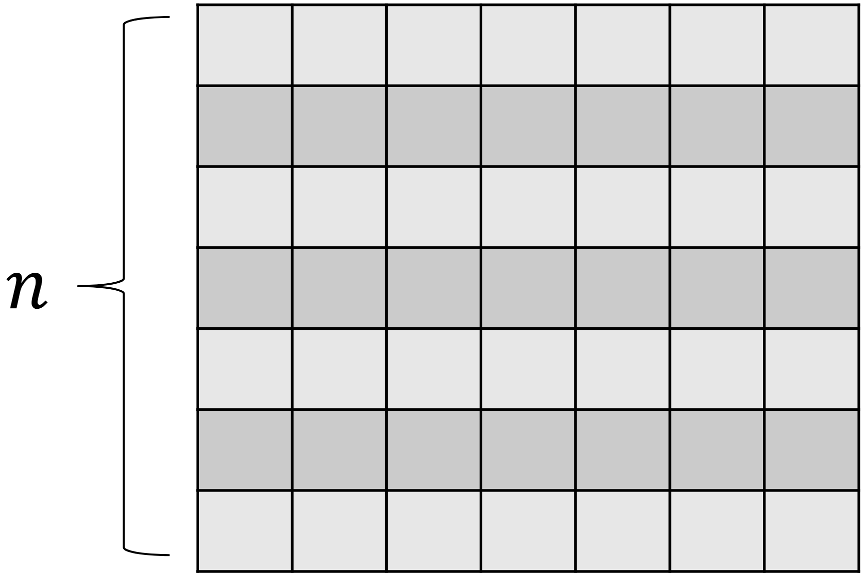
$k > 3$

Distance-Based Outlier Detection

- An Instance-Specific Definition
  - The distance-based outlier score of an object O is its distance to its k-th nearest neighbor
  - Sometimes, average distance is used
- High-computational Cost $O(n^6)$
  - Index structure
    - Effective when the dimensionality is low
  - Pruning tricks
    - Designed for the case that only the top-N outliers are needed

The Naïve Approach for Finding Top r-Outliers

1. Evaluate the $n \times n$ distance matrix



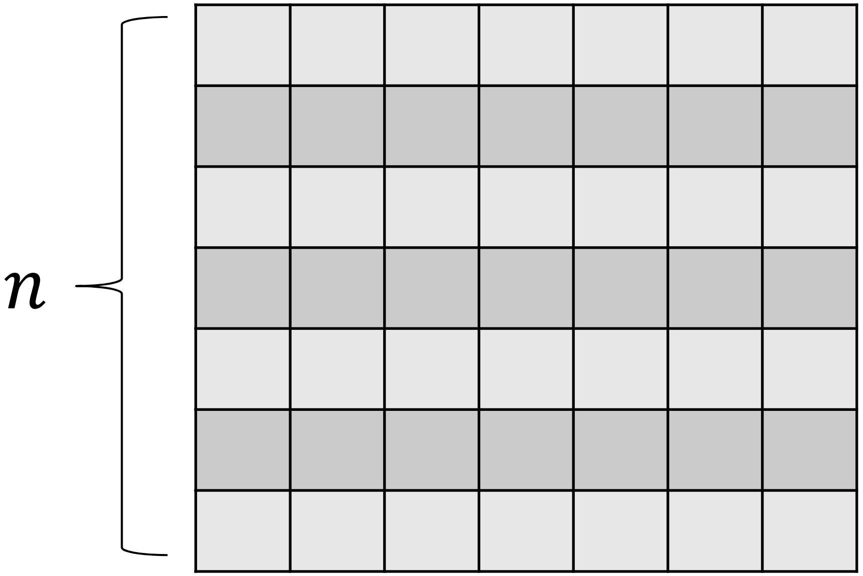
The Naïve Approach for Finding Top r-Outliers

1. Evaluate the $n \times n$ distance matrix

2. Find the k-th smallest value in each row

The Naïve Approach for Finding Top r-Outliers

1. Evaluate the $n \times n$ distance matrix



8Þ: : :5;

8Þ: : :6;

8Þ: : :7;

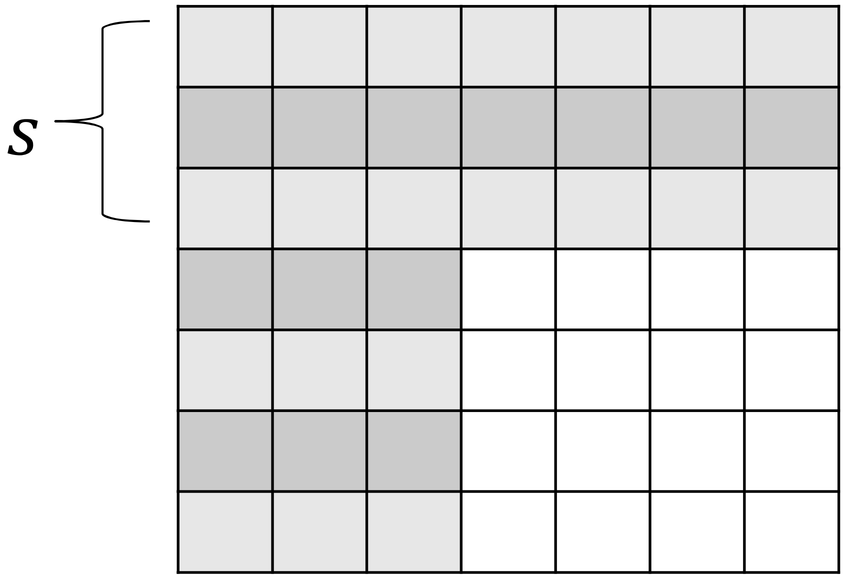
®

8Þ: : :á;

2. Find the k-th smallest value in each row
3. Choose r data points with largest $V_Þ(\cdot)$

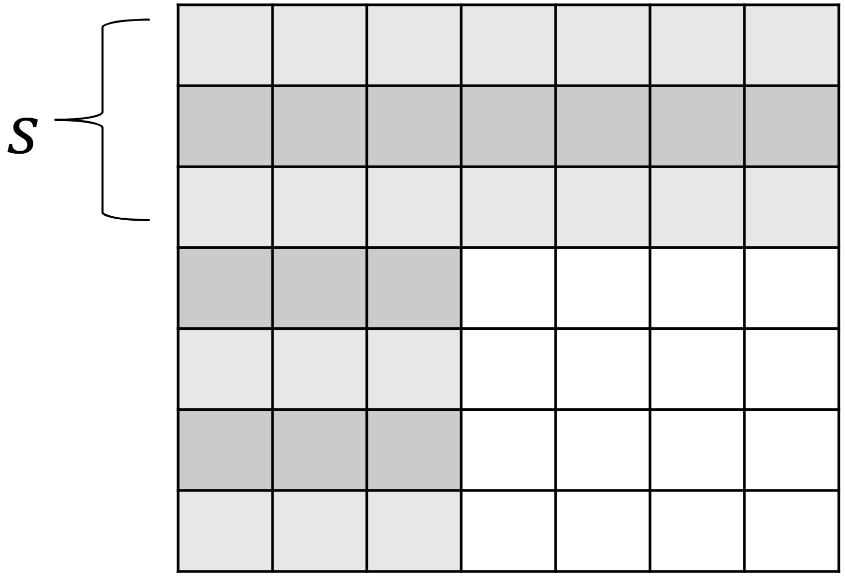
Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix



Pruning Methods—Sampling

1. Evaluate a $s \times n$ distance matrix



8Þ: : :5 ;

®

8Þ: : :æ;

âÞ(: æ >)5 R Þ(8: æ >)5

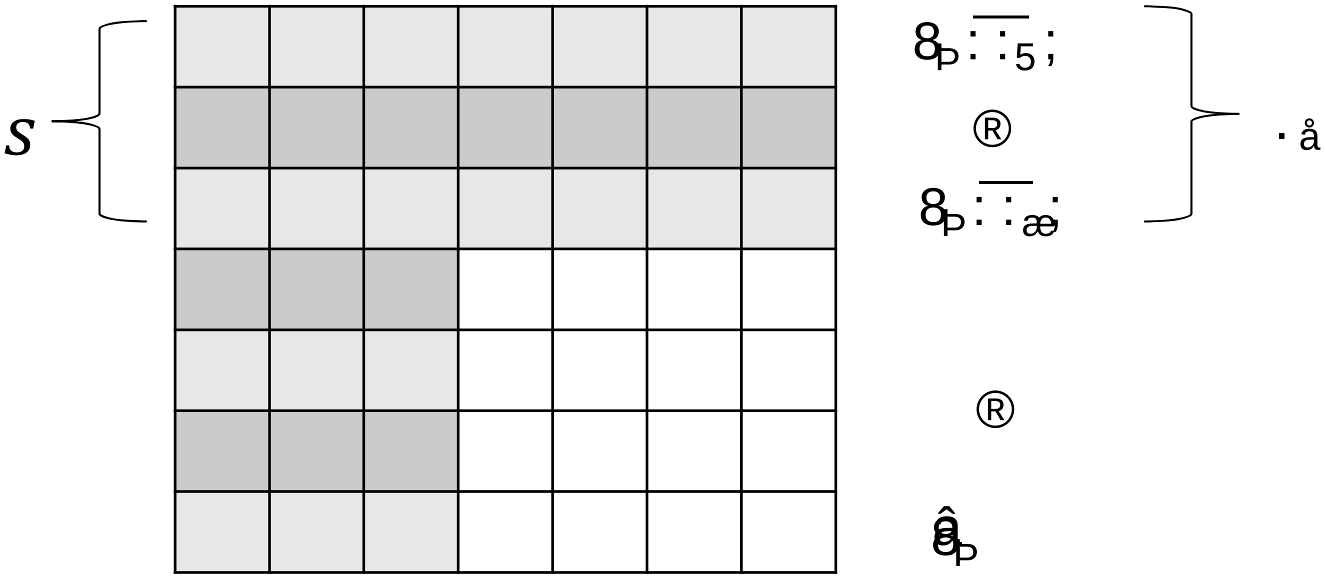
®

âÞ(: á) R Þ8: :á ;

2. Find the k-th smallest value in each row

Pruning Methods—Sampling

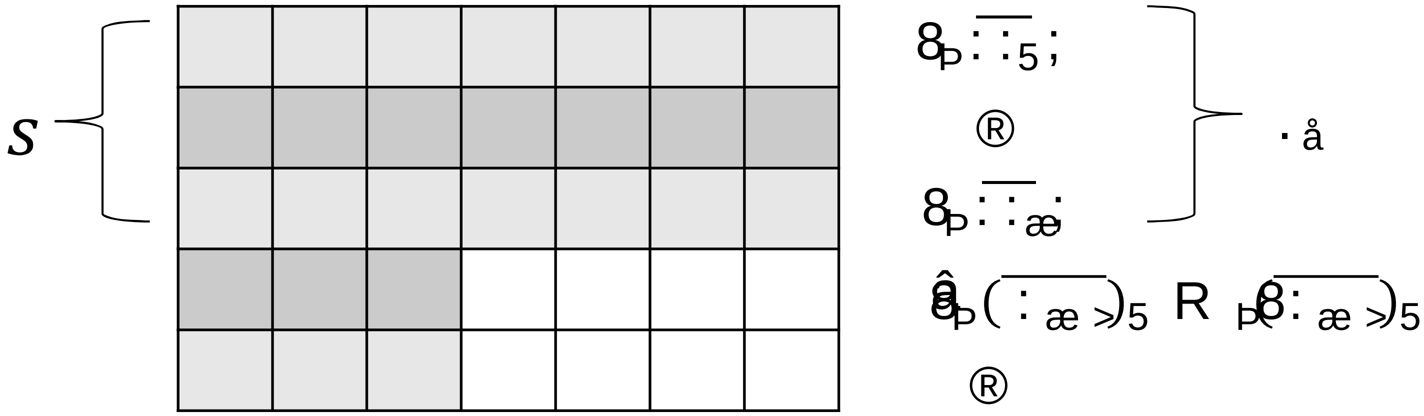
1. Evaluate a $s \times n$ distance matrix



2. Find the k-th smallest value in each row
3. Identify the r-th score in top s-rows

Pruning Methods—Sampling

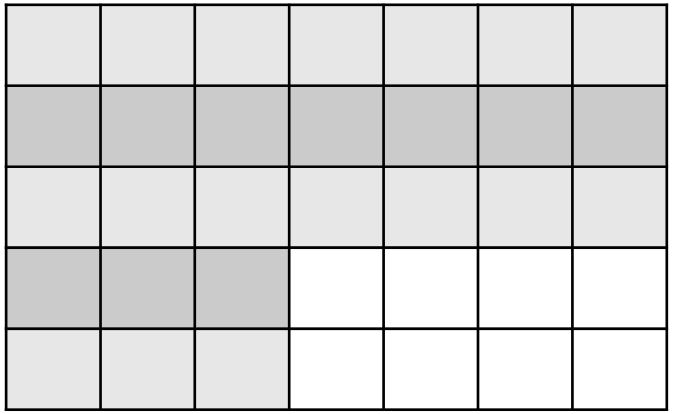
1. Evaluate a $s \times n$ distance matrix



2. Find the k-th smallest value in each row
3. Identify the r-th score in top s-rows
4. Remove points with $\widehat{V}_{Þ}(\cdot) \leq L_{å}$

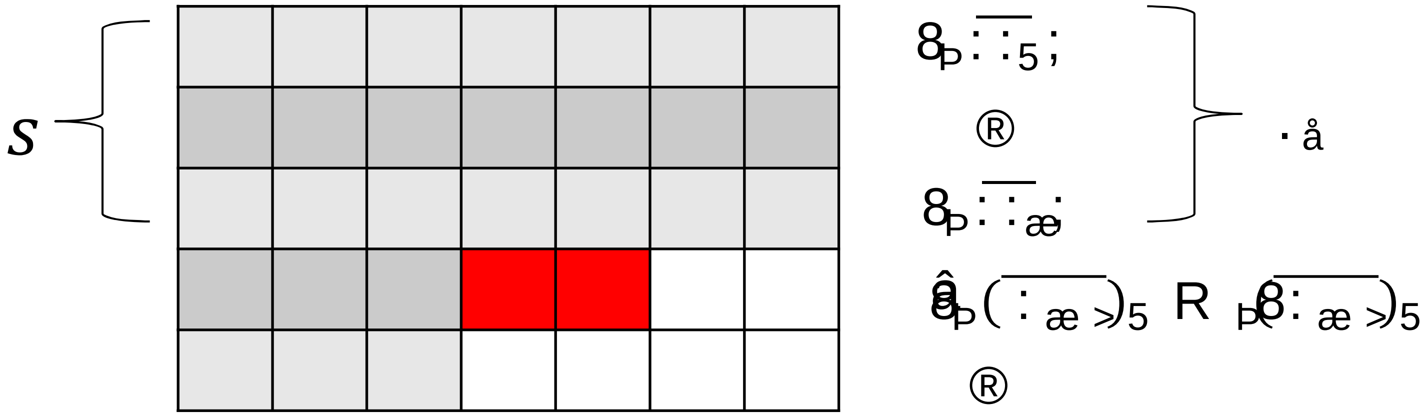
Pruning Methods—Early Termination

- When completing the empty area



Pruning Methods—Early Termination

- When completing the empty area



- Update $\widehat{V}_{Þ}(\cdot)$ when more distances are known
- Stop if $\widehat{V}_{Þ}(\cdot) \leq L_{å}$
- Update $L_{å}$ if necessary

Local Distance Correction Methods

- Impact of Local Variations

Local Outlier Factor (LOF)

- Let $V^{Þ}(\bar{X})$ be the distance of $\bar{X}$ to its k-nearest neighbor
- Let $L_{Þ}(\bar{X})$ be the set of points within the k-nearest neighbor distance of $\bar{X}$
- Reachability Distance

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

  - Not symmetric between $\bar{X}$ and $\bar{Y}$
  - If $Dist(\bar{X},\bar{Y})$ is large, $R_{Þ}(\bar{X},\bar{Y}) = Dist(\bar{X},\bar{Y})$
  - Otherwise, $R_{Þ}(\bar{X},\bar{Y}) = V^{Þ}(\bar{Y})$
    - Smoothed out by 8Þ: \$;„ more stable

Local Outlier Factor (LOF)

- Average Reachability Distance

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

- Local Outlier Factor

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

  - Larger for Outliers
  - Close to 1 for Others

- Outlier Score

$$\max_{Þ} LOF_{Þ}(\bar{X})$$

Instance-Specific Mahalanobis Distance (1)

- Define a local Mahalanobis distance for each point
  - Based on the covariance structure of the neighborhood of a data point

- The Challenge
  - Neighborhood of a data point is hard to define with the Euclidean distance
  - Euclidean distance is biased toward capturing the circular region around that point

Instance-Specific Mahalanobis Distance (2)

- An agglomerative approach for neighborhood construction
  - Add $\bar{X}$ to $L^{Þ}(\bar{X})$
  - Data points are iteratively added to $L^{Þ}(\bar{X})$ that have the smallest distance to $L^{Þ}(\bar{X})$

ƒ ” ‰ •¥ $Ð Þ • $Ð • ‹ Ö Ä: ¢ N; • @ E Ö$ P :§ „

- Instance-specific Mahalanobis score

$$LMaha_k(\overline{X}) = Maha(\overline{X}, \overline{\mu_k(X)}, \Sigma_k(\overline{X}))$$

- Outlier score • ƒ Þ š. / = D Þ : :$;

Instance-Specific Mahalanobis Distance (3)

- Can be applied to both cases

- Relation to clustering-based approaches

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Density-Based Methods

- The Key Idea
  - Determine sparse regions in the underlying data
- Limitations
  - Cannot handle variations of density

Histogram- and Grid-Based Techniques

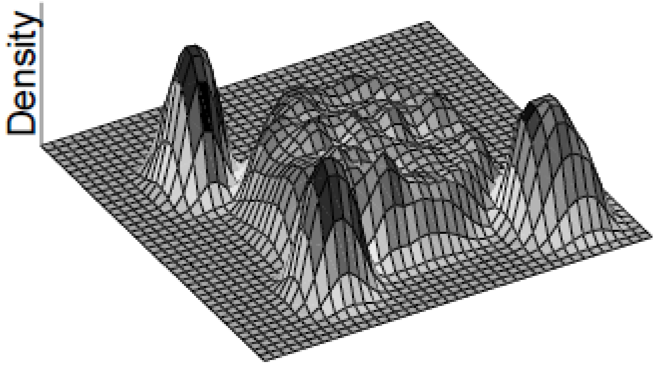
- Histogram for 1-dimensional data
  - Data points that lie in bins with very low frequency are reported as outliers

https://www.mathsisfun.com/data/histograms.html

- Grid for high-dimensional data
- Challenges
  - Size of grid
  - Too local
  - Sparsity

Kernel Density Estimation

- Given n data points $\overline{X}_5, \ldots, \overline{X}_á$



$$f(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} K(\overline{X} - \overline{X_i}).$$

  - $K(\cdot)$ is a kernel function

$$K(\overline{X} - \overline{X_i}) = \left(\frac{1}{h\sqrt{2\pi}}\right)^{d} e^{-\frac{||\overline{X}-\overline{X_i}}{2\cdot h^2}} .$$

- The density at each data point
  - Computed without including the point itself in the density computation
  - Low values of the density indicate greater tendency to be an outlier

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Information-Theoretic Models

- An Example

  ABABABABABABABABABABABABABABABABAB

  ABABACABABABABABABABABABABABABABAB

  - The 1st One: “AB 17 times”
  - C in 2nd string increases its minimum description length
- Conventional Methods
  - Fix model, then calculate the deviation
- Information-Theoretic Models
  - Fix the deviation, then learn the model
  - Outlier score of :$ increase of the model size when :$is present

Probabilistic Models

- The Conventional Method
  - Learn the parameters of generative model with a fixed size
  - Use the fit of each data point as the outlier score
- Information-Theoretic Method
  - Fix a maximum allowed deviation (a minimum value of fit)
  - Learn the size and values of parameters
  - Increase of size is used as the outlier score

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Outlier Validity

- Methodological Challenges
  - Internal criteria are rarely used in outlier analysis
  - A particular validity measure will favor an algorithm using a similar objective function criterion
  - Magnified because of the small sample solution space
- External Measures
  - The known outlier labels from a synthetic data set
  - The rare class labels from a real data set

Receiver Operating Characteristic (ROC) curve

† $\mathcal{G}$ is the set of outliers (ground-truth)

† An algorithm outputs a outlier score

† Given a threshold t, we denote the set of outliers by $\mathcal{S}(t)$

  „ True-positive rate (recall)

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

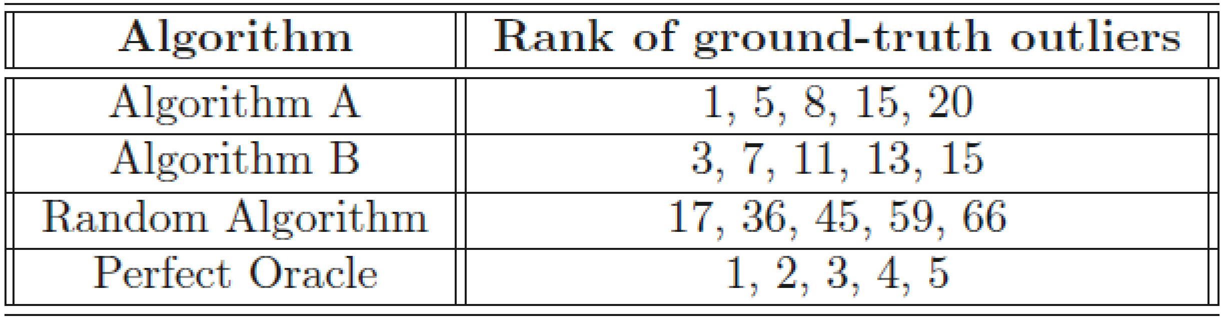
  „ The false positive rate

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

† ROC Curve

  „ Plot 6 2 (4 P) versus (2 4 : P ;

An Example



Algorithm	**Rank of ground-truth outliers**
Algorithm A	1, 5, 8, 15, 20
Algorithm B	3, 7, 11, 13, 15
Random Algorithm	17, 36, 45, 59, 66
Perfect Oracle	1, 2, 3, 4, 5

Outline

- Introduction
- Extreme Value Analysis
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Methods
- Information-Theoretic Models
- Outlier Validity
- Summary

Summary

- Extreme Value Analysis
  - Univariate, Multivariate, Depth-Based
- Probabilistic Models
- Clustering for Outlier Detection
- Distance-Based Outlier Detection
  - Pruning, LOF, Instance-Specific
- Density-Based Methods
  - Histogram- and Grid-Based, Kernel Density
- Information-Theoretic Models
- Outlier Validity
  - ROC curve