Agenda

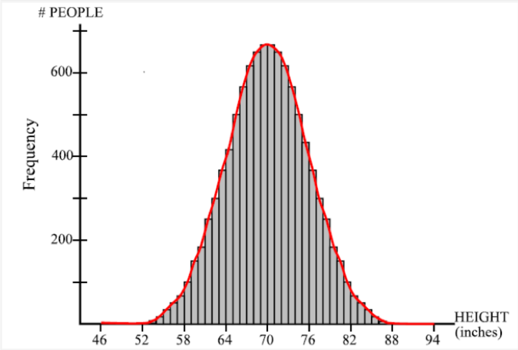
Agenda

Introduction

What is an outlier?

Hawkins (1980) "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."

- Statistics-based intuition:
  - Normal data objects follow a "generating mechanism", e.g. some given statistical process
  - Abnormal objects deviate from this generating mechanism



Introduction

Applications

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, ...)
  - Unusual symptoms or test results may indicate potential health problems of a patient
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
  - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Introduction

Applications (cont'd)

- Sports statistics
  - In many sports, various parameters are recorded for players in order to evaluate the players' performances
  - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
  - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
- Detecting measurement errors
  - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
  - Abnormal values could provide an indication of a measurement error
  - Removing such errors can be important in other data mining and data analysis tasks
  - *"One person's noise could be another person's signal."*

Introduction

Important Properties of Outlier Models

- Global vs. local approach
  - "Outlierness" regarding whole dataset (global) or regarding a subset of data (local)?
- Labeling vs. Scoring
  - Binary decision or outlier degree score?
- Assumptions about "Outlierness"
  - What are the characteristics of an outlier object?

- An object is a cluster-based outlier if it does not strongly belong to any cluster.

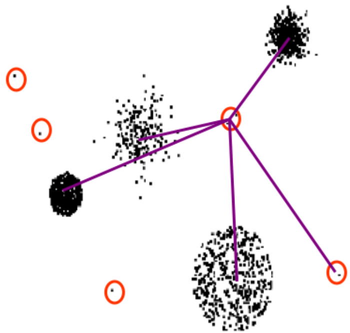
Agenda

Clustering-based Outliers

An object is a cluster-based outlier if it does not strongly belong to any cluster.

Basic Idea

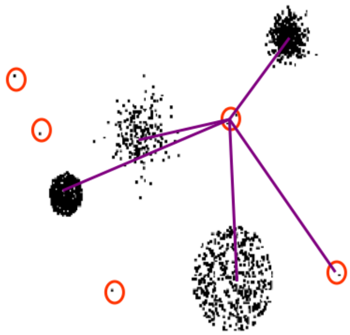
- Cluster the data into groups
- Choose points in small clusters as candidate outliers.
- Compute the distance between candidate points and non-candidate clusters.
- If candidate points are far from all other non-candidate points and clusters, they are outliers



Clustering-based Outliers

More Systematic Approaches

- Find clusters and then assess the degree to which a point belongs to any cluster
  - E.g. for k-Means, use distance to the centroid
- If eliminating a point results in substantial improvement of the objective function, we could classify it as an outlier
  - Clustering creates a model of the data and the outliers distort that model.
  - Applicable to clustering algorithms optimizing some objective function (e.g. k-means)

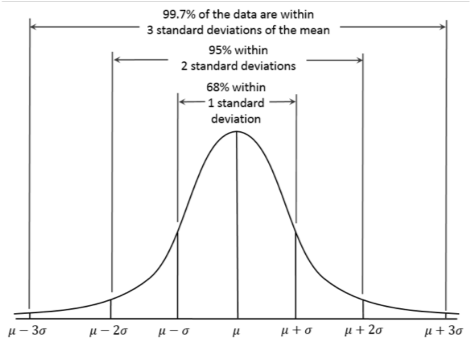


Agenda

Statistical Tests

General Idea

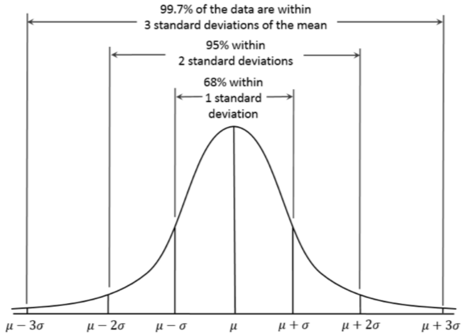
- Given a certain kind of statistical distribution (e.g., Gaussian)
- Compute the parameters assuming all data points have been generated by such a statistical distribution (e.g., mean and standard deviation)
- Outliers are points that have a low probability to be generated by the overall distribution (e.g., deviate more than 3 times the standard deviation from the mean)



Statistical Tests

Basic Assumption

- Normal data objects follow a (known) distribution and occur in a high probability region of this model
- Outliers deviate strongly from this distribution



Statistical Tests

A huge number of different tests are available differing in

- Type of data distribution (e.g. Gaussian)
- Number of variables, i.e., dimensions of the data objects (univariate/multivariate)
- Number of distributions (mixture models)
- Parametric versus non-parametric (e.g. histogram-based)

Example on the Following Slides

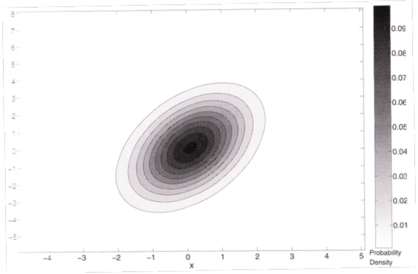
- Gaussian distribution
- Multivariate
- Single model
- Parametric

Statistical Outliers: Gaussian Distribution

Probability Density Function of a Multivariate Normal Distribution

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- μ is the mean value of all points (usually data are normalized such that $\mu = 0$)
- Σ is the covariance matrix from the mean



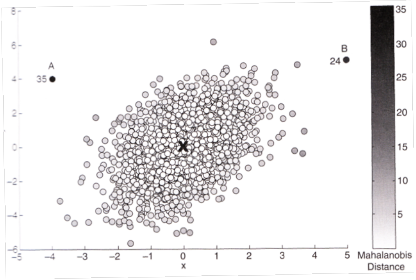
Statistical Outliers: Mahalanobis Distance

Mahalanobis Distance

Mahalanobis distance of point x to μ :

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

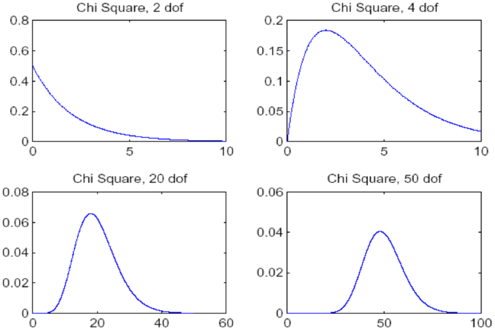
- $MDist$ follows a χ^2-distribution with d degrees of freedom (d = data dimensionality)
- Outliers: All points x, with $MDist(x, \mu) > \chi^2(0.975)$ ($\approx 3\sigma$)



Statistical Outliers: Problems

Problems

- Curse of dimensionality: The larger the degree of freedom, the more similar the *MDist* values for all points

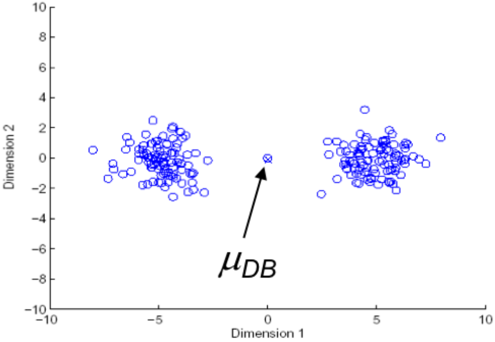


- x-axis = observed *MDist* values
- y-axis = frequency of observation

Statistical Outliers: Problems

Problems (cont'd)

- Robustness
  - Mean and standard deviation are very sensitive to outliers
  - These values are computed for the complete data set (including potential outliers)
  - The *MDist* is used to determine outliers although the *MDist* values are influenced by these outliers



Statistical Outliers: Problems

Problems (cont'd)

- Data distribution is fixed
- Low flexibility (if no mixture models)
- Global method

Agenda

Distance-Based Approaches

General Idea

Judge a point based on the distance(s) to its neighbors (Several variants proposed)

Basic Assumption

- Normal data objects have a dense neighborhood
- Outliers are far apart from their neighbors, i.e., have a less dense neighborhood

Distance-Based Approaches

D(ϵ, π)-Outlier [6]

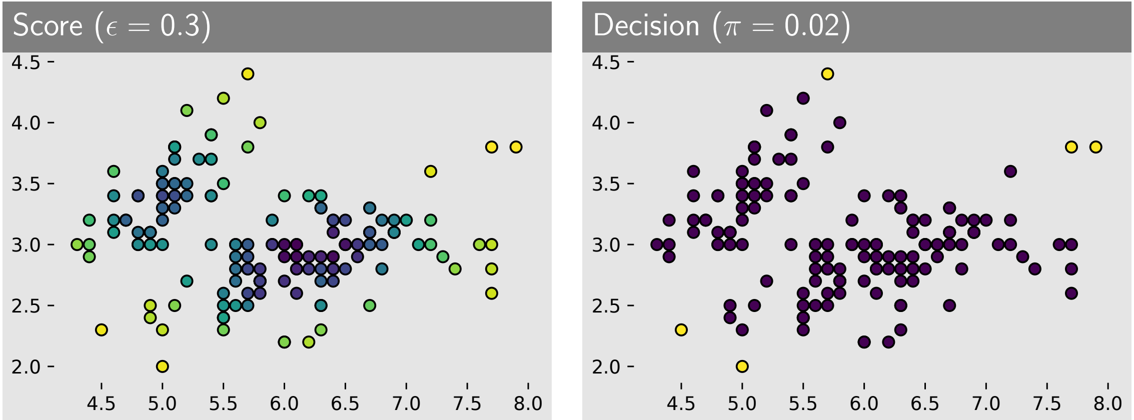
- Given: distance threshold $\epsilon \geq 0$, fraction threshold $0 < \pi \leq 1$
- A point p is considered an outlier if at most fraction π of all points in D have a distance to q less than or equal ϵ.

$$OutlierSet(\epsilon, \pi) = \left\{ p \in D \;\middle|\; \frac{|\{q \in D \mid dist(p,q) < \epsilon\}|}{|D|} \leq \pi \right\}$$

where $dist(\cdot, \cdot)$ is a distance measure.

[6] Han, J., Kamber, M. & Pei, J. (2012). *Data mining concepts and techniques*

Distance-Based Approaches: $D(\epsilon, \pi)$ Example



Distance-Based Approaches: kNN

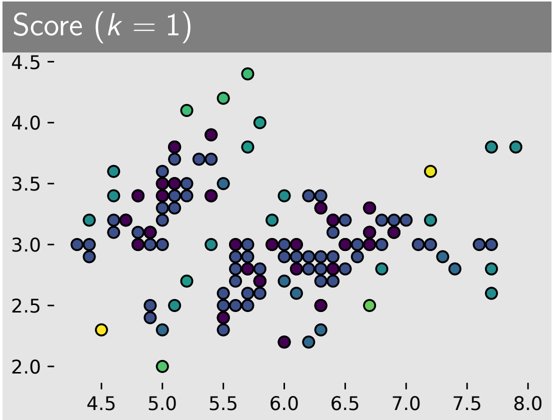
Outlier scoring based on kNN distances

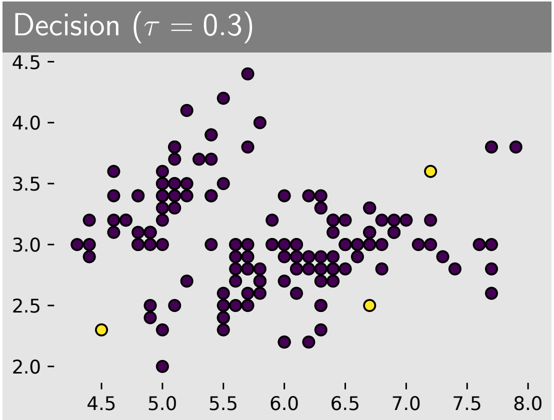
General models: Take the kNN distance of a point as its outlier score

Decision

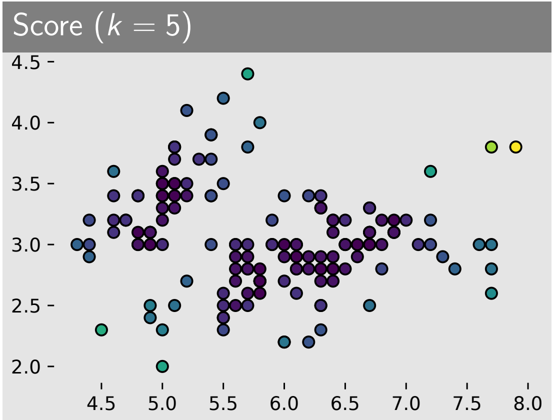
k-distance above some threshold $\tau \iff$ Outlier.

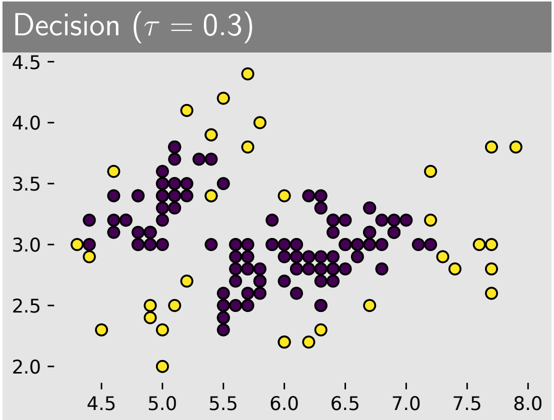
Distance-Based Approaches: kNN Example





Distance-Based Approaches: kNN Example





kNN: Problems

Problems

- Highly sensitive towards k:
  - Too small k: small number of close neighbors can cause low outlier scores.
  - Too large: all objects in a cluster with less than k objects might become outliers.
- cannot handle datasets with regions of widely different densities due to the global threshold

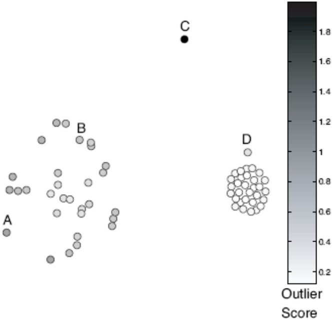


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Image Source: P. Tan, M. Steinbach, V. Kumar (2006). *Classification: basic concepts, decision trees, and model evaluation. Introduction to data mining*, 1, 145-205.

Agenda

Density-Based Approaches

General Idea

- Compare the density around a point with the density around its local neighbors.
- The relative density of a point compared to its neighbors is computed as an outlier score.
- Approaches also differ in how to estimate density.

Basic Assumption

- The density around a normal data object is similar to the density around its neighbors.
- The density around an outlier is considerably different to the density around its neighbors.

Density-Based Approaches

Problems

- Different definitions of density: e.g., #points within a specified distance ϵ from the given object
- The choice of ϵ is critical (too small $\implies$ normal points considered as outliers; too big $\implies$ outliers considered normal)
- A global notion of density is problematic (as it is in clustering); fails when data contain regions of different densities

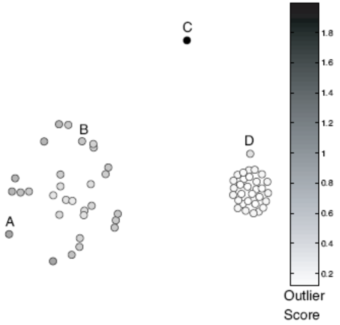


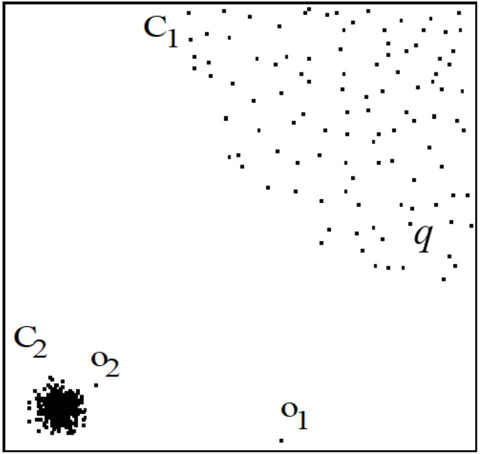
Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

D has a higher absolute density than A but compared to its neighborhood, Ds density is lower.

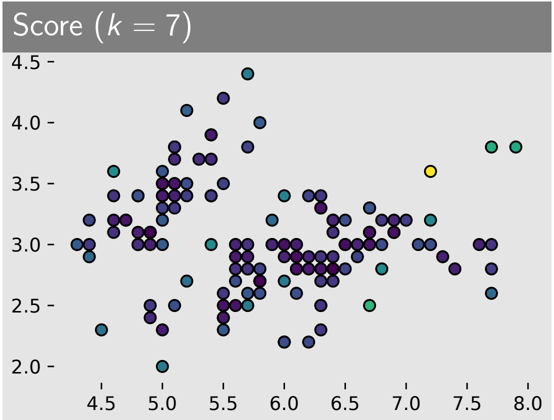
Density-Based Approaches

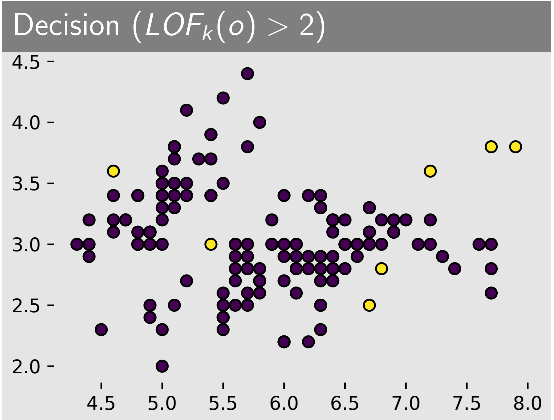
Failure Case of Distance-Based

- $D(\epsilon, \pi)$: parameters ϵ, π cannot be chosen s.t. o_2 is outlier, but none of the points in C_1 (e.g. q)
- kNN-distance: kNN-distance of objects in C_1 (e.g. q) larger than the kNN-distance of o_2.



Density-Based Approaches





Density-Based Approaches

Solution

Consider the relative density w.r.t. to the neighbourhood.

Model

- Local Density (ld) of point p (inverse of avg. distance of kNNs of p)

$$ld_k(p) = \left(\frac{1}{k} \sum_{o \in kNN(p)} dist(p, o) \right)^{-1}$$

- Local Outlier Factor (LOF) of p (avg. ratio of lds of kNNs of p and ld of p)

$$LOF_k(p) = \frac{1}{k} \sum_{o \in kNN(p)} \frac{ld_k(o)}{ld_k(p)}$$

Density-Based Approaches

Extension (Smoothing factor)

- Reachability "distance"

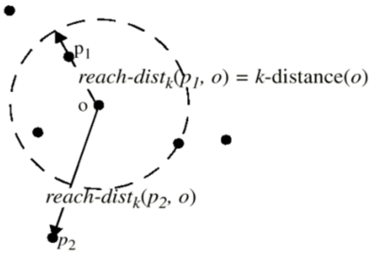
$$rd_k(p, o) = \max\{kdist(o), dist(p, o)\}$$

- Local reachability distance lrd_k

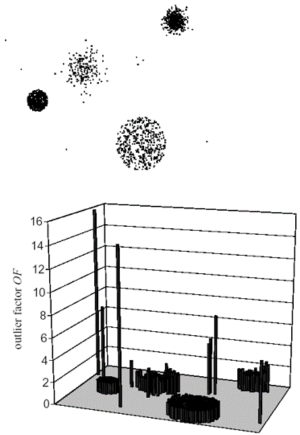
$$lrd_k(p) = \left(\frac{1}{k} \sum_{o \in kNN(p)} rd(p, o) \right)^{-1}$$

- Replace ld by lrd

$$LOF_k(p) = \frac{1}{k} \sum_{o \in kNN(p)} \frac{lrd_k(o)}{lrd_k(p)}$$



Density-Based Approaches



Discussion

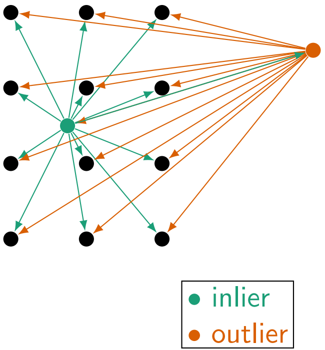
- $LOF \approx 1 \implies$ point in cluster
- $LOF \gg 1 \implies$ outlier.
- Choice of k defines the reference set

Agenda

Angle-Based Approach

General Idea

- Angles are more stable than distances in high dimensional spaces
- *o outlier* if most other objects are located in similar directions
- *o no outlier* if many other objects are located in varying directions



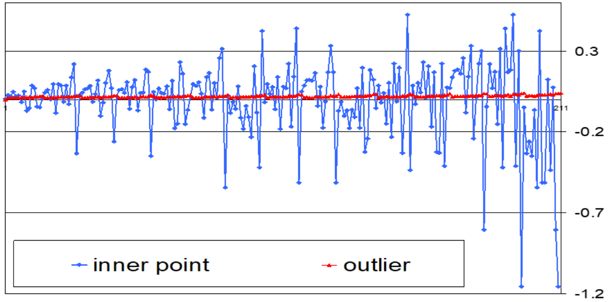
Basic Assumption

- Outliers are at the border of the data distribution
- Normal points are in the center of the data distribution

Angle-Based Approach

Model

- Consider for a given point p the angle between $\overrightarrow{px}$ and $\overrightarrow{py}$ for any two x, y from the database
- Measure the variance of the angle spectrum



Angle-Based Approach

Model (cont'd)

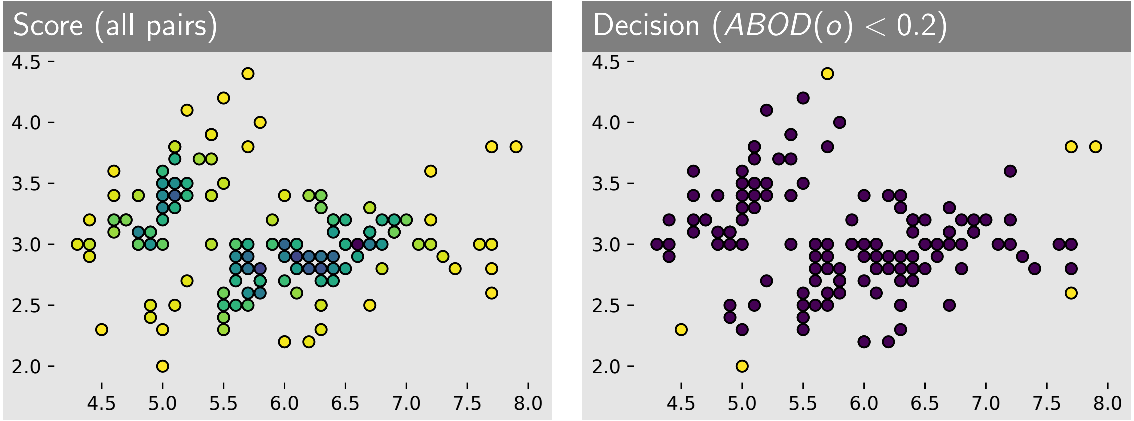
- Weighted by the corresponding distances (for lower dimensional data sets where angles are less reliable)
  Angle-based Outlier Detection[7]:

$$ABOD(p) = \mathrm{VAR}_{x,y \in D}\left(\frac{1}{\|\overrightarrow{xp}\|_2 \|\overrightarrow{yp}\|_2} \cos\left(\overrightarrow{xp}, \overrightarrow{yp}\right)\right) = \mathrm{VAR}_{x,y \in D}\left(\frac{\langle \overrightarrow{xp}, \overrightarrow{yp} \rangle}{\|\overrightarrow{xp}\|_2^2 \|\overrightarrow{yp}\|_2^2}\right)$$

- Small ABOD $\iff$ outlier

[7] Kriegel, Hans-Peter, Matthias Schubert, and Arthur Zimek. "Angle-based outlier detection in high-dimensional data." Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008.

Angle-Based Approaches



Agenda

Tree-Based Approaches: Isolation Forest

General Idea

Outlierness = how easy it is to separate a point from the rest by random space splitting?

Basic Assumption

- Anomalies are the minority consisting of fewer instances
- Anomalies have attribute-values that are very different from those of normal instances

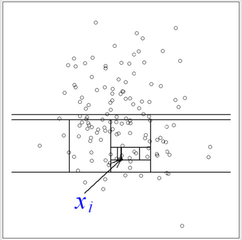
Tree-Based Approaches

Isolation Tree - Training

1. Randomly select one dimension
2. Randomly select a split position in that dimension
3. Repeat until: a) only one point left or b) height reaches predefined threshold h

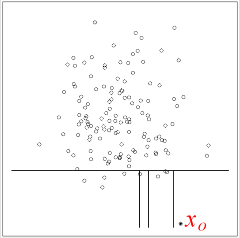
Example

Normal point path length=10 splits



Example

Outlier point path length=4 splits

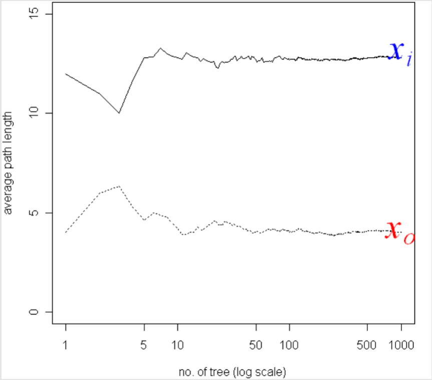


Tree-Based Approaches: Training

Isolation Forest - Training

1. Random sample ψ points, build an isolation tree
2. Repeat for t times $\Rightarrow$ a forest of t isolation trees

Average path lengths converge



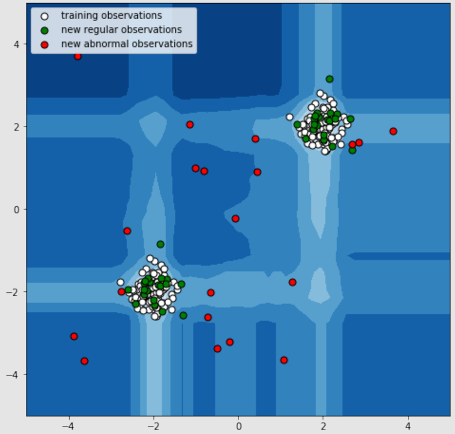
Tree-Based Approaches: Anomaly Score

- Let $h(x)$ be the path length of x on an isolation tree, and estimate $E(h(x))$ by the *average path length* among t isolation trees.
- Let $c(\psi) = 2H(\psi - 1) - 2(\psi - 1)/\psi$, which is the expected path length of unsuccessful search in BST of ψ points; $H(\cdot)$ is the harmonic number.
- Define the anomaly score of a point x as $s(x) = 2^{-\frac{E(h(x))}{c(\psi)}}$
- Observe $s(x) \in (0, 1)$
  - $E(h(x)) \to c(\psi)$ yields $s \to 0.5$,
  - $E(h(x)) \to 0$ yields $s \to 1$,
  - $E(h(x)) \to n - 1$ yields $s \to 0$.
- Usually, set $s = 0.5$ as threshold, i.e. the average of the expected path length

Tree-Based Approaches: Discussion

- Advantages:
  - Anomaly score between 0 and 1
  - Very efficient, especially on large dataset
  - A model (the forest) is learned from the training dataset
  - Easy for parallelization
  - Can be adapted to categorical data
- Disadvantages:
  - Only detects global outliers (of course, follow-up approaches are available)
  - Not efficient on high-dimensional data

iForest anomaly score contour



Recap - Outlier Detection

- Properties: global vs. local, labeling vs. scoring
- *Clustering-Based* Outliers: Identification as non-(cluster-members)
- *Statistical* Outliers: Assume probability distribution; outliers = unlikely to be generated by distribution
- *Distance-Based* Outliers: Distance to neighbors as outlier metric
- *Density-Based* Outliers: Relative density around the point as outlier metric
- *Angle-Based* Outliers: Angles between outliers and random point pairs vary only slightly