Outlier Detection

Contents of this Chapter

Introduction

Extreme value analysis [Aggarwal section 8.3]

Probabilistic methods [section 8.3]

Clustering-based methods [section 8.4]

Distance-based methods [section 8.5]

Density-based methods [section 8.6]

Introduction

Overview

"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. " (Hawkins)

- Clustering discovers groups of similar objects. Outlier detection seeks individual objects that are different from all clusters.
- Most outlier detection methods create a model of normal patterns. Outliers are the objects that do not fit within this normal model.
- Objects get assigned an outlier score: the higher, the more likely the object is an outlier.

🔍 Outlier detection is unsupervised, i.e. we have no examples of outliers.

Introduction

Overview

Outliers are also referred to as anomalies, abnormalities , discordants, deviants.

Applications of outlier detection

- Data cleaning
  remove outliers.
- Fraud detection
  abnormal behavior may indicate fraud.
- Network intrusion detection
  abnormal traffic may indicate intrusion.

Introduction

Overview

- If examples of outliers are available, can formulate the problem as a classification task.
- Supervised classification tends to be more accurate than unsupervised outlier detection, but it cannot adjust so well to changing patterns of normal data and outliers.
- Two alternative tasks of outlier detection:

  (1) Detect all outliers
  convert outlier score into a Boolean decision.

  (2) Rank objects in decreasing order of their outlier score
  typically, only interested in top k outlier scores.

Introduction

Types of Outliers

- Point outliers
- Contextual outliers
- Collective outliers

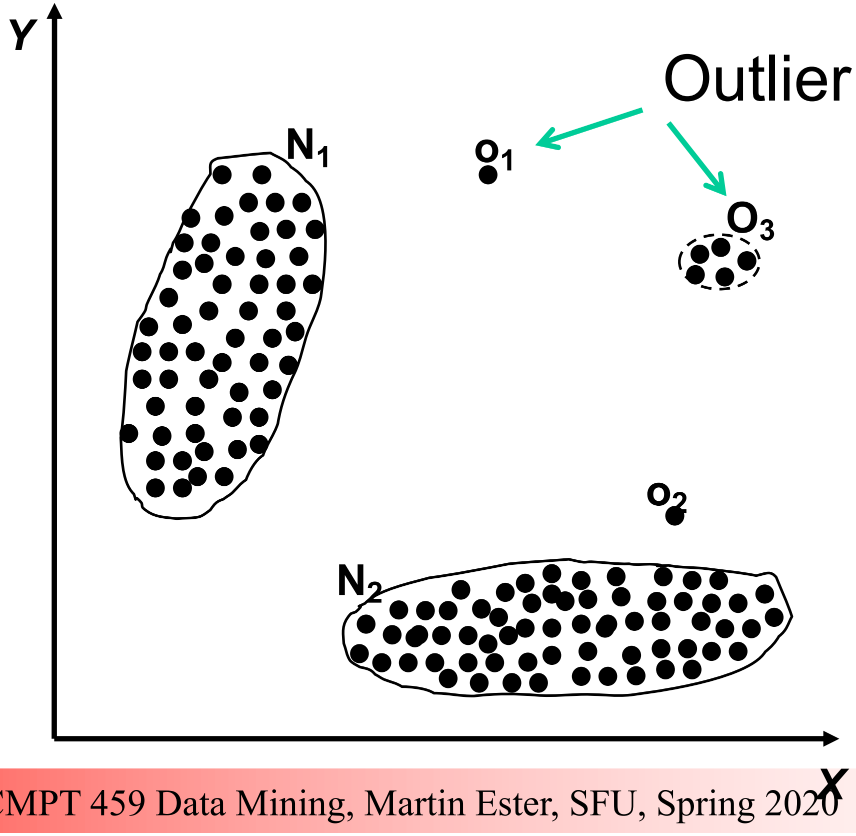
→ Most methods detect point outliers.

Introduction

Point Outliers

- Point outlier

An individual object that deviates significantly from the rest of the data set.

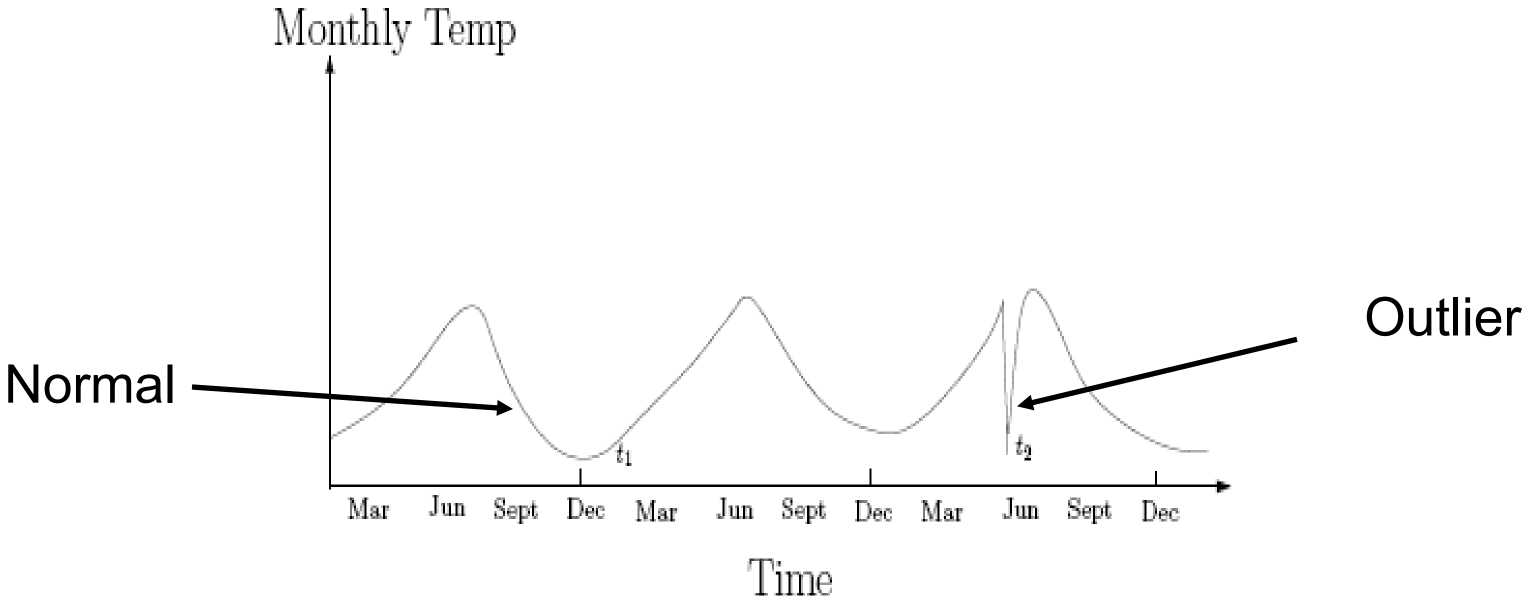


Introduction

Contextual Outliers

- Contextual outlier

An individual object that deviates significantly from its context within the data set.

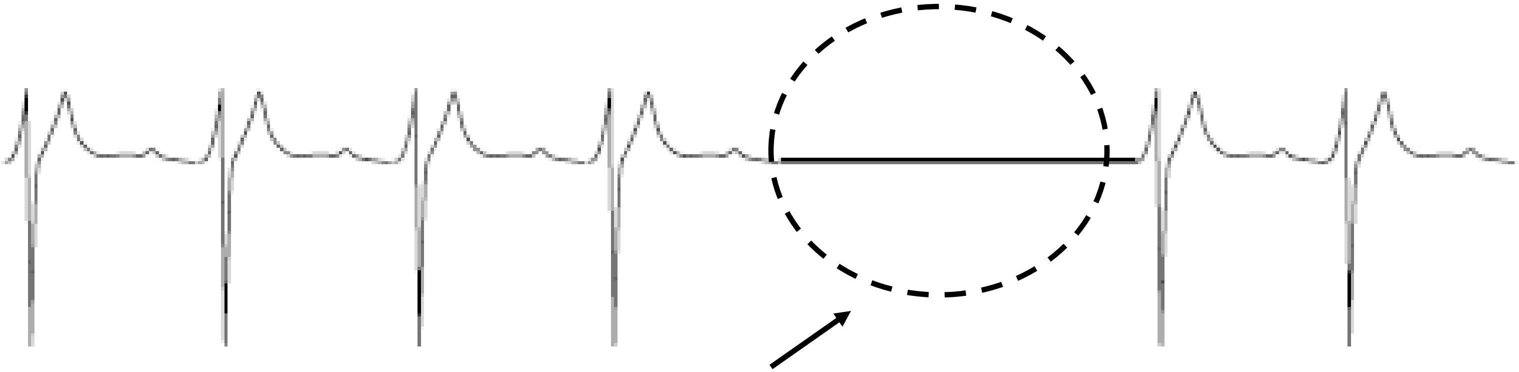


Introduction

Collective Outliers

- Collective outliers

A set of related objects that deviates significantly from the data set.



Outlying subsequence

Extreme Value Analysis

Method

- Assume that all data has been generated from a probability distribution of known type, e.g. Gaussian distribution:

$$P(x) = \frac{1}{\sqrt{(2\pi)^d \, |\sum|}} e^{\frac{1}{2} \cdot (x-\mu)^T \cdot (\sum)^{-1} \cdot (x-\mu)}$$

- Parameters of the distribution known from domain knowledge or estimated from data.

- Outlier: object within tail of probability distribution.

→ Data may not follow the assumed distribution!

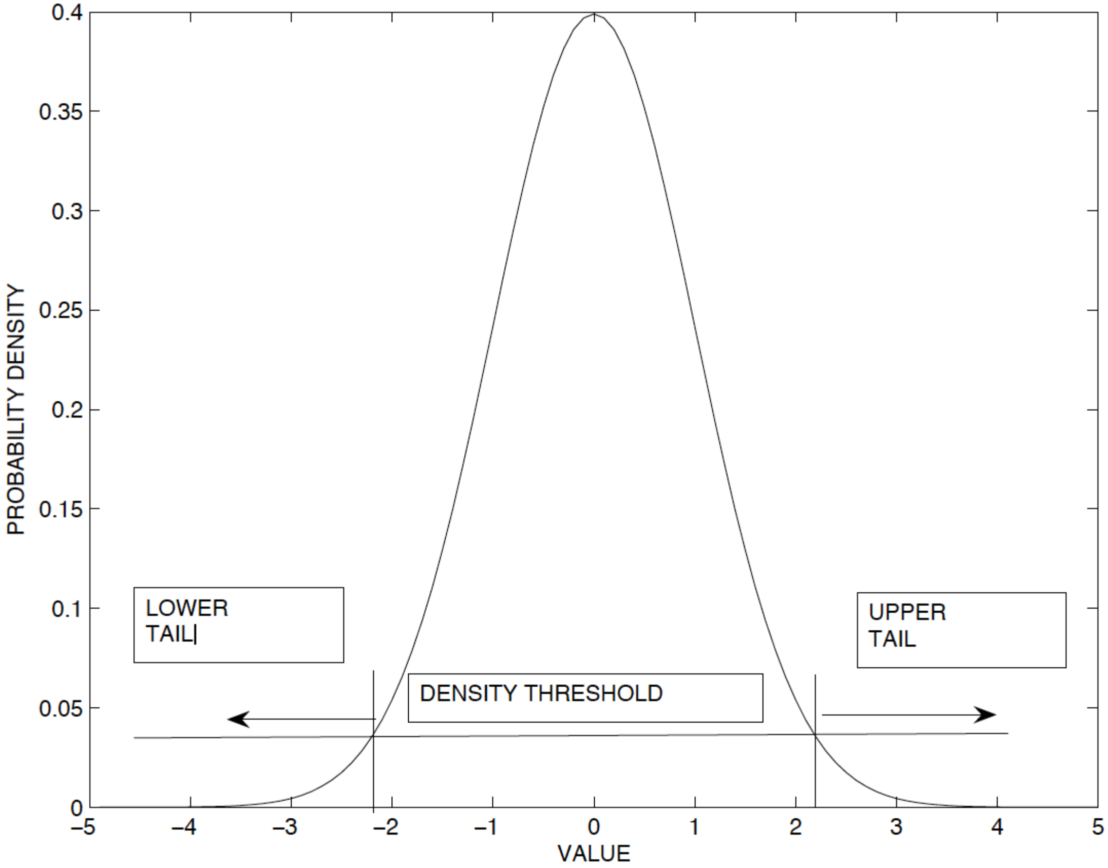
Extreme Value Analysis

Method

- Z-number as outlier score

$$z = \frac{x - \mu}{\sigma}$$

•If the absolute value of the Z – number is greater than 3, the object is considered as an extreme value (outlier).

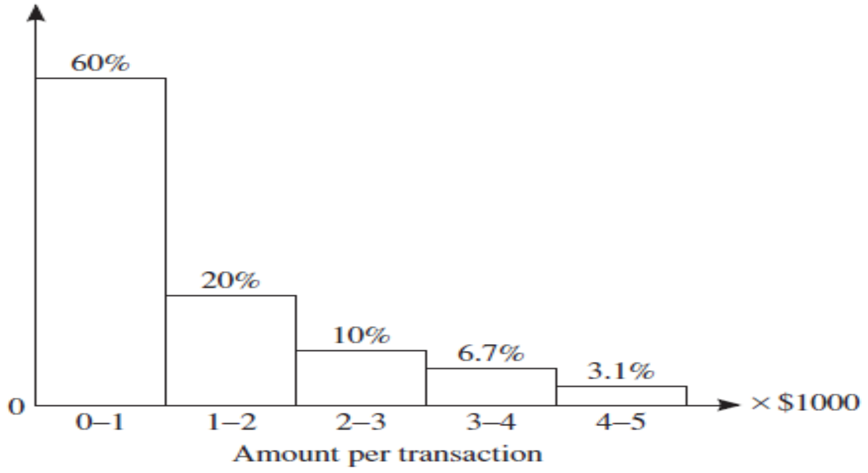


Extreme Value Analysis

Histogram-based Method

- Non-parametric, i.e. does not assume a given probability distribution.

- Outlier: object that has less than t% other objects with more extreme values.



→ Hard to choose an appropriate bin size for histogram.

Probabilistic Outlier Detection

Method

- Generalizes Extreme Value Analysis.
- Assume that data has been generated from a mixture of multiple probability distributions.

$$P(x) = \sum_{i=1}^{k} P(C_i) \cdot P(x \mid C_i)$$

- Often, assume a mixture of *k* Gaussians.
- Outliers are defined as those objects that are highly unlikely to be generated by this model.

Clustering-based Outlier Detection

Method

- Assumption

  Normal objects belong to large clusters, while outliers do not belong to any of the clusters or form very small clusters.
- Density-based clustering

  Outliers are not included in any cluster.

  outlier score: (depends on) density in neighborhood
- Distance-based clustering

  Outliers are objects with largest distances from cluster representative.

  outlier score: distance from nearest cluster representative

→ Result depends on chosen clustering algorithm.

Distance-based Outlier Detection

Introduction

- Outliers data are far away from the “crowded regions”.
- Approach 1

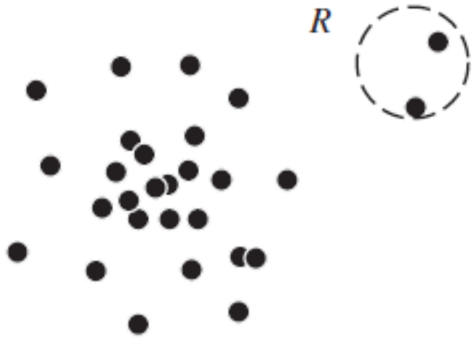
  Count percentage of objects within given distance *r,*
  the fewer objects, the higher the outlier score.
- Approach 2
  Compute distance to *k*th-nearest neighbor,
  the higher the distance, the higher the outlier score.

Distance-based Outlier Detection

Methods

- Definition 1: o is distance-based outlier

$$\frac{\|\{\mathbf{o}' \mid dist(\mathbf{o}, \mathbf{o}') \leq r\}\|}{\|D\|} \leq \pi,$$



- Definition 2: o is distance-based outlier

$$knn\_dist(o) \geq \omega$$

knn_dist(*o*) is the distance from *o* to its *k*th-nearest neigbor.

→ There can be multiple objects that have *knn_dist*(*o*) from *o*.

Distance-based Outlier Detection

Challenges

- How to set the parameters of the method?

  Results depend on proper parameter setting.

- How to efficiently retrieve the neighborhood of an object?

  Without suitable index support, complexity is O(n).

- How to efficiently score only the top outliers?
  Reduce the time required for the k-nearest neighbor distance computations by ruling out objects quickly that are obviously non-outliers (even with approximate computation).

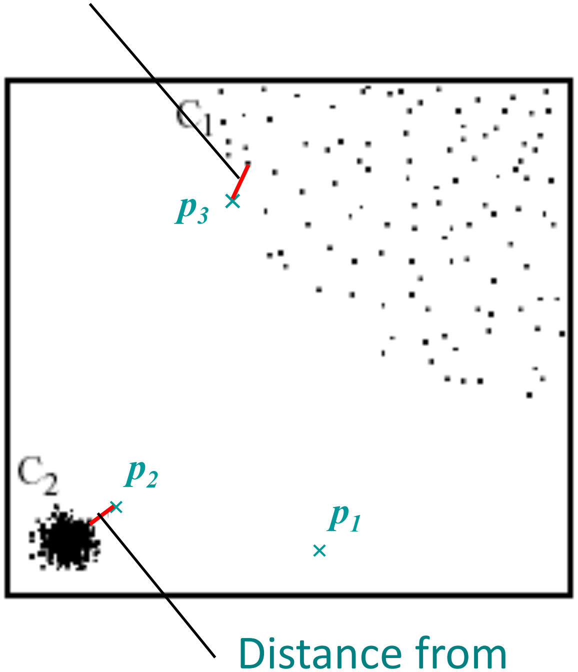
Distance-based Outlier Detection

Challenges

- How to deal with locally different densities?

The nearest neighbor distance of many objects in the sparser cluster C_1 is at least as large as the nearest neighbor distance of outlier p_2.

Distance from p_3 to nearest neighbor



Distance from p_2 to nearest neighbor

Density-based Outlier Detection

Local Outlier Factor

- Relevant distance of objects should be computed in a normalized way, relative to its local distance distribution.
- Inspiration from OPTICS algorithm.
- *Reachability distance* of object *p* relative to object *o*

$$R_k(p,o) = \max\{dist(p,o), knn\_\text{dist}\ (o)\}$$

- If *o* is in dense region and the distance between *o* and *p* is large, the reachability distance of *p* to *o* is the actual distance.
- If the distance between *p* and *o* is small, then the reachability distance is "smoothed out" by the *k*-nearest neighbor distance of *o*.

Density-based Outlier Detection

Local Outlier Factor

- Average reachability distance of object *p* to its neighboring objects

$$AR_k(p) = \left. \sum_{o \in L_k(p)} \mathrm{R}_k(p,o) \middle/ |L_k(p)| \right.$$

where $L_k(p)$ contains all objects within *kNN* distance from p

- Local Outlier Factor: mean ratio of $\mathrm{AR_k(p)}$ and $\mathrm{AR_k}$ (o) for *o* in the neigborhood of *p*.

$$LOF_k(p) = \frac{\sum_{o \in L_k(p)} AR_k(p) / AR_k(o)}{|L_k(p)|}$$

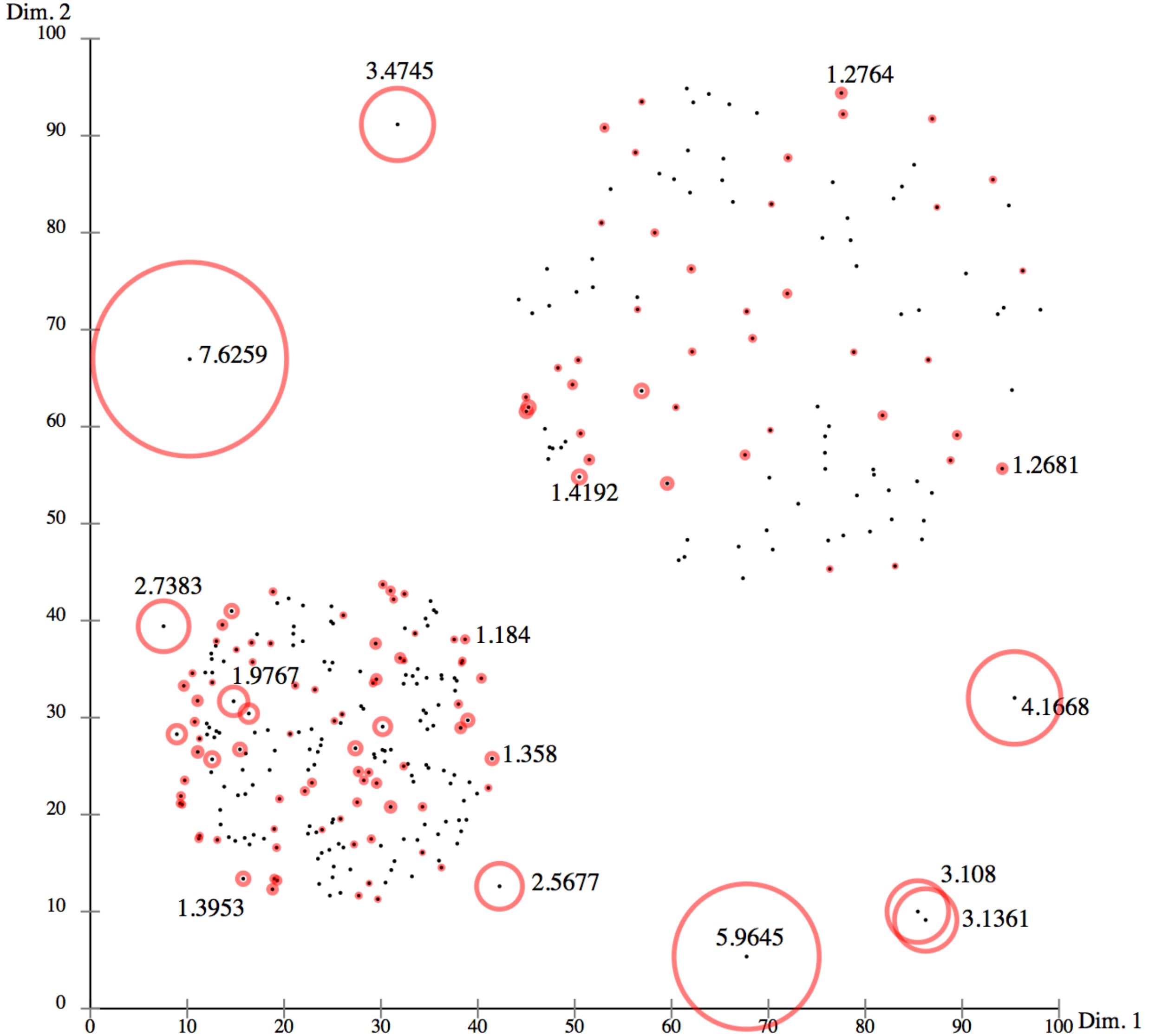
Density-based Outlier Detection

Local Outlier Factor

- LOF values for the objects in a cluster are often close to 1 when the cluster objects are homogeneously distributed.
- LOF values of outliers will be much higher, the more the density of the object deviates from its neigborhood.
- How to determine the parameter *k*?
  k too small: not robust enough
  k too large: not local enough
- If some outlier *p* is known, choose *k* that maximizes value of LOFk (*p*).

→ LOF detects contextual outliers

Density-based Outlier Detection



Density-based Outlier Detection

Kernel Density Estimation

- Non-parametric method, i.e. makes no assumption on the data distribution.
- Estimate the density in a neigborhood by the mean influence of all objects x_i on that neigborhood.
- Influence measured by kernel function K_h.
- Density at a given point in the data space is estimated as the sum of the smoothed values of kernel functions:

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)$$

Density-based Outlier Detection

Kernel Density Estimation

- Kernel functions define the relevant neigborhood of a point in data space.
- For most smooth kernel functions, with increasing data set size, the estimate converges to the true density.
- Gaussian kernel

$$K_h(x - x_i) = \left(\frac{1}{\sqrt{2\pi} \cdot h} \right)^d e^{-\|x - x_i\|^2 / (2h^2)}$$

- Which kernel function? Which kernel width *h*?

Density-based Outlier Detection

Kernel Density Estimation

- The density at each object is computed without including the object itself.
- Outlier score: (depends on) the density estimate.
- Low values of the density indicate greater tendency to be an outlier.

Density-based Outlier Detection

Discussion

- Kernel density estimation does not require data to follow a known distribution.
- Is usually robust to the choice of the kernel width *h*.
- Does not work well for high-dimensional data
  accuracy of the density estimation degrades with increasing dimensionality.
- Does not work well if local density varies greatly
  kernel width is global.