STEVENS INSTITUTE of TECHNOLOGY
School of Business

FE 582

Lecture 9: Outlier Detection

Dragos Bozdog

Outlier Analysis - Introduction

- An outlier is a data point that is very different from most of the remaining data.

"*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.*"

- Outliers complementary concept to clusters.

- *Abnormalities*, *discordants*, *deviants*, or *anomalies.*

Outlier Analysis - Introduction

- Outliers have numerous applications in many data mining scenarios:

  - *Data cleaning*
  - *Credit card fraud*
  - *Network intrusion detection*

- Most outlier detection methods create a model of normal patterns.

- Outliers are defined as data points that do not naturally fit within this normal model.

  - quantified by a numeric value: *outlier score*.

Outlier Analysis - Introduction

- Most outlier detection algorithms produce an output that can be one of two types:

1. *Real-valued outlier score:*

Higher values of the score make it more (or, in some cases, less) likely that a given data point is an outlier.

2. *Binary label:*

Binary output, indicating whether or not a data point is an outlier.

Outlier Analysis - Introduction

- The generation of an outlier score requires the construction of a model of the normal patterns.

- Key models for outlier analysis:

1. Extreme values:

- A data point is an extreme value, if it lies at one of the two ends of a probability distribution (univariate).
- Can be defined for multidimensional data by using a multivariate probability distribution.
- These are very *specialized* types of outliers.

Outlier Analysis - Introduction

2. *Clustering models:*
- Clustering is considered a complementary problem to outlier analysis
- Many clustering models determine outliers as a side-product of the algorithm.

3. *Distance-based models:*
- *k*-nearest neighbor distribution of a data point is analyzed to determine whether it is an outlier.
- Distance-based models can be considered a more fine-grained and instance-centered version of clustering models.

4. *Density-based models:*
- local density of a data point is used to define its outlier score.
- the local density at a given data point is low only when its distance to its nearest neighbors is large.

Outlier Analysis - Introduction

5. Probabilistic models:

- The steps are similar to those of clustering algorithms, except:
- EM algorithm is used for clustering
- The probabilistic fit values are used to quantify the outlier scores of data points (instead of distance values).

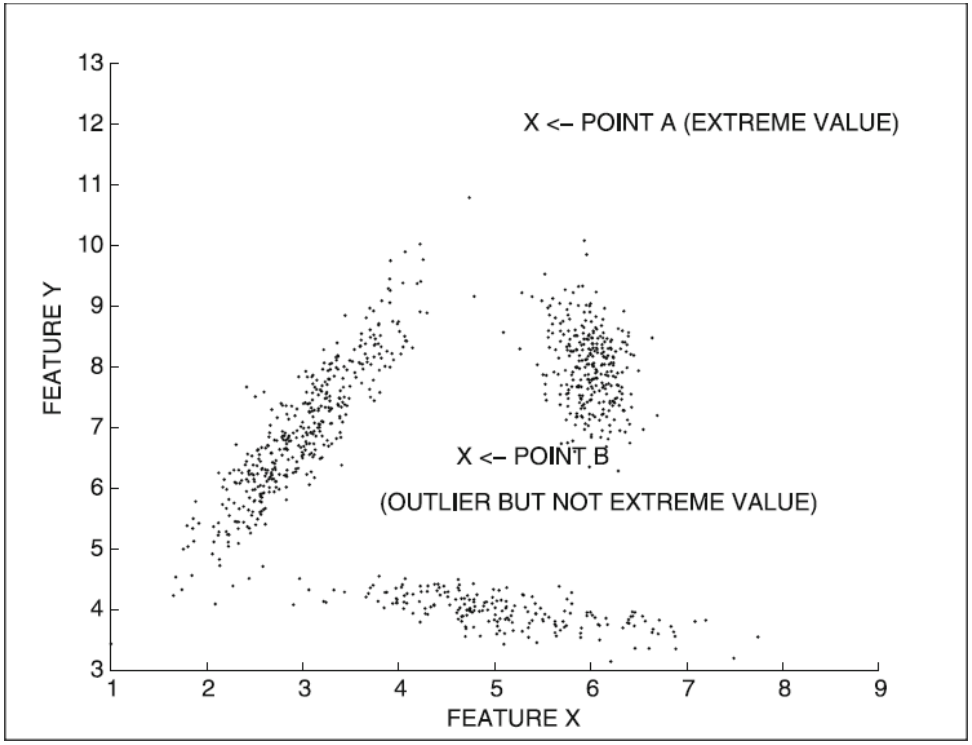
6. *Information-theoretic models:*

- constrain the maximum deviation allowed from the normal model and then examine the difference in space requirements for constructing a model with or without a specific data point.
- If the difference is large, then this point is reported as an outlier.

Outlier Analysis - Extreme Value Analysis

- These type of outliers correspond to the *statistical tails* of probability distributions.

- The most *isolated* point in the data set should be considered an outlier from a generative perspective.



Outlier Analysis - Extreme Value Analysis

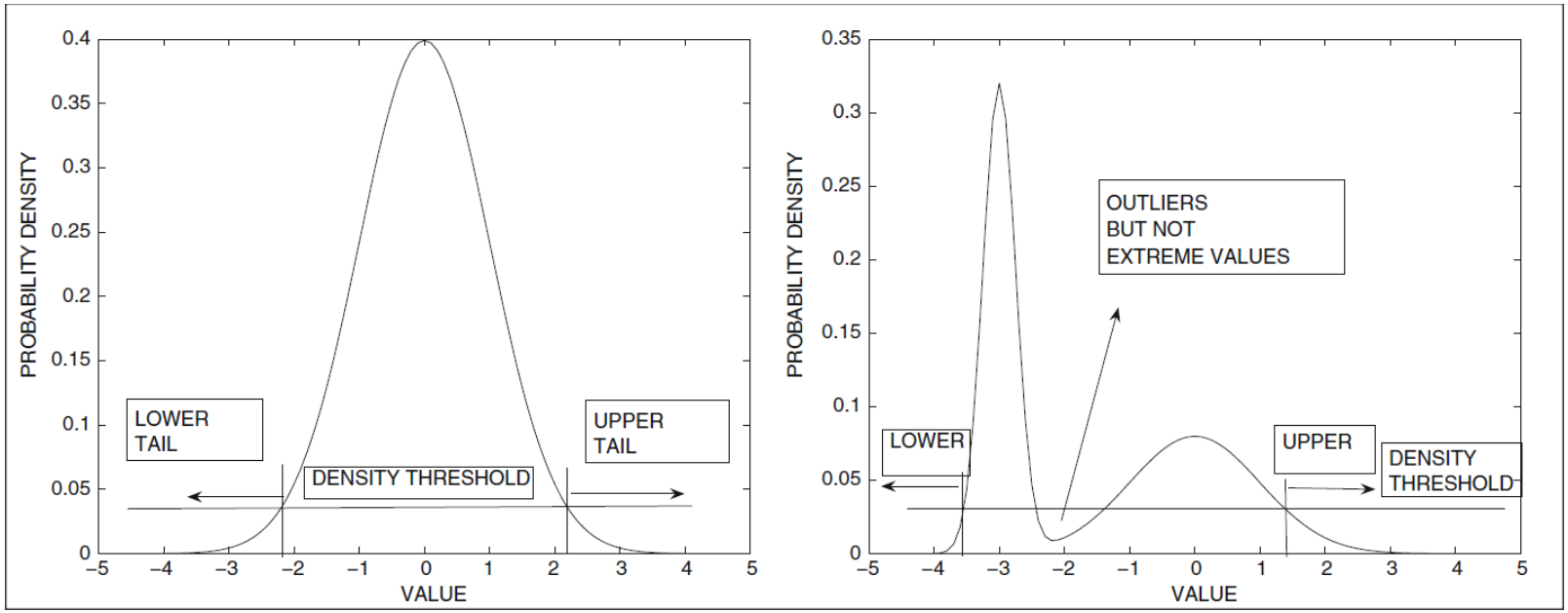
Univariate Extreme Value Analysis

- Typically, statistical tail confidence tests assume that the 1-dimensional data are described by a specific distribution.
- These methods attempt to determine the fraction of the objects expected to be more extreme than the data point
- The quantification provides a level of confidence about whether or not a specific data point is an extreme value.

Outlier Analysis - Extreme Value Analysis

How is the “tail” of a distribution defined?

- upper tail / lower tail
- extreme regions of the distribution for which $f_X(x) \leq \theta$ for some defined threshold θ.



Outlier Analysis - Extreme Value Analysis

Commonly used model is the normal distribution.

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

The *Z*-number z_i of an observed value x_i can be computed as follows:

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

- Large positive values correspond to the upper tail
- Large negative values correspond to the lower tail.

Outlier Analysis - Extreme Value Analysis

Multivariate Extreme Values

- The implicit modeling assumption: probability distribution with a single peak (i.e., single Gaussian cluster)

- Let $\bar{\mu}$ be the *d*-dimensional mean vector of a *d*-dimensional data set, and Σ be its *d* × *d* covariance matrix.

- The probability distribution $f(\bar{X})$ for a *d*-dimensional data point *X* can be defined as follows:

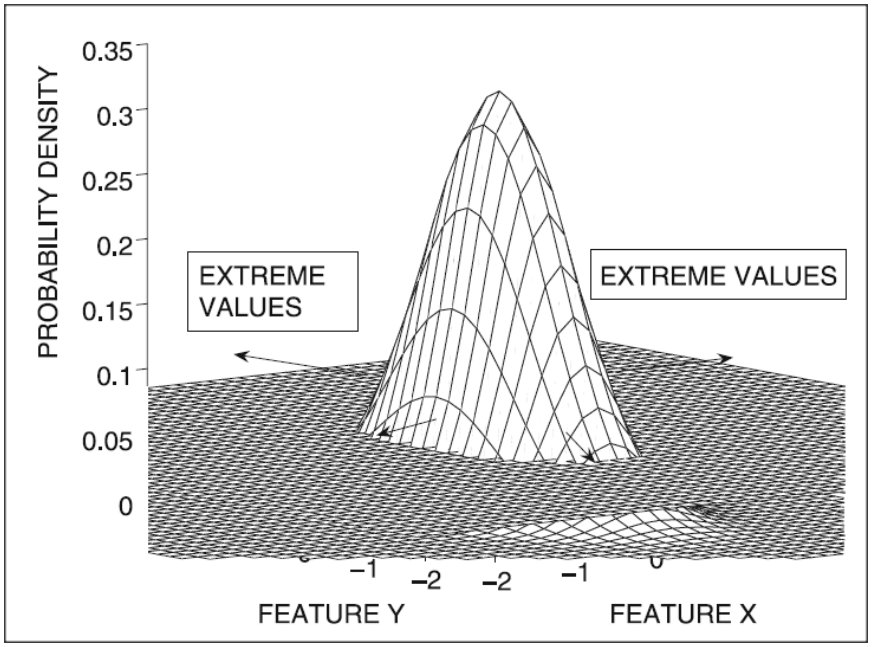
$$f(\bar{X}) = \frac{1}{\sqrt{|\Sigma|}(2\pi)^{(d/2)}} e^{-\frac{1}{2}(\bar{X}-\bar{\mu})\Sigma^{-1}(\bar{X}-\bar{\mu})^T}$$

Outlier Analysis - Extreme Value Analysis

Multivariate Extreme Values

- if *Maha*(*X, μ,* Σ) represents the Mahalanobis distance between *X* and *μ*, with respect to the covariance matrix Σ, then

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



Outlier Analysis - Depth-Based Methods

- Depth-based methods are based on the general principle that the convex hull of a set of data points represents the pareto-optimal extremes of this set.

- A depth-based algorithm proceeds in an iteratively

  - during the k^{th} iteration, all points at the corners of the convex hull of the data set are removed.

- The index of the iteration k provides an outlier score.

Outlier Analysis - Depth-Based Methods

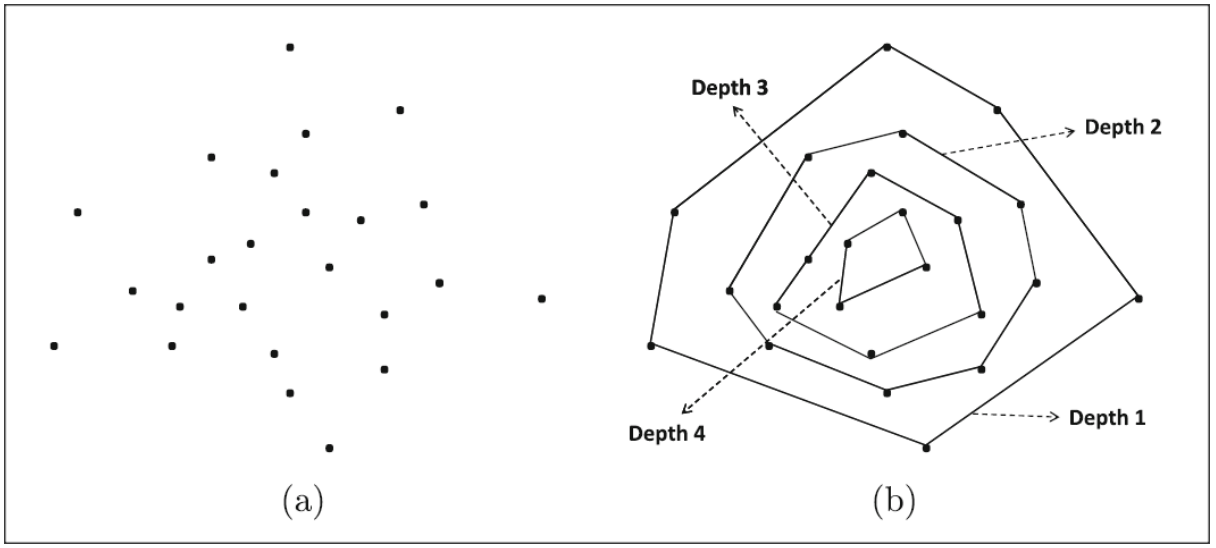
Algorithm *FindDepthOutliers* (Data Set: *D*, Score Threshold: *r*)

begin

    k = 1;
    repeat
        Find set *S* of corners of convex hull of *D*;
        Assign depth *k* to points in *S*;
        D = *D*−*S*;
        k = *k* + 1;
    until (*D* is empty);
    Report points with depth at most *r* as outliers;

end

Outlier Analysis - Depth-Based Methods



- Depth-based methods similar goals as the multivariate method
- Not dependent on parametrization of distributions
- All data points at the corners of a convex hull are treated equally.

Outlier Analysis - Probabilistic Models

- Probabilistic models are based on a generalization of the multivariate extreme values analysis methods.
- By generalizing this model to multiple mixture components, it is possible to determine general outliers, rather than multivariate extreme values.
- This idea is related to the EM-clustering algorithm discussed in previous lectures.
- At an intuitive level, data points that do not naturally fit any cluster in the probabilistic sense may be reported as outliers.

Outlier Analysis - Probabilistic Models

- The broad principle of a mixture-based generative model is to assume that the data were generated from a mixture of *k* distributions with the probability distributions $G_1 \cdots G_k$.

- This generative model will be denoted by *M*

  - it generates each point in the data set *D*.

- Data set *D* is used to estimate the parameters of the model.

- After the parameters of the model have been estimated, outliers are defined as those data points in *D* that are highly unlikely to be generated by this model.

Outlier Analysis - Probabilistic Models

- For data set *D* containing *n* data points, denoted by $X_1 \cdots X_n$, the probability density of the data set being generated by model *M* is the product of the various point-specific probability densities:

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

- The log-likelihood fit *L*(*D*|*M*) :

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

- This log-likelihood fit needs to be optimized to determine the model parameters.

Outlier Analysis - Probabilistic Models

- After the parameters of the model have been determined, the value of $f^{point}(X_j|M)$ (or its logarithm) may be reported as the outlier score.

- The major advantage of such mixture models is that the mixture components can also incorporate domain knowledge about the shape of each individual mixture component.

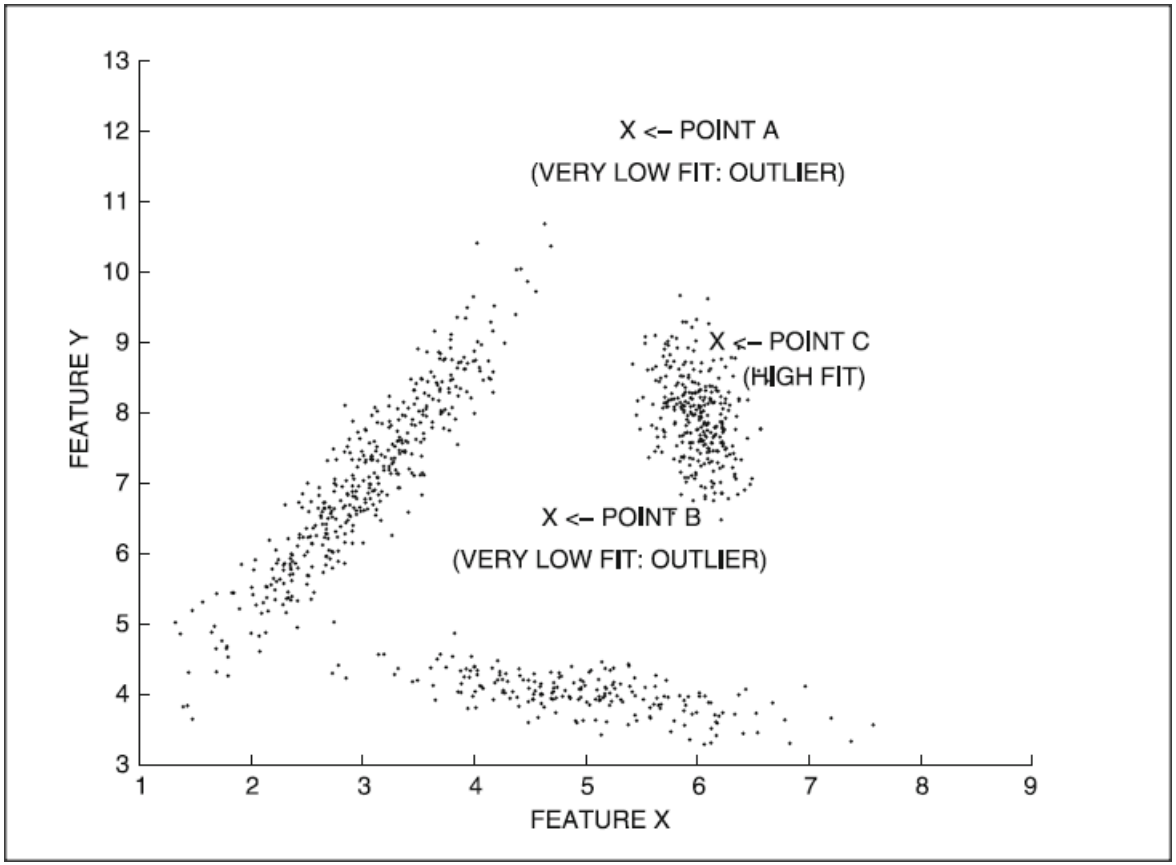
- For example:

  - if it is known that the data points in a particular cluster are correlated in a certain way,

  - then it can be incorporated in the mixture model by fixing the appropriate parameters of the covariance matrix

Outlier Analysis - Probabilistic Models

Likelihood fit values versus outlier score



Outlier Analysis - Clustering for Outlier Detection

- Clustering and outlier detection have complementary relationship.
- A simplistic view: every data point is either a member of a cluster or an outlier.

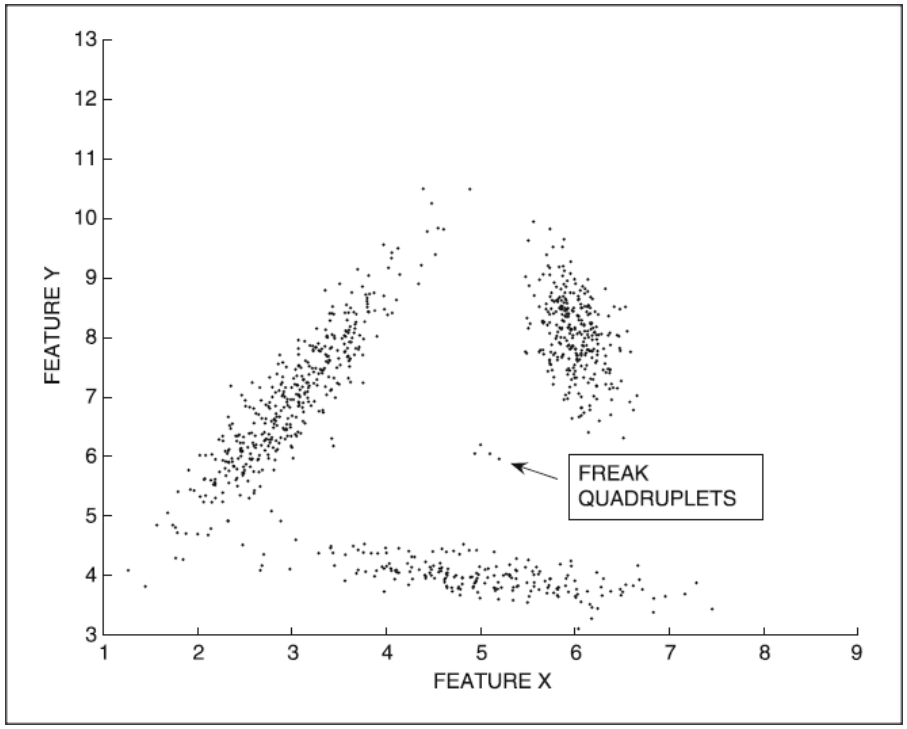
    - clustering algorithms are not optimized for outlier detection.

- Clustering models have some advantages.

    - Outliers often tend to occur in small clusters of their own.

Outlier Analysis - Clustering for Outlier Detection

- An example of a small set of isolated outliers is illustrated in figure.



- A simple way of defining the outlier score of a data point is to use the raw distance of the data point to its closest cluster centroid.

Outlier Analysis - Distance-Based Outlier Detection

- Because outliers are defined as data points that are far away from the "crowded regions" (or clusters) in the data, a natural and *instance-specific* way of defining an outlier is as follows:

The distance-based outlier score of an object O is its distance to its k^{th} nearest neighbor.

- *k*-nearest neighbor distance is the most common.
- Variation: the average distance to the *k*-nearest neighbors.
- Selecting a value of *k* larger than 1 helps identify isolated groups of outliers.

Outlier Analysis - Distance-Based Outlier Detection

- For example, in the previous figure, as long as *k* is fixed to any value larger than 3, all data points within the small groups of closely related points will have a high outlier score.

- Distance-based methods typically use a finer granularity of analysis than clustering methods

    - Distinguish between ambient noise and truly isolated anomalies.

    - Noise will typically have a lower *k*-nearest neighbor distance than a truly isolated anomaly.

Outlier Analysis - Distance-Based Outlier Detection

- The price of this better granularity is higher computational complexity. Consider a data set *D* containing *n* data points.

- The determination of the *k*-nearest neighbor distance requires $O(n)$ time *for each data point*, when a sequential scan is used.

- The determination of the outlier scores of all data points may require $O(n^2)$ time. This is clearly not a feasible option for very large data sets.

- Variety of methods are used to speed up the computation:

1. *Index structures*

2. *Pruning tricks*

Outlier Analysis - Density-Based Methods

- Density-based methods based on similar principles as density-based clustering. The idea is to determine sparse regions in the underlying data in order to report outliers.

- Histogram-based, grid-based, or kernel density-based methods can be used.

- Histogram- and Grid-Based Techniques

  - Histograms are simple and easy to construct for univariate data

  - In the context of multivariate data, a natural generalization is the use of a grid-structure.

  - Data points that have density less than τ in any particular grid region are reported as outliers.

Outlier Analysis - Density-Based Methods

- Kernel density estimation methods are similar to histogram techniques.
- The value of the density at a given point is estimated as the sum of the smoothed values of kernel functions $K_h(\cdot)$ associated with each point in the data set.
- The kernel estimation *f*(*X*) based on *n* data points of dimensionality *d*, and kernel function $K_h(\cdot)$ is defined as follows:

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

Outlier Analysis - Density-Based Methods

- An example of such a distribution is the Gaussian kernel with width *h*.

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

- The value of the density is reported as the outlier score.
  - Low values of the density indicate greater tendency to be an outlier.

Outlier Analysis - Information-Theoretic Models

- Outliers are data points that do not naturally fit the remaining data distribution. The outliers would increase the minimum code length required to describe the patterns in distribution.

- For example, consider the following two strings:

  ABABABABABABABABABABABABABABABABAB
  ABABACABABABABABABABABABABABABABAB

- The first string can be described concisely as “AB 17 times.”

- The second string has a single position corresponding to the symbol C. Therefore, the second string can no longer be described as concisely.

Outlier Analysis - Information-Theoretic Models

- In other words, the presence of the symbol C somewhere in the string increases its *minimum description length.*

- Information-theoretic models are based on this general principle because they measure the increase in model size required to describe the data as concisely as possible.

- Information-theoretic models almost equivalent to conventional deviation-based models

    - the outlier score is defined by the model size for a fixed deviation

Outlier Analysis - Information-Theoretic Models

- For example, if a clustering model is used, then a larger number of cluster centroids (model size) will result in lowering the maximum deviation of any data point (including the outlier) from its nearest centroid.

- A different way of computing the outlier score

- fix the maximum allowed deviation (instead of the number of cluster centroids)

- compute the number of cluster centroids required to achieve the same level of deviation, with and without a particular data point.

Detecting Fraudulent Transactions

- The case study addresses an instantiation of the general problem of detecting unusual observations of a phenomena, that is, finding rare and quite different observations.
- The driving application has to do with transactions of a set of products that are reported by the salespeople of some company.
- The goal is to find unusual transaction reports that may indicate fraud attempts by some of the salespeople.

Detecting Fraudulent Transactions

- We want to provide a kind of fraud probability ranking as outcome of the process. These rankings should allow the company to apply its inspection resources in an optimal way.

- This general resource-bounded inspection activity is frequent in many fields, such as credit card transactions, tax declarations inspection, etc.

- This section addresses several data mining tasks:

- (1) outlier or anomaly detection,
- (2) clustering
- (3) semi-supervised prediction models.

Problem Description and Objectives

- These salespeople sell a set of products of the company and report these sales with a certain periodicity. The data we have available concerns these reports over a short period of time.

- The salespeople are free to set the selling price according to their own policy and market. At the end of each month, they report back to the company their transactions.

- The goal of this data mining application is to help in the task of verifying the veracity of these reports given past experience of the company that has detected both errors and fraud attempts in these transaction reports.

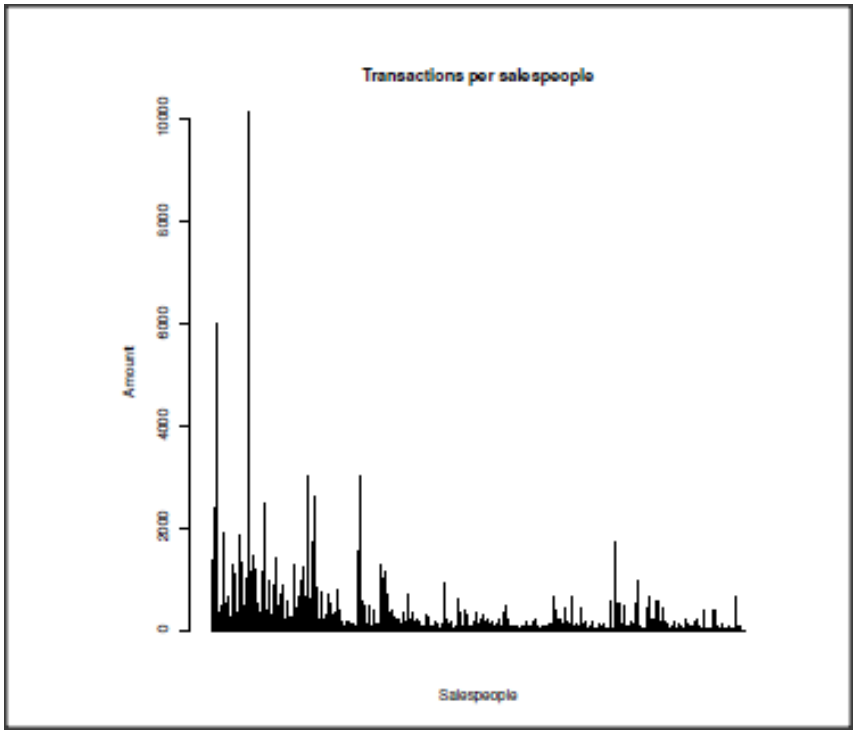
Loading the Data into R

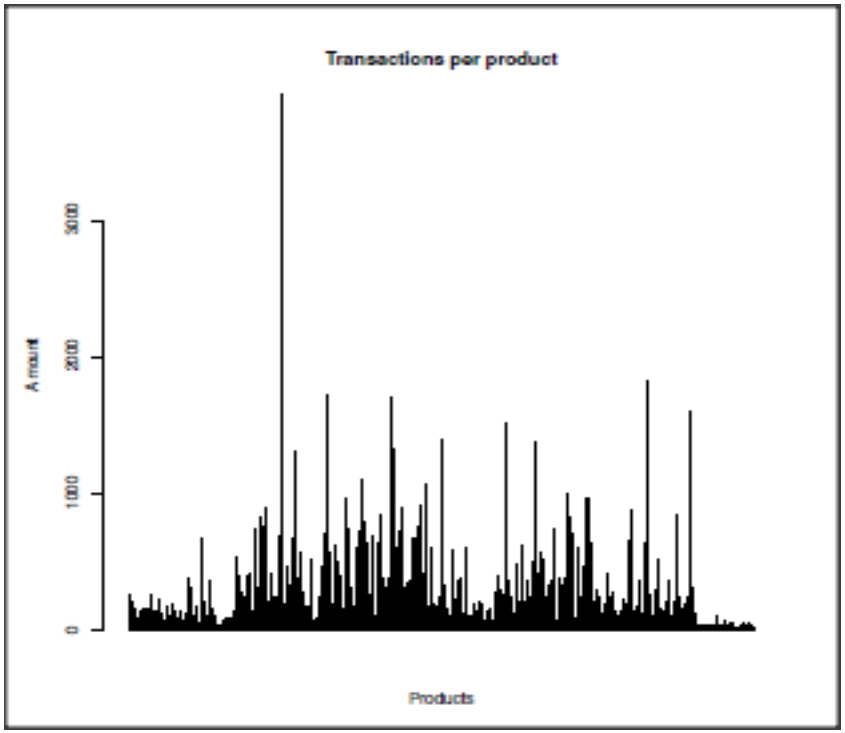
```
load("sales.Rdata")

> head(sales)
ID Prod Quant Val Insp
1 v1 p1 182 1665 unkn
2 v2 p1 3072 8780 unkn
3 v3 p1 20393 76990 unkn
4 v4 p1 112 1100 unkn
5 v3 p1 6164 20260 unkn
6 v5 p2 104 1155 unkn
```

Exploring the Dataset

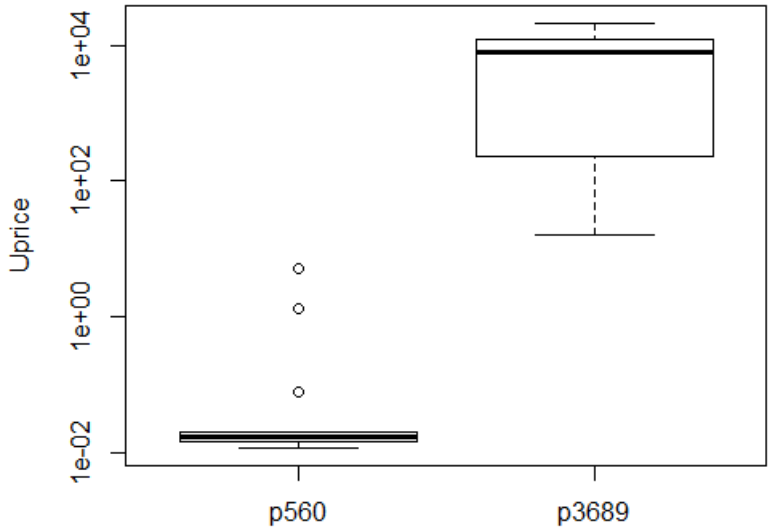
See R code





Exploring the Dataset

- The distribution of the unit prices of the cheapest and most expensive products.



Exploring the Dataset

- We can carry out a similar analysis to discover which salespeople are the ones who bring more (less) money to the company

	Most	Least
[1,]	"v431"	"v3355"
[2,]	"v54"	"v6069"
[3,]	"v19"	"v5876"
[4,]	"v4520"	"v6058"
[5,]	"v955"	"v4515"

- It may be interesting to note that the top 100 salespeople on this list account for almost 40% of the income of the company, while the bottom 2,000 out of the 6,016 salespeople generate less than 2% of the income.

Exploring the Dataset

If we carry out a similar analysis in terms of the quantity that is sold for each product, the results are even more unbalanced:

```
     Most    Least
[1,] "p2516" "p2442"
[2,] "p3599" "p2443"
[3,] "p314"  "p1653"
[4,] "p569"  "p4101"
[5,] "p319"  "p3678"
```

Exploring the Dataset

- One of the main assumptions we will be making in our analysis to find abnormal transaction reports is that the unit price of any product should follow a near-normal distribution.

- This means that we expect that the transactions of the same product will have roughly the same unit price with some small variability, possibly caused by some strategies of the salespeople to achieve their commercial goals.

STEVENS
INSTITUTE of TECHNOLOGY
School of Business

Thank You

Dragos Bozdog

For academic use only.