Data Mining
Anomaly Detection

Master Soft Computing y Sistemas Inteligentes

Curso: Modelos avanzados en Minería de Datos

Universidad de Granada

Juan Carlos Cubero

JC.Cubero@decsai.ugr.es

Transparencias realizadas a partir de las confeccionadas por Tan, Steinbach, Kumar, para el libro:

Introduction to Data Mining

Data Mining
Anomaly Detection

- Motivation and Introduction
- Graphical and Statistical Approaches
- Distance-based Approaches
  - Nearest Neighbor
  - Density-based
  - Cluster-based
- Post processing: Outlier detection rate
- Abnormal regularities

Importance of Anomaly Detection

Bacon, writing in Novum Organum about 400 years ago said:

"Errors of Nature, Sports and Monsters correct the understanding in regard to ordinary things, and reveal general forms. For whoever knows the ways of Nature will more easily notice her deviations; and, on the other hand, whoever knows her deviations will more accurately describe her ways."

Anomaly/Outlier Detection

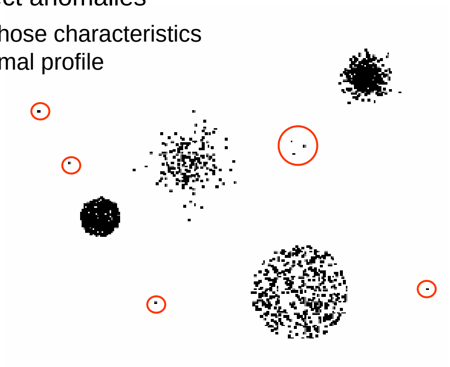
- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Variants of Anomaly/Outlier Detection Problems
  - Given a database D, find all the data points **x** ∈ D with anomaly scores greater than some threshold t
  - Given a database D, find all the data points **x** ∈ D having the top-n largest anomaly scores f(**x**)
  - Given a database D, containing mostly normal (but unlabeled) data points, and a test point **x**, compute the anomaly score of **x** with respect to D
- Applications:
  - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

Anomaly Detection

- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data
- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack

Anomaly Detection Schemes

- General Steps
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile
- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based

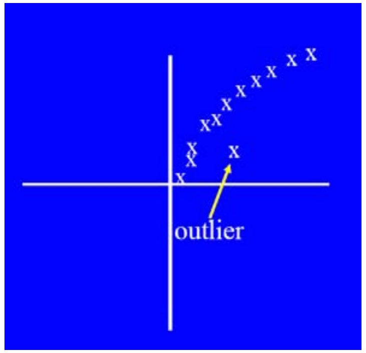


Data Mining Anomaly Detection

- Motivation and Introduction
- Graphical and Statistical Approaches
- Distance-based Approaches
  - Nearest Neighbor
  - Density-based
  - Cluster-based
- Post processing: Outlier detection rate
- Abnormal regularities

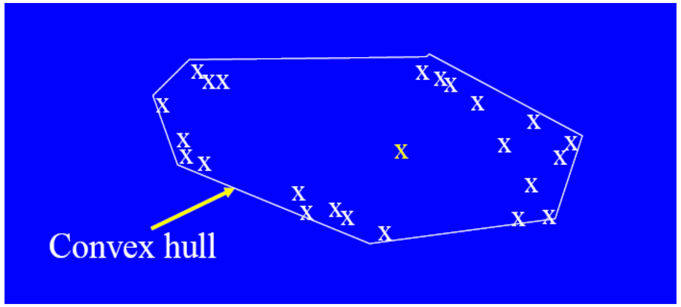
Graphical Approaches

- Limitations
  - Time consuming
  - Subjective



Convex Hull Method

- Extreme points are assumed to be outliers
- Use convex hull method to detect extreme values



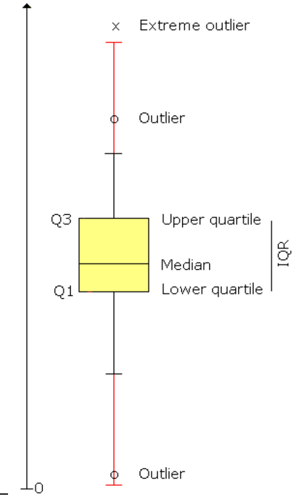
- What if the outlier occurs in the middle of the data?

Statistical Approaches

- Without assuming a parametric model describing the distribution of the data (and only 1 variable)

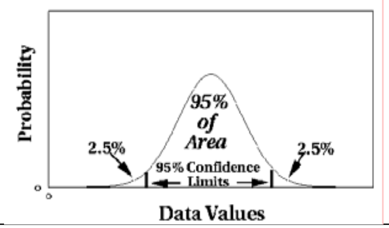
IQR = Q3 - Q1

P is an Outlier if P > Q3 + 1.5 IQR
P is an Outlier if P < Q1 - 1.5 IQR
P is an Extreme Outlier if P > Q3 + 3 IQR
P is an Extreme Outlier if P < Q1 - 3 IQR



Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Grubbs' Test

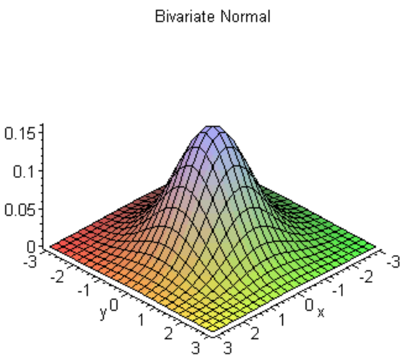
- Detect outliers in univariate data
- Assume data comes from normal distribution
- Detects one outlier at a time, remove the outlier, and repeat
  - H_0: There is no outlier in data
  - H_A: There is at least one outlier
- Grubbs' test statistic: $$G = \frac{\max \left| X - \overline{X} \right|}{s}$$
- Reject H_0 if:

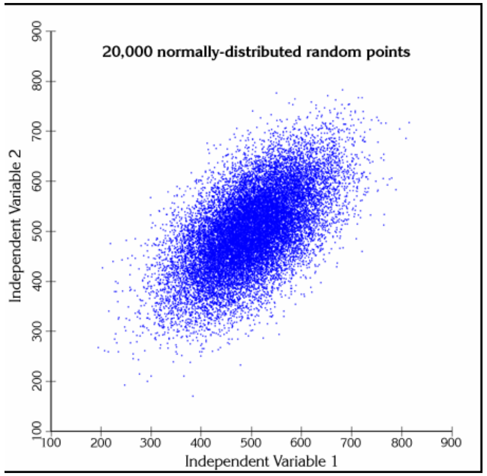
$$G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$$

http://www.graphpad.com/quickcalcs/Grubbs1.cfm

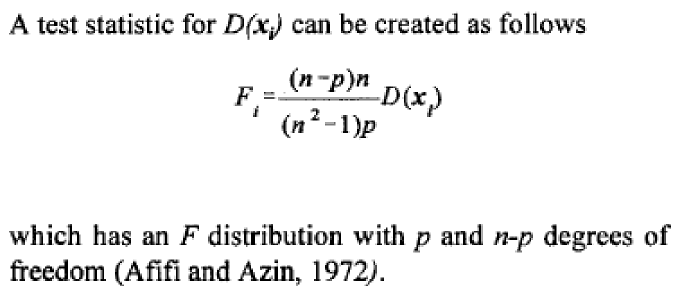
Multivariate Normal Distribution

- Working with several dimensions





Multivariate Normal Distribution



A test statistic for $D(x_i)$ can be created as follows

$$F_i = \frac{(n-p)n}{(n^2-1)p} D(x_i)$$

which has an F distribution with p and n-p degrees of freedom (Afifi and Azin, 1972).

Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

Data Mining Anomaly Detection

- Motivation and Introduction
- Graphical and Statistical Approaches
- Distance-based Approaches
  - Nearest Neighbor
  - Density-based
  - Cluster-based
- Post processing: Outlier detection rate
- Abnormal regularities

Distance-based Approaches (DB)

- Data is represented as a vector of features. We have a distance measure to evaluate nearness between two points

- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

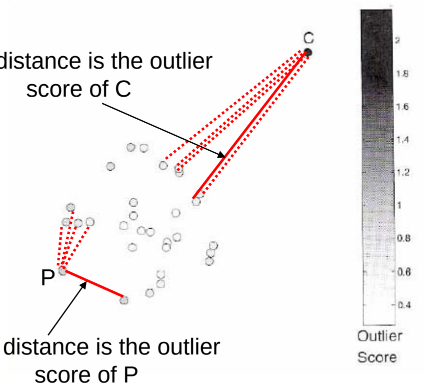
- The first two methods work directly with the data. The third one requires an a priori clustering

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance (proximity) between every pair of data points
  - Fix a magic number k representing the k-th nearest point to another point
  - For a given point P, compute its *outlier score* as the distance of P to its k-nearest neighbor. There are no clusters. Neighbor refers to a point
  - Consider as outliers those points with *high* outlier score.

Nearest-Neighbor Based Approach

k = 5

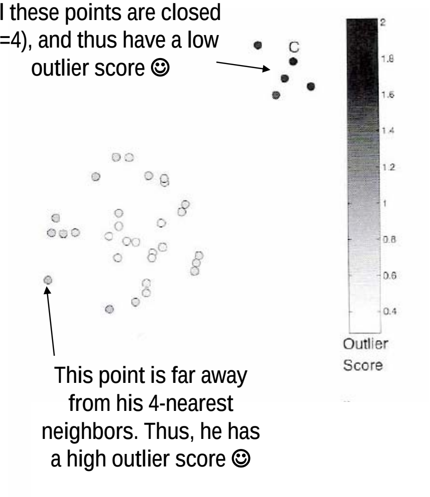


This distance is the outlier score of C

This distance is the outlier score of P

Nearest-Neighbor Based Approach

k = 4



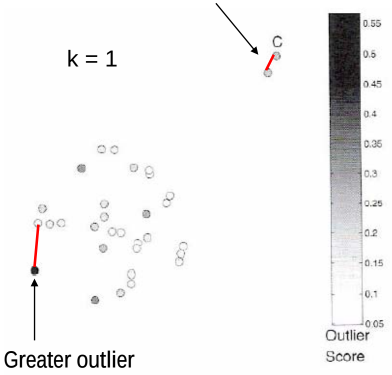
All these points are closed (k=4), and thus have a low outlier score ☺

This point is far away from his 4-nearest neighbors. Thus, he has a high outlier score ☺

Nearest-Neighbor Based Approach

Choice of k is problematic

Low outlier scores ☹



Greater outlier score than C ☹

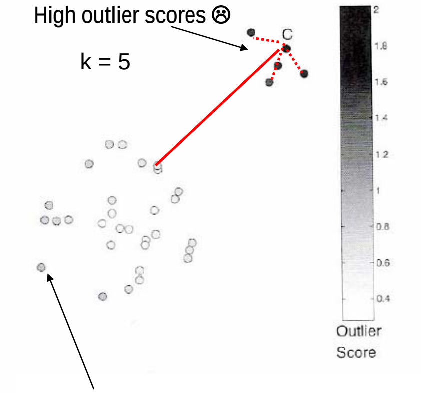
Nearest-Neighbor Based Approach

Choice of k is problematic

All the points in any isolated natural cluster with fewer points than k, have high outlier score

We could mitigate the problem by taking the average distance to the k-nearest neighbors but is still poor

High outlier scores ☹



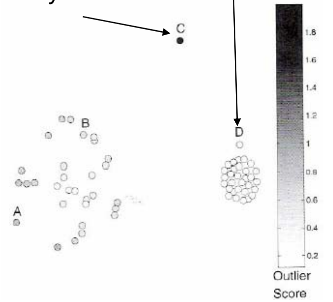
Medium-High outlier score 😐
(Could be greater)

Nearest-Neighbor Based Approach

Density should be taken into account

C has a high outlier score ☺ for every k

D has a low outlier score ☹ for every k



Density-based Approach

Density should be taken into account

Let us define the *k-density around a point as:*

Alternative a) *k-density of a point is the inverse of the average sum of the distances to its k-nearest neighbors.*

Alternative b) *d-density of a point P is the number P_i of points which are d-close to P (distance(P_i,P) ≤ d)*
Used in DBSCAN
Choice of *d* is problematic

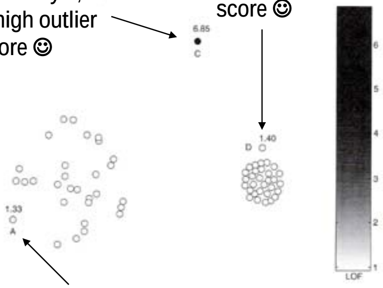
Density-based Approach

Density should be taken into account

Compute the *k-density* of every point

Define the *k-relative density* of a point P as the ratio between its *k-density* and the average *k-densities* of its *k*-nearest neigbhors

The outlier score of a point P (called LOF for this method) is its *k-relative density*



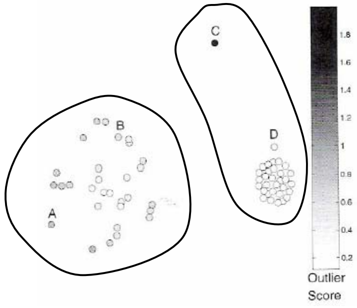
C has a low k-relative density for every k, and thus a high outlier score ☺

D has a low-medium k-relative density for every k, and thus a high outlier score ☺

A has a low k-relative density for every k, and thus a high outlier score ☺

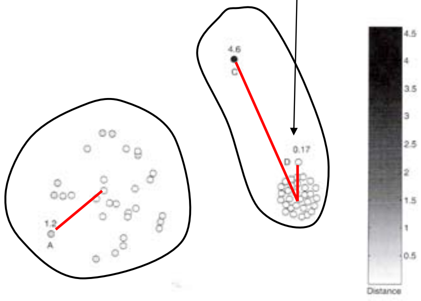
Clustering-Based Approach

- Basic idea:
  - We have already constructed a set of clusters by any clustering method.
  - An object is a *cluster-based outlier* if the object does not strongly belong to any cluster.
  - How do we measure it?



Clustering-Based Approach

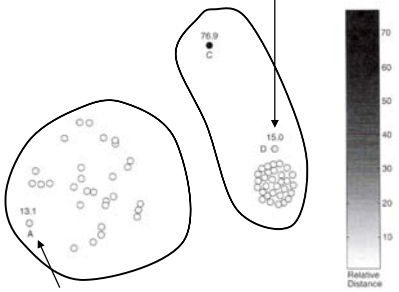
- Alternative a)
  - By measuring the distance to its closest centroid



D its near to its centroid, and thus a low outlier score ☹

Clustering-Based Approach

- Alternative b)
  - By measuring the **relative** distance to its closest centroid.
  - Relative distance is the ratio of the point′s distance from the centroid to the median distance of all the points in the cluster from the centroid.



D has a medium-high relative distance to its centroid, and thus a high outlier score ☺

A has a medium-high relative distance to its centroid, and thus a high outlier score ☺

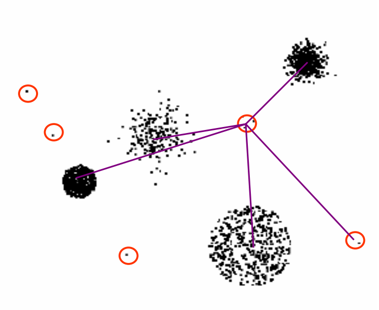
Clustering-Based

Choice of k is problematic

(k is now the number of clusters)

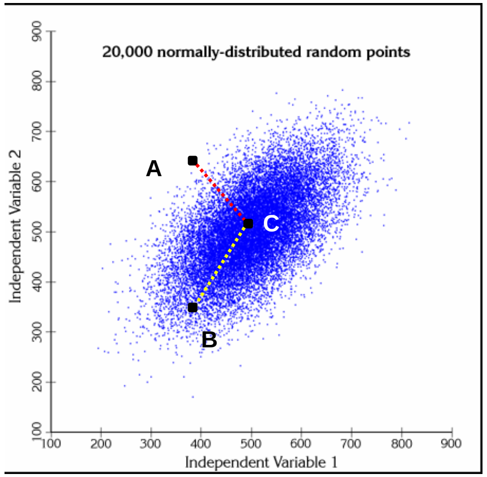
Usually, it’s better to work with a large number of small clusters.

An object identified as outlier when there are a large number of small clusters, it’s likely to be a true outlier.



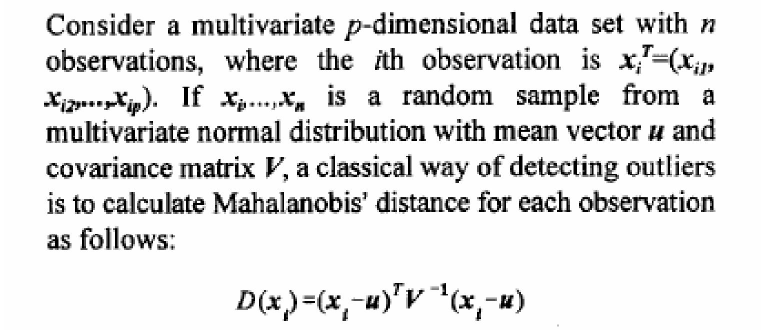
Distance Measure

B is *closest* to the centroid C than A, but its Euclidean distance is higher



Distance Measure

- Replace Euclidean distance by Mahalanobis distance:

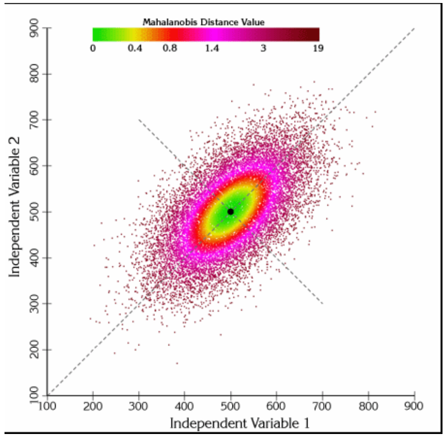


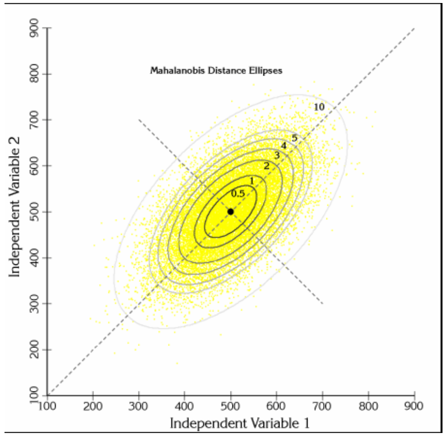
Consider a multivariate p-dimensional data set with n observations, where the ith observation is $x_i^T=(x_{i1}, x_{i2},\dots,x_{ip})$. If $x_i,\dots,x_n$ is a random sample from a multivariate normal distribution with mean vector u and covariance matrix V, a classical way of detecting outliers is to calculate Mahalanobis’ distance for each observation as follows:

$$D(x_i)=(x_i-u)^T V^{-1}(x_i-u)$$

Usually, V is unknown and is replaced by the sample Covariance matrix

Distance Measure





Outliers in High Dimensional Problems

- In high-dimensional space, data is sparse and notion of proximity becomes meaningless
  - Every point is an almost equally good outlier from the perspective of proximity-based definitions
- Lower-dimensional projection methods
  - A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density

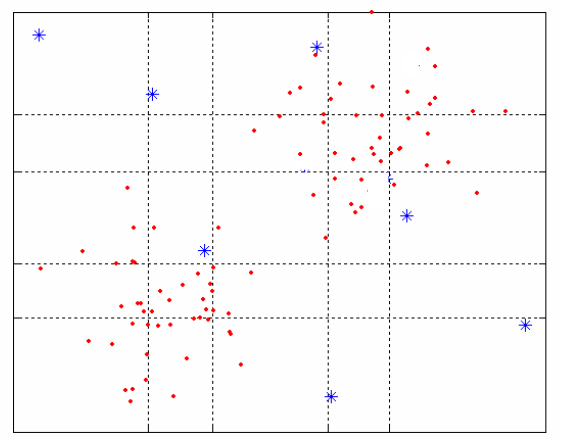
Outliers in High Dimensional Problems

- Approach by Aggarwal and Yu.
- Divide each attribute into ϕ equal-depth intervals
  - Each interval contains a fraction f = 1/ϕ of the records
- Consider a k-dimensional cube created by picking grid ranges from k different dimensions
  - If attributes are independent, we expect region to contain a fraction f^k of the records
  - If there are N points, we can measure sparsity of a cube D as:

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

Outliers in High Dimensional Problems

- k=2, N=100, ϕ = 5, f = 1/5 = 0.2, N × f^2 = 4



Outliers in High Dimensional Problems

- Algorithm:
  - Try every k-projection (k=1,2,...Dim)
  - Compute the sparsity of every Cube in such k – projection
  - Retain the cubes with the most negative sparsity
- The authors use a genetic algorithm to compute it
- This is still an open problem for future research

Data Mining
Anomaly Detection

- Motivation and Introduction
- Graphical and Statistical Approaches
- Distance-based Approaches
  - Nearest Neighbor
  - Density-based
  - Cluster-based
- Post processing: Outlier detection rate
- Abnormal regularities

Base Rate Fallacy (Axelsson, 1999)

Suppose that your physician performs a test that is 99% accurate, i.e. when the test was administered to a test population all of which had the disease, 99% of the tests indicated disease, and likewise, when the test population was known to be 100% free of the disease, 99% of the test results were negative.

Upon visiting your physician to learn of the results he tells you he has good news and bad news. The bad news is that indeed you tested positive for the disease. The good news however, is that out of the entire population the rate of incidence is only 1=10000, i.e. only 1 in 10000 people have this ailment.

What, given the above information, is the probability of you having the disease?

Base Rate Fallacy

- Bayes theorem:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

- More generally:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{\sum_{i=1}^{n} P(A_i) \cdot P(B|A_i)}$$

Base Rate Fallacy

- Call *S=Sick, Pt=Positive*
  P(S)=1/10000 P(Pt|S)=0.99 P(Pt|¬S)=1- P(¬Pt|¬S)
- Compute *P(S|P)*

$$P(S|P) = \frac{P(S) \cdot P(P|S)}{P(S) \cdot P(P|S) + P(\neg S) \cdot P(P|\neg S)}$$

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} =$$

$$= 0.00980\ldots \approx 1\%$$

- Even though the test is 99% certain, your chance of having the disease is 1/100, because the population of healthy people is much larger than sick people

Base Rate Fallacy in Outlier Detection

- Outlier detection as a Classification System:
  Two classes: Outlier, Not an outlier
- A typical problem: Intrusion Detection
  I : real intrusive behavior,
  ¬I : non-intrusive behavior
  A : alarm (outlier detected)
  ¬A : no alarm
- A good classification system will have:
  - A high Detection rate (true positive rate): *P(A|I)*
  - A low False alarm rate: *P(A|¬I)*
- We should also obtain high values of:
  – Bayesian detection rate, *P(I|A)* (If the alarm fires, its an intrusion)
  – *P(¬I| ¬A)* (if the alarm does not fire, it is not an intrusion)

Base Rate Fallacy in Outlier Detection

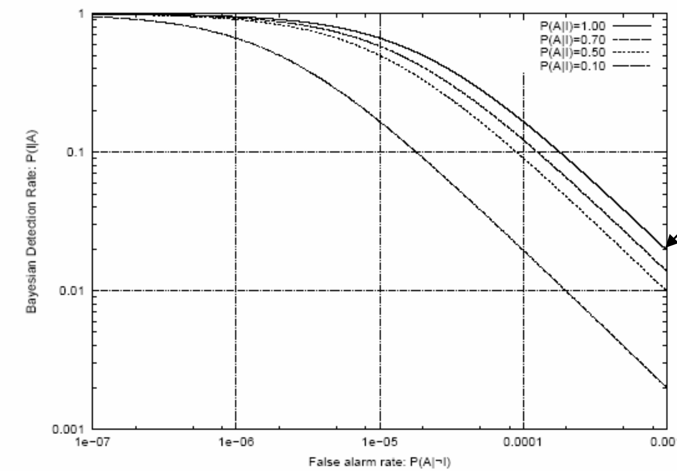
$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)}$$

- In intrusion (outlier in general) detection systems, its usually to have very low *P(I)* values (10^{-5}). So, *P(¬I)* is very high

$$P(I|A) = \frac{2 \cdot 10^{-5} \cdot P(A|I)}{2 \cdot 10^{-5} \cdot P(A|I) + 0.99998 \cdot P(A|\neg I)}$$

- The final value of *P(I|A)* is dominated by the false alarm rate *P(A|¬I).*
  P(A|¬I) should have a very low value (as 10^{-5}) to compensate 0.99998.
  BUT even a very good classification system, does not have such a false alarm rate. ☹

Base Rate Fallacy in Outlier Detection



Consider a classification with the best possible accuracy P(A|I)=1and an extremely good false alarm rate of 0.001

In this case, P(I|A)=0.02 (the scale is logarithmic) So, If the alarm fires 50 times, ony one is a real intrusion

- Conclusion: Outlier Classification systems must be carefully designed when applied to data with a very low positive rate (outlier).

Data Mining
Anomaly Detection

▪ Motivation and Introduction
▪ Graphical and Statistical Approaches
▪ Distance-based Approaches
  - Nearest Neighbor
  - Density-based
  - Cluster-based
▪ Post processing: Outlier detection rate
▪ Abnormal regularities

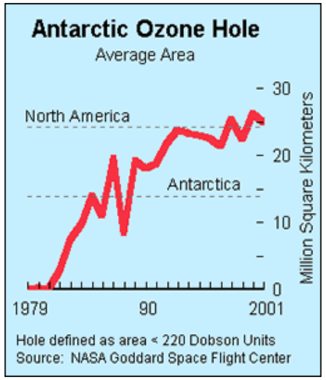
Abnormal Regularities

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- It could be better to talk about:
  - Outlier: A point is an outlier if it's considerably different than the remainder of the data
  - Abnormal regularity: A small set of closed points which are considerably different than the remainder of the data

Abnormal Regularities

Ozone Depletion History

- In 1985 three researchers (Farman, Gardinar and Shanklin) were puzzled by data gathered by the British Antarctic Survey showing that ozone levels for Antarctica had dropped 10% below normal levels
- Why did the Nimbus 7 satellite, which had instruments aboard for recording ozone levels, not record similarly low ozone concentrations?
- The ozone concentrations recorded by the satellite were so low they were being treated as outliers by a computer program and discarded!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Abnormal Regularities

- Some definitions of abnormal regularities:
  - *Peculiarities:* Association rules between infrequent items (Zhong et al)
  - *Exceptions:* Occur when a value interacts with another one, in such a way that changes the behavior of an association rule (Suzuki et al)
  - *Anomalous Association Rules:* Occur when there are two behaviors: the typical one, and the abnormal one.