

Data Mining Anomaly Detection

Lecture Notes for Chapter 10

Introduction to Data Mining

by

Tan, Steinbach, Kumar

Edited by J. Taylor for STATS202, Stanford University, Winter 2009

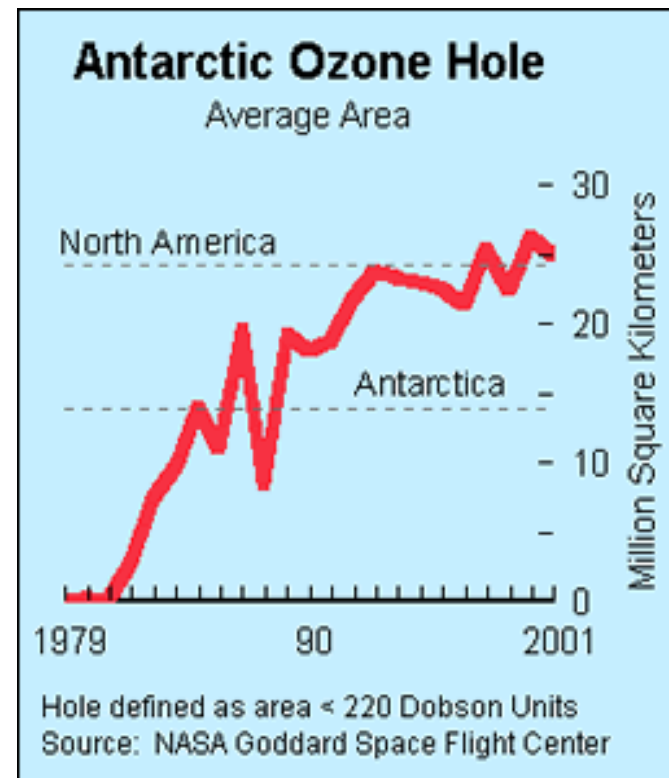
Anomaly/Outlier Detection

- 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 $\mathbf{x} \in D$ with anomaly scores greater than some threshold t
 - Given a database D , find all the data points $\mathbf{x} \in D$ having the top- n largest anomaly scores $f(\mathbf{x})$
 - Given a database D , containing mostly normal (but unlabeled) data points, and a test point \mathbf{x} , compute the anomaly score of \mathbf{x} with respect to D
- Applications:
 - Credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

Importance of Anomaly Detection

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>

Anomaly Detection

- 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

- Working assumption:

- There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data

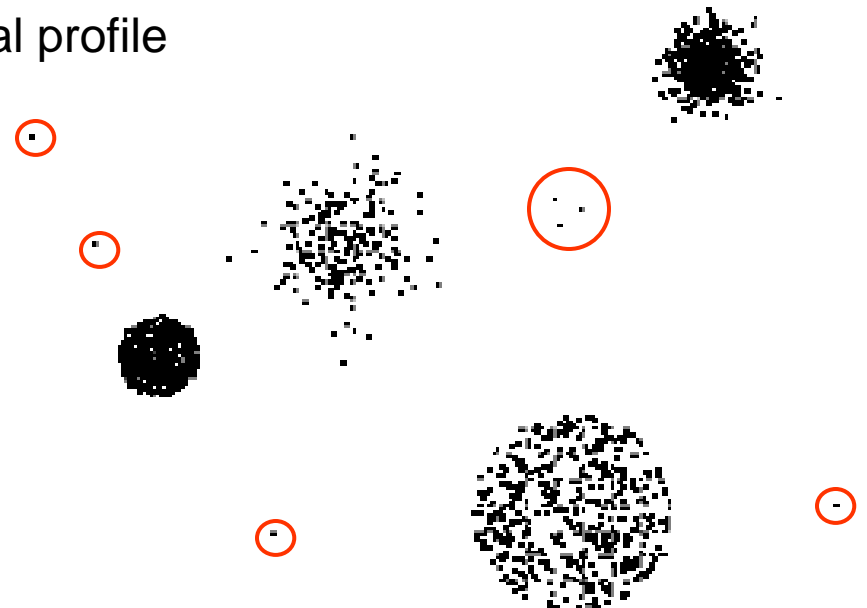
Anomaly Detection Schemes

- 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

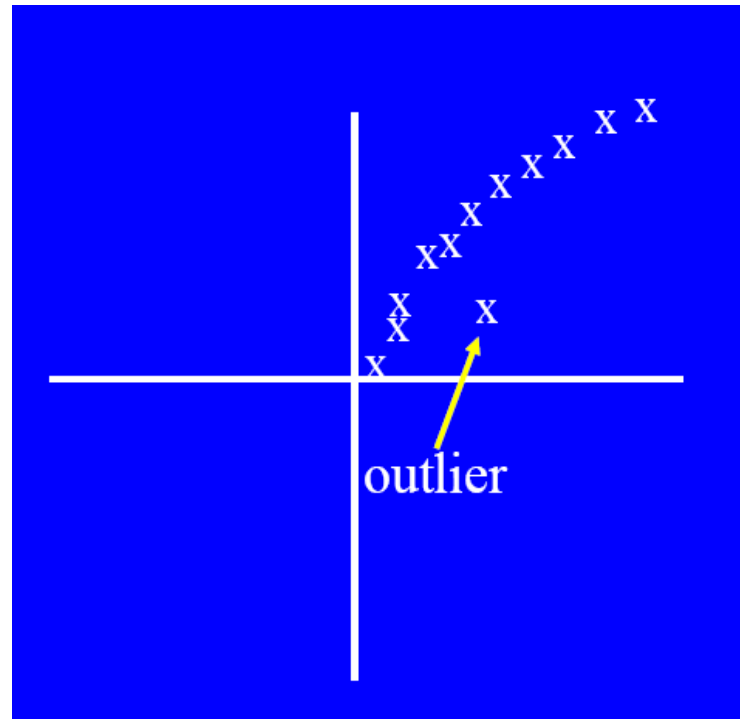
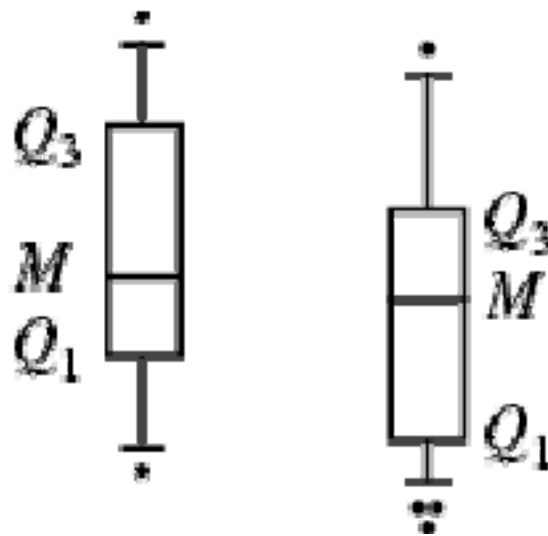


Graphical Approaches

- Boxplot (1-D), Scatter plot (2-D), Spin plot (3-D)

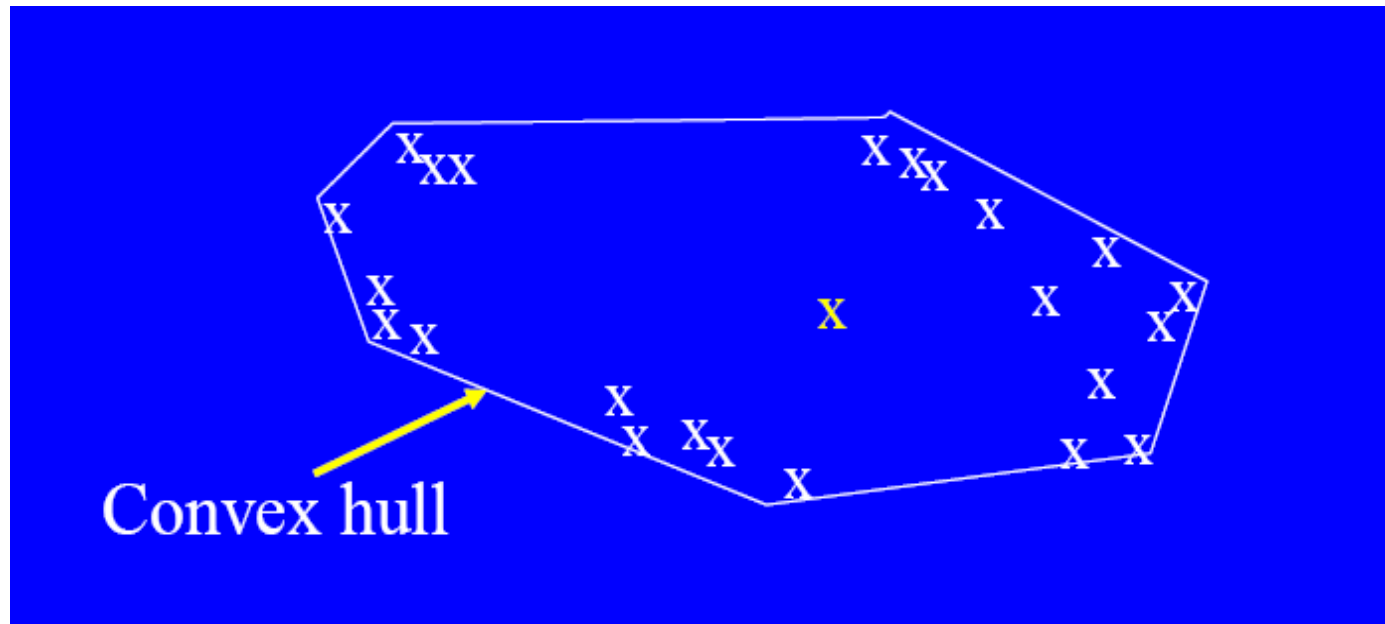
- Limitations

- Time consuming
- Subjective



Convex Hull Method

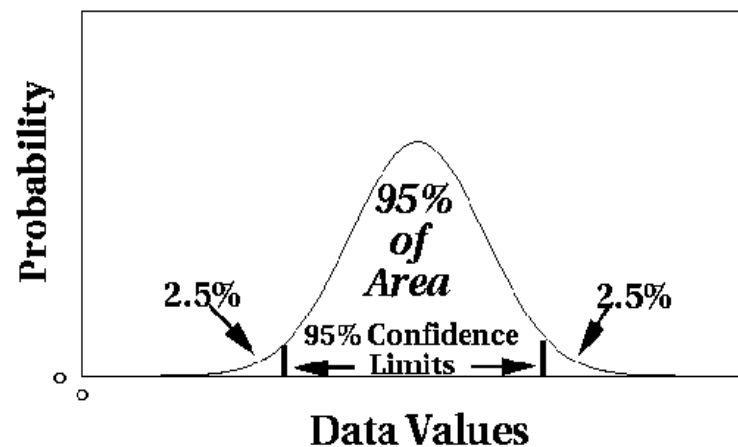
- Extreme points are assumed to be outliers
- Use convex hull method to detect extreme values



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

Statistical Approaches

- 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

- A common method of detecting outliers for a single attribute is to look for observations more than a large number of standard deviations above or below the mean
- The “z score” is the number of standard deviations above or below the mean (p. 661)
- For the normal (bell-shaped) distribution we know the exact probabilities for the z scores
- For non-normal distributions this approach is still useful and valid
- A z score of 3 or -3 is a common cut off value

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

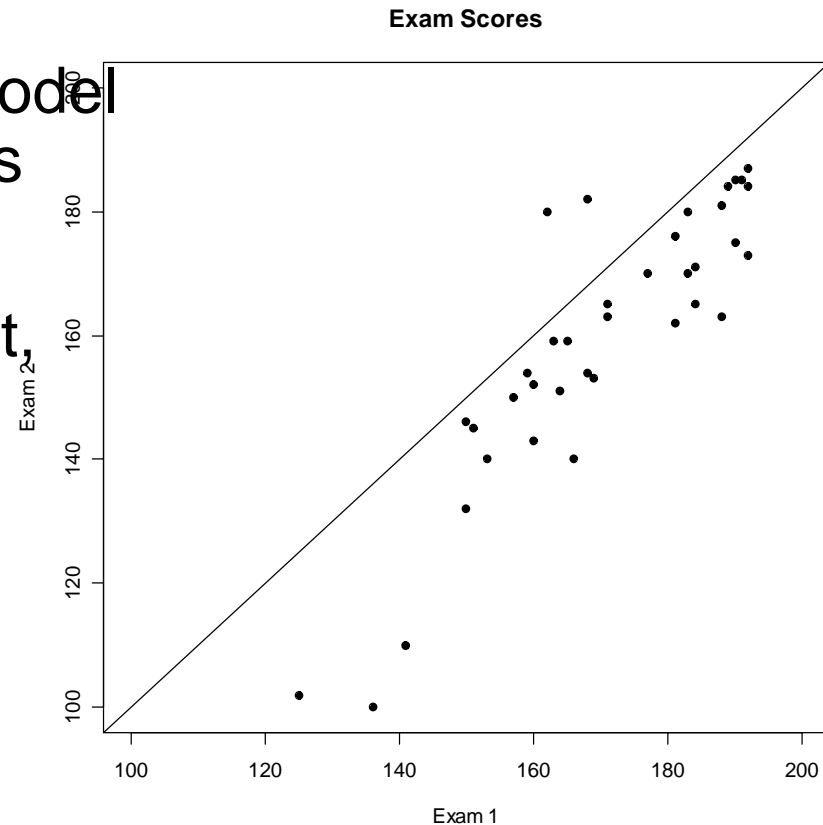
- Grubbs' test statistic:
$$G = \frac{\max |X - \bar{X}|}{s}$$

- Reject H_0 if:

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

Model based techniques

- First build a model
- Points which don't fit the model well are identified as outliers
- For the example at the right, a *least squares regression* model would be appropriate
- Residuals can be fed in to Grubbs' test.



Statistical-based – Likelihood Approach

- Assume the data set D contains samples from a mixture of two probability distributions:
 - M (majority distribution)
 - A (anomalous distribution)
- General Approach:
 - Initially, assume all the data points belong to M
 - Let $L_t(D)$ be the log likelihood of D at iteration t
 - For each point x_t that belongs to M , move it to A
 - ◆ Let $L_{t+1}(D)$ be the new log likelihood.
 - ◆ Compute the difference, $\Delta = L_t(D) - L_{t+1}(D)$
 - ◆ If $\Delta > c$ (some threshold), then x_t is declared as an anomaly and moved permanently from M to A

Statistical-based – Likelihood Approach

- Data distribution, $D = (1 - \lambda) M + \lambda A$
- M is a probability distribution estimated from data
 - Can be based on any modeling method (naïve Bayes, maximum entropy, etc)
- A is often assumed to be uniform distribution
- Likelihood at iteration t :

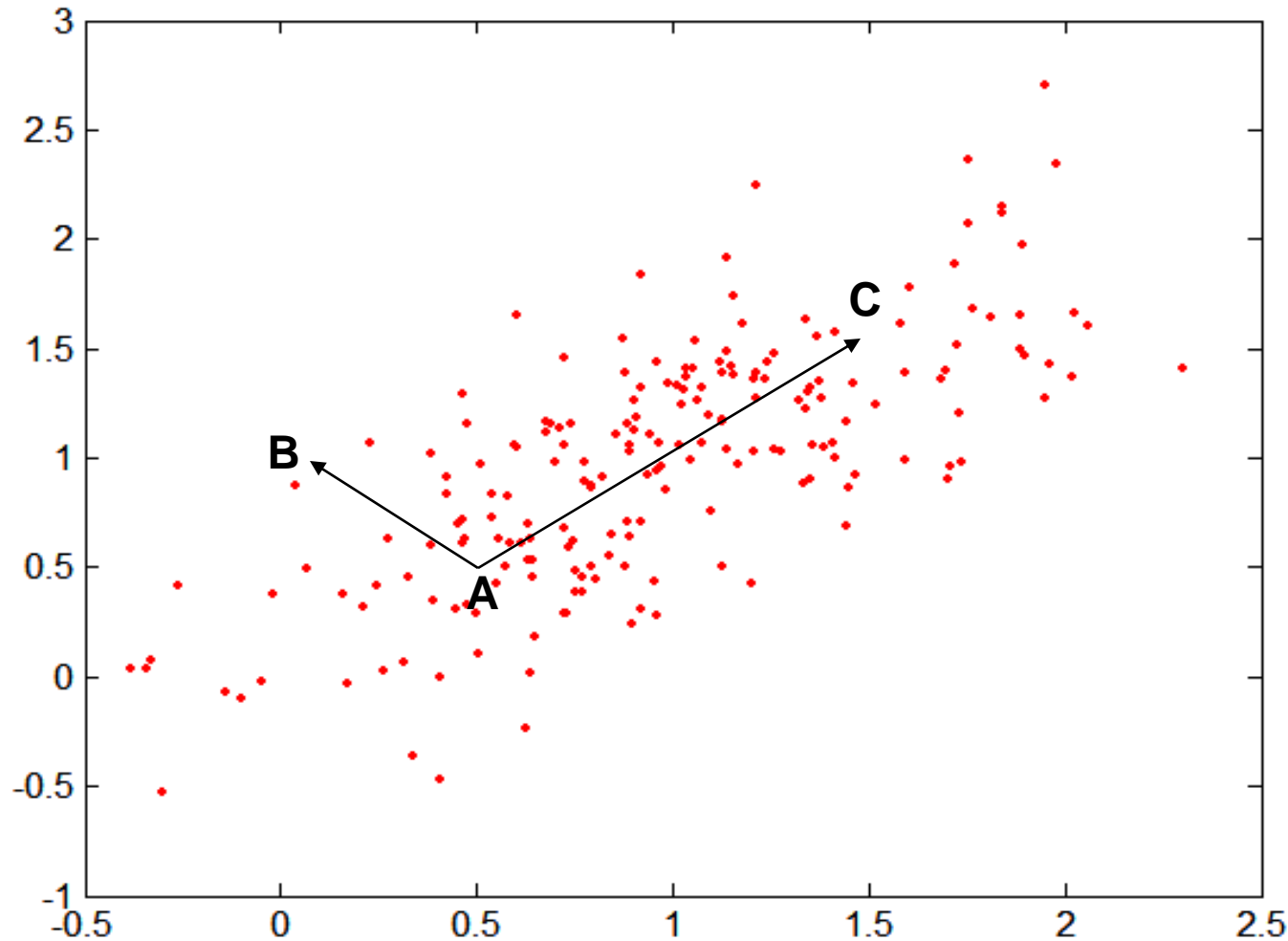
$$L_t(D) = \prod_{i=1}^N P_D(x_i) = \left((1 - \lambda)^{|M_t|} \prod_{x_i \in M_t} P_{M_t}(x_i) \right) \left(\lambda^{|A_t|} \prod_{x_i \in A_t} P_{A_t}(x_i) \right)$$

$$LL_t(D) = |M_t| \log(1 - \lambda) + \sum_{x_i \in M_t} \log P_{M_t}(x_i) + |A_t| \log \lambda + \sum_{x_i \in A_t} \log P_{A_t}(x_i)$$

Limitations of Statistical Approaches

- Choice of difference, c
- 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
- If distribution is Gaussian, statistical approach related to Mahalanobis distance of points to mean.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Distance-based Approaches

- Data is represented as a vector of features
- Three major approaches
 - Nearest-neighbor based
 - Density based
 - Clustering based

Nearest-Neighbor Based Approach

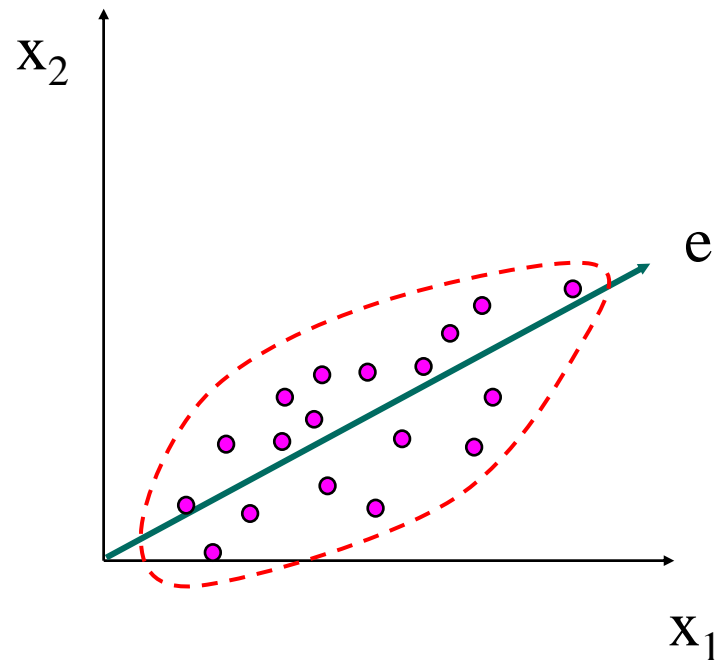
- Approach:
 - Compute the distance between every pair of data points
 - There are various ways to define outliers:
 - ◆ Data points for which there are fewer than p neighboring points within a distance D
 - ◆ The top n data points whose distance to the k th nearest neighbor is greatest
 - ◆ The top n data points whose average distance to the k nearest neighbors is greatest

Outliers in Lower Dimensional Projection

- 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
 - E.g., multidimensional scaling
 - A point is an outlier if it an outlier when considered a point in lower dimensions or it is present in a local region of abnormally low density in lower dimensions

Dimensionality Reduction: PCA

- By taking k PCA components, we construct a mapping from n -dimensions to k -dimensions.



Outliers in Lower Dimensional Projection

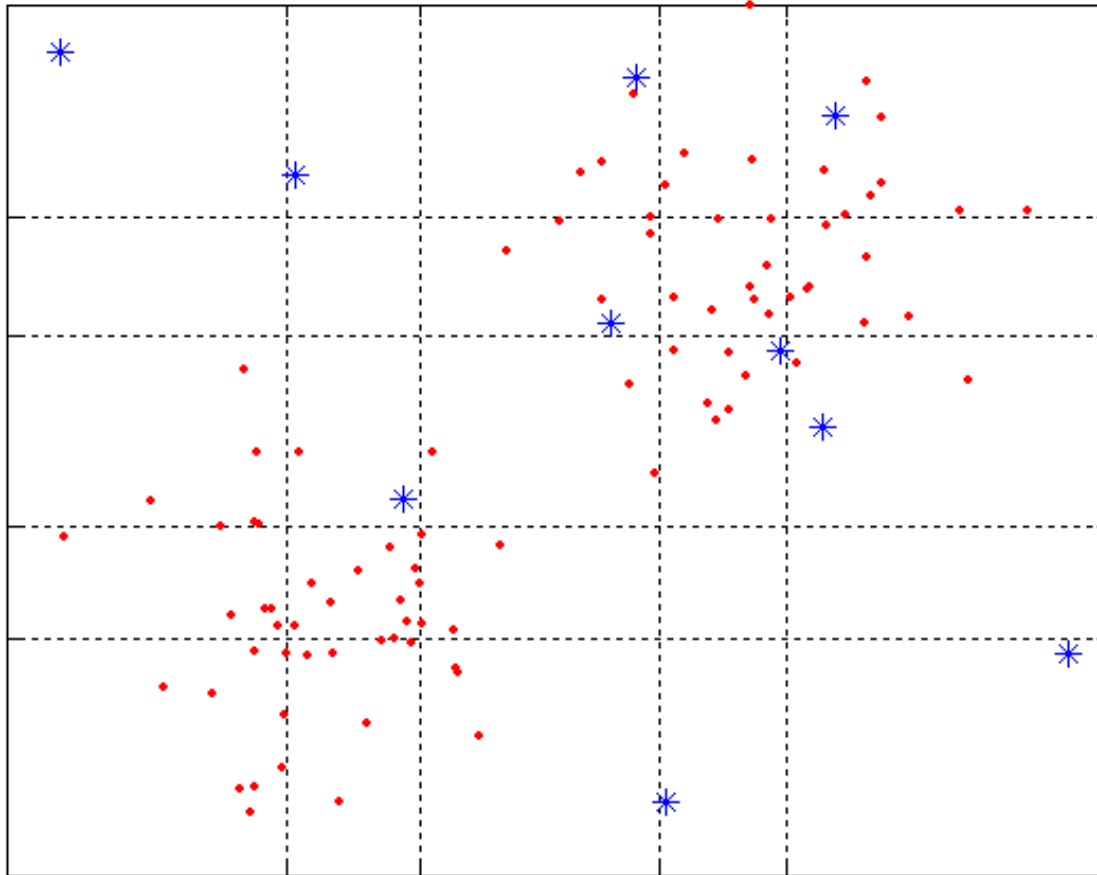
- Divide each attribute into ϕ equal-depth intervals
 - Each interval contains a fraction $f = 1/\phi$ of the records
- Consider a k -dimensional cube created by picking grid ranges from k different dimensions
 - If attributes are independent, we expect region to contain a fraction f^k of the records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

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

- Negative sparsity indicates cube contains smaller number of points than expected

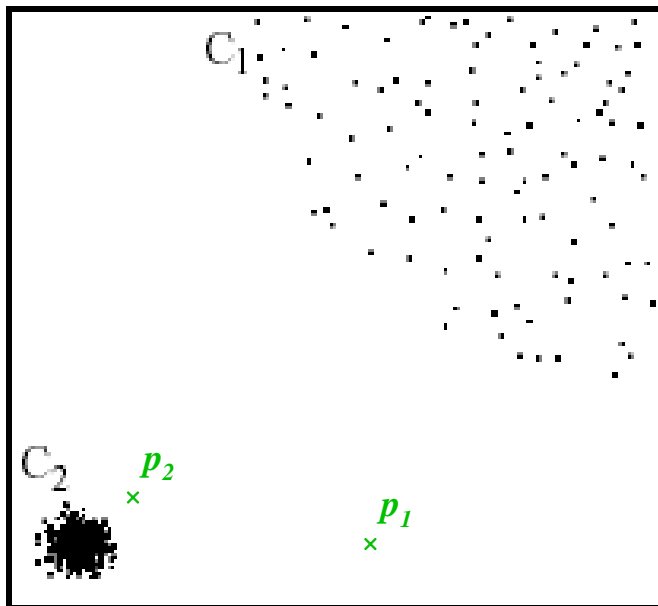
Example (but independence fails...)

- $N=100$, $\phi = 5$, $f = 1/5 = 0.2$, $N \times f^2 = 4$



Density-based: LOF approach

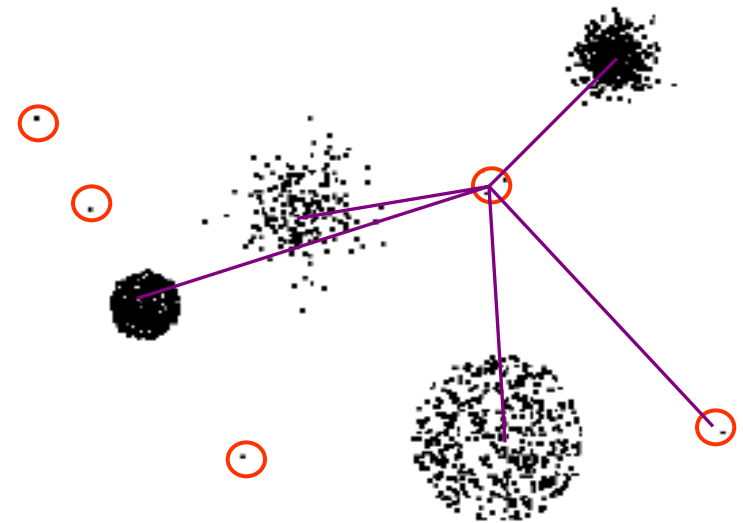
- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample p as the average of the ratios of the density of sample p and the density of its nearest neighbors
- Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Clustering-Based

- Basic idea:
 - Cluster the data into groups of different density
 - Choose points in small cluster as candidate outliers
 - Compute the distance between candidate points and non-candidate clusters.
 - ◆ If candidate points are far from all other non-candidate points, they are outliers



Interpreting outliers

- Given an outlier detection procedure, our goal is to detect “true outliers” and avoid “false alarms”.
- If our rule is too lax, we will get many “false alarms” (potentially costly).
- If the rule too strict, we will miss many “true outliers”.
- Can give some sense of what’s happening using Bayes’ rule.
- To obtain a high “yield” for our procedure, we need to be quite strict.

Base Rate Fallacy

- Recall 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 (Axelsson, 1999)

The base-rate fallacy is best described through example.² Suppose that your doctor performs a test that is 99% accurate, i.e. when the test was administered to a test population all of whom 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 doctor to learn 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 this information, is the probability of you having the disease? The reader is encouraged to make a quick “guesstimate” of the answer at this point.

Base Rate Fallacy

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

$$\begin{aligned} P(S|P) &= \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = \\ &= 0.00980 \dots \approx 1\% \end{aligned}$$

- 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 Intrusion Detection

- I: intrusive behavior,
¬I: non-intrusive behavior
A: alarm
¬A: no alarm
- Detection rate (true positive rate): $P(A|I)$
- False alarm rate: $P(A|\neg I)$
- Goal is to maximize both
 - Bayesian detection rate, $P(I|A)$
 - $P(\neg I|\neg A)$

Detection Rate vs False Alarm Rate

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

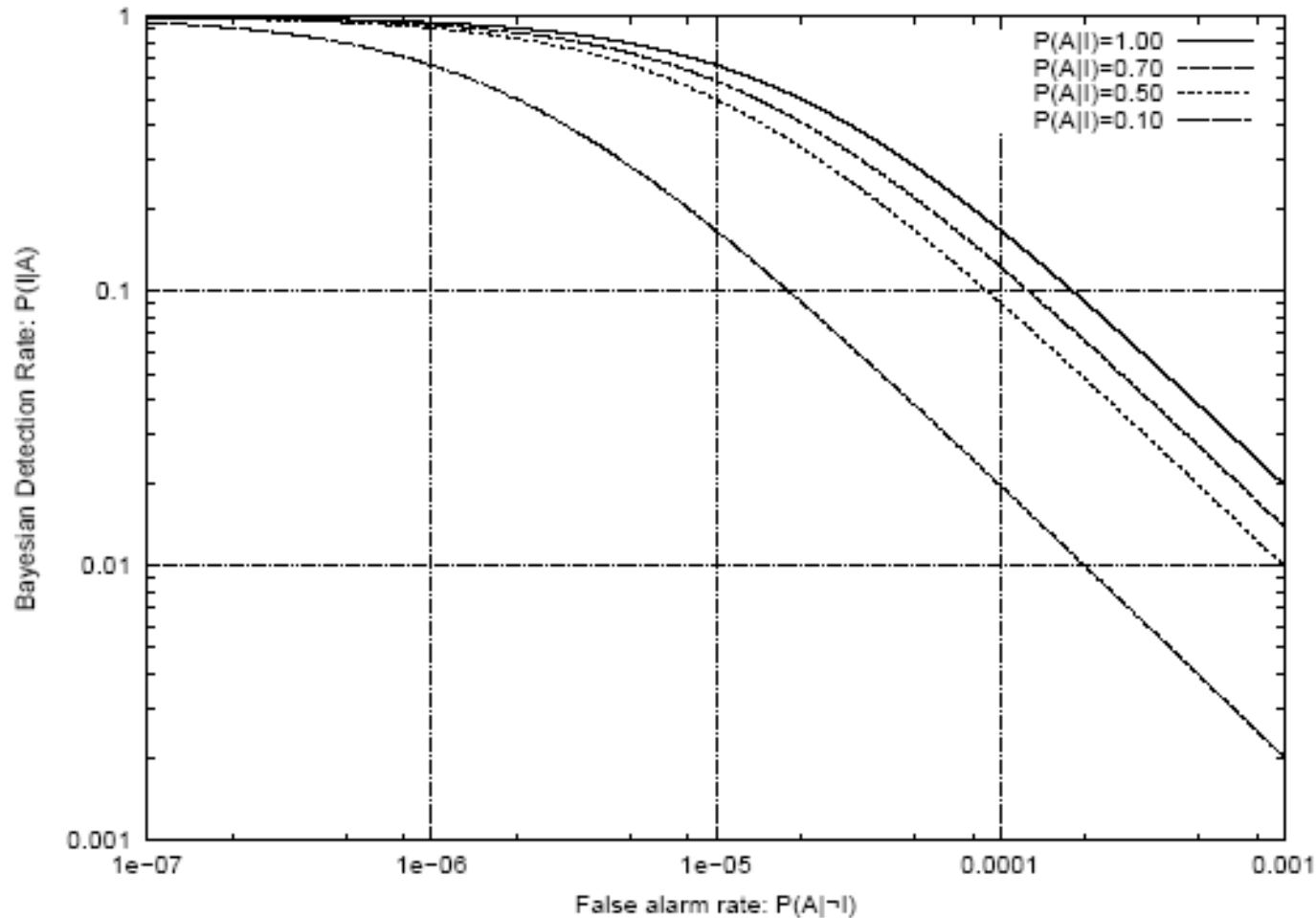
- Suppose: $P(I) = 1 / \frac{1 \cdot 10^6}{2 \cdot 10} = 2 \cdot 10^{-5};$
 $P(\neg I) = 1 - P(I) = 0.99998$

- Then:

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

- False alarm rate becomes more dominant if $P(I)$ is very low

Detection Rate vs False Alarm Rate



- Conclusion: We need a very low false alarm rate to achieve a reasonable Bayesian detection rate