Statistical Learning and Data Mining
CS 363D/ SSC 358

Lecture: Anomaly Detection

Prof. Pradeep Ravikumar
pradeepr@cs.utexas.edu

Adapted From: Pang-Ning Tan, Steinbach, Kumar

Anomaly 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, 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

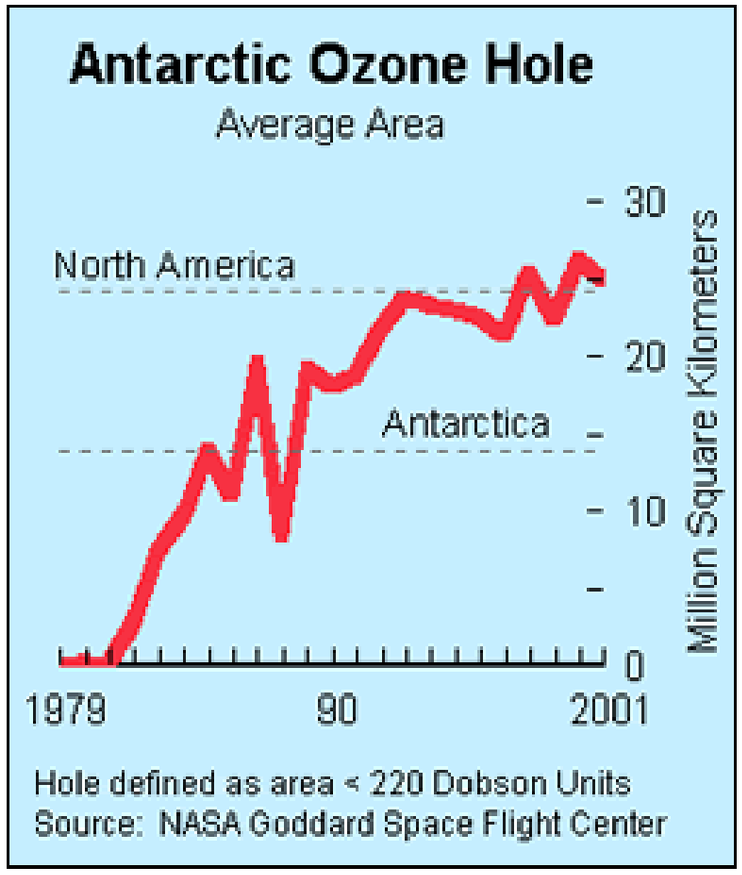
Importance of Anomaly Detection

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

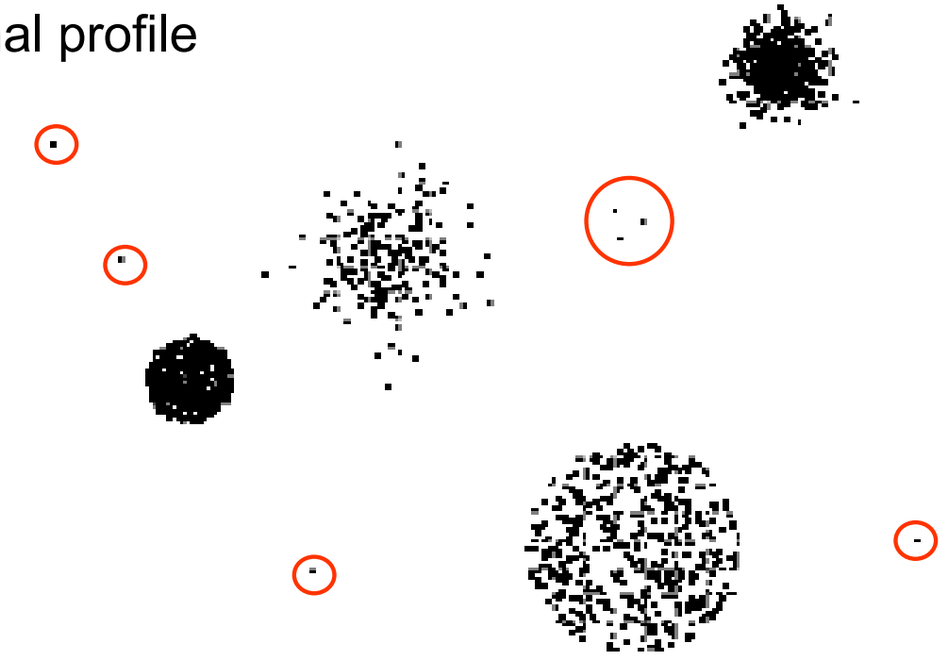
- Challenges
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging (just like for clustering)
  - Finding needle in a haystack

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

Anomaly Detection

- 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

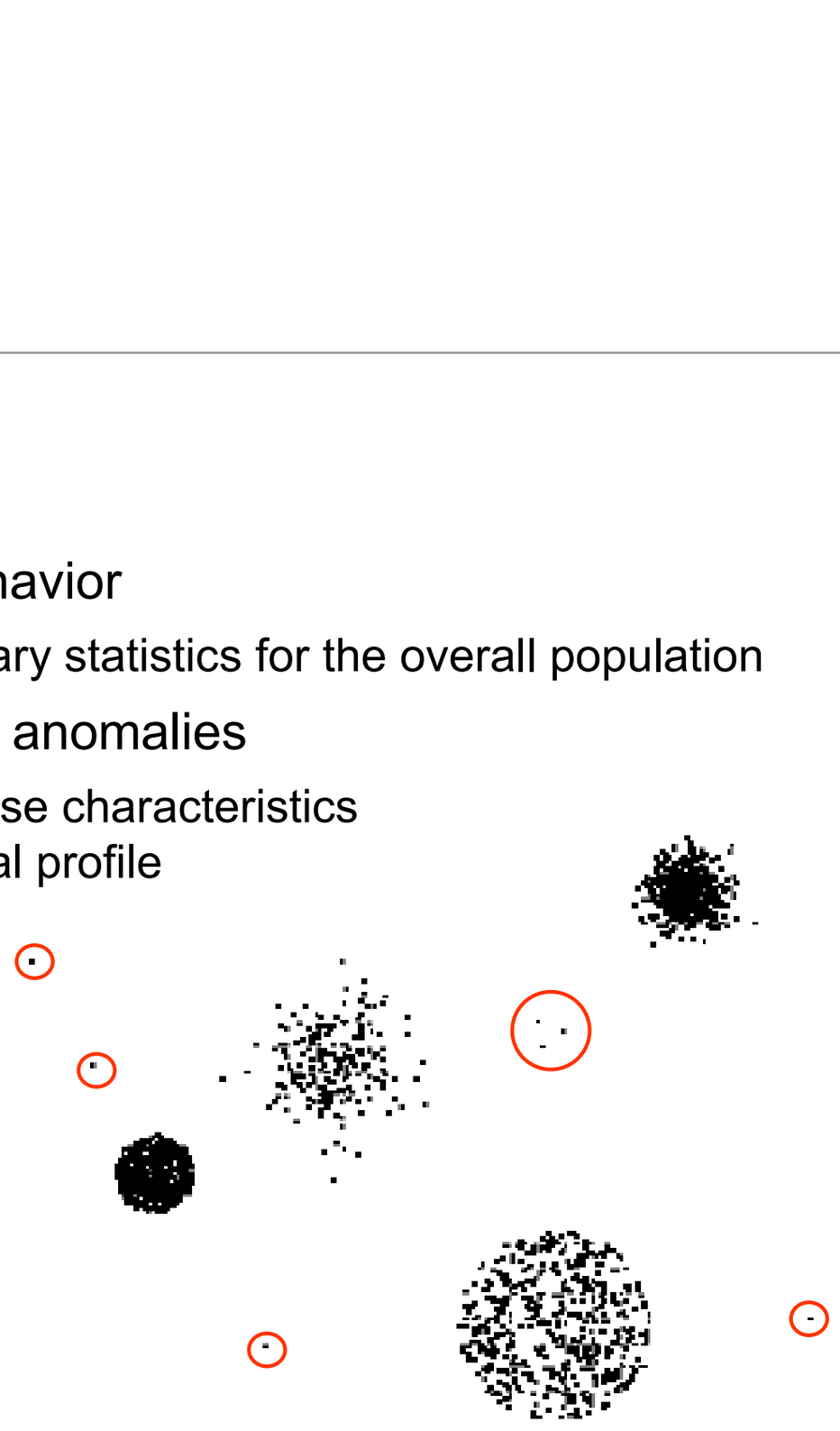
- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based



Anomaly Detection

- 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

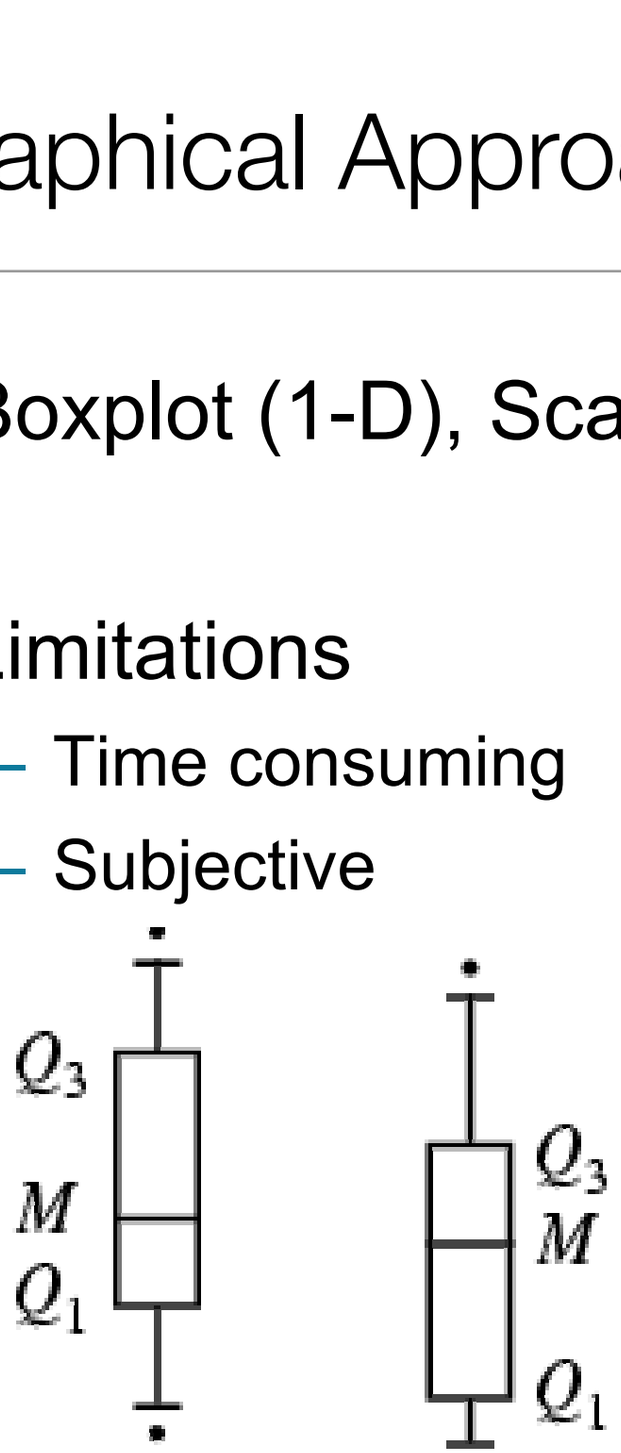
- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based
  - Model-based

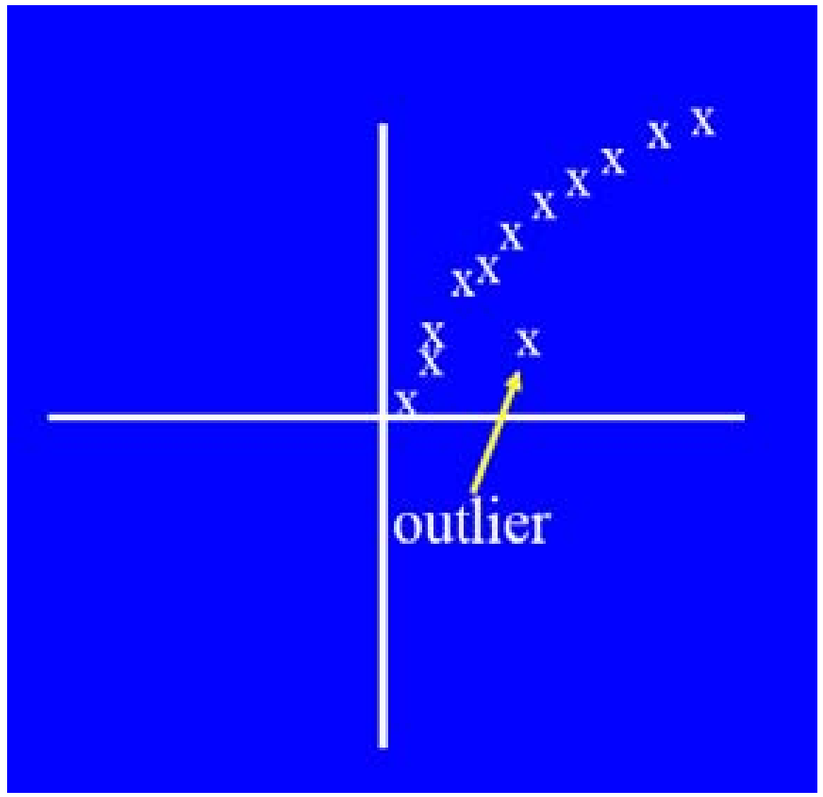


Graphical Approaches

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

- Limitations
  - Time consuming
  - Subjective





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:

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*

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 kth nearest neighbor is greatest

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 kth nearest neighbor is greatest

    - The top n data points whose average distance to the k nearest neighbors is greatest

Density-based: LOF Approach

- For each point, compute the density of its local neighborhood

Density-based: LOF Approach

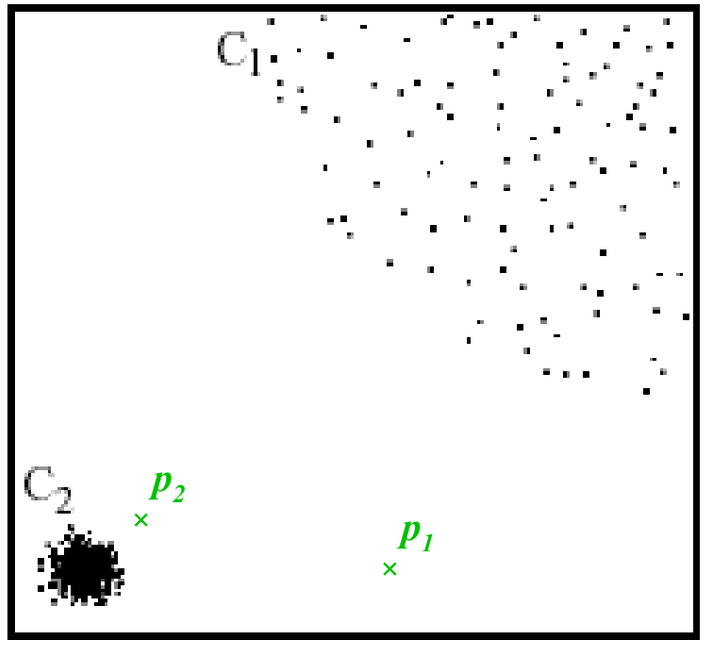
- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors

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

Density-based: LOF Approach

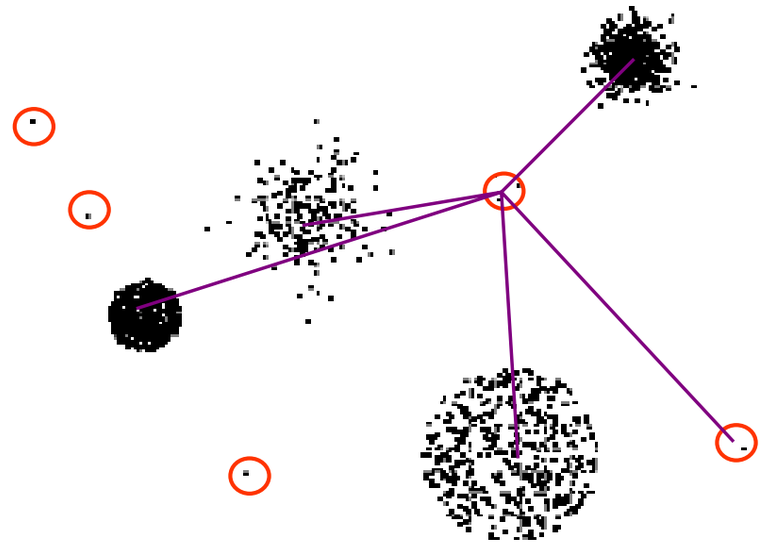
- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



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

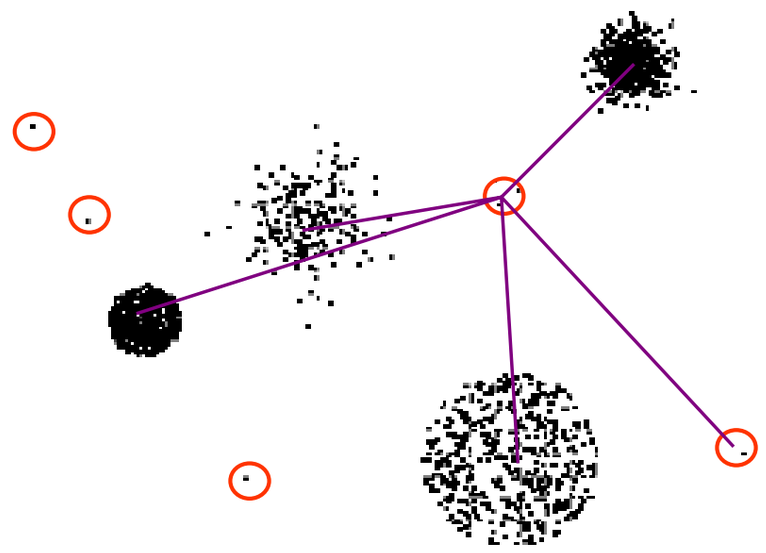
Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density



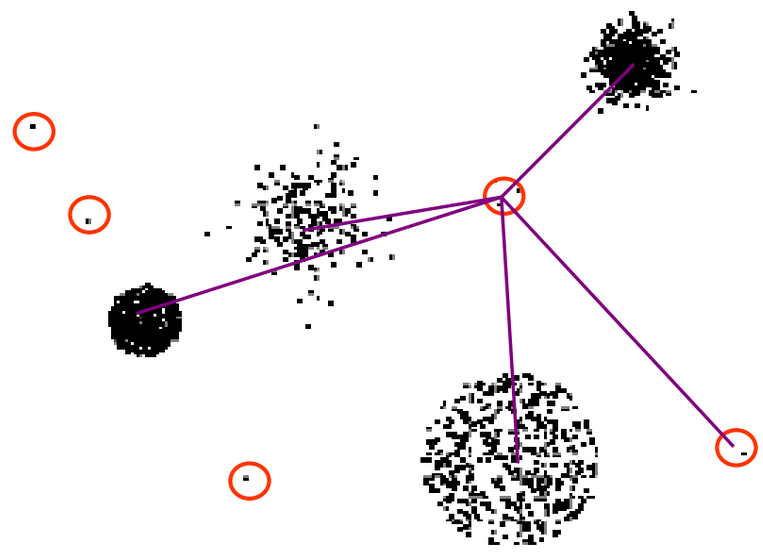
Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers



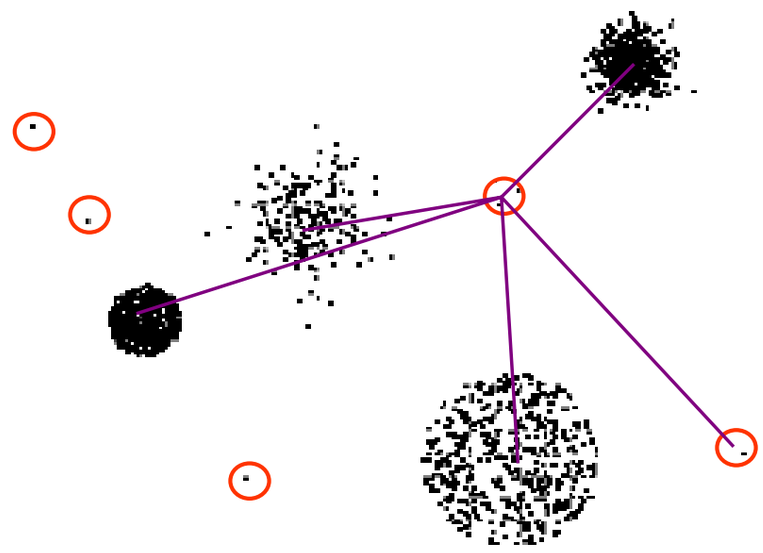
Clustering-Based

- 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.



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



Supervised Anomaly Detection: Classification

- If we have training data consisting of (attributes, anomaly-class-label) tuples, we could use classification techniques to predict the anomaly class label given any test data point

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 (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 (Axelsson, 1999)

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

$$P(S|P) = \frac{1/10000 \cdot 0.99}{1/10000 \cdot 0.99 + (1 - 1/10000) \cdot 0.01} = \\ = 0.00980\ldots \approx 1\%$$

- 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 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|¬I)

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