Lecture 3-4
Anomaly Detection using Distance-based Methods

Ref: Outlier Analysis, Charu C Agrawal
Ref: Chandola, Varun, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey." *ACM computing surveys (CSUR)* 41.3 (2009): 15.

Problem Diversity

- Case I: No labels are available
- Case II: Only normal data points are available
- Case III: Only abnormal data points are available
- Case IV: Both type of labelled data is available- delegate to CV/ML people :-)

Data Types

- Categorical - good, bad, and ugly
- Numerical - numbers
- Mixed - numbers as well as categories

Relationship Among Data Points

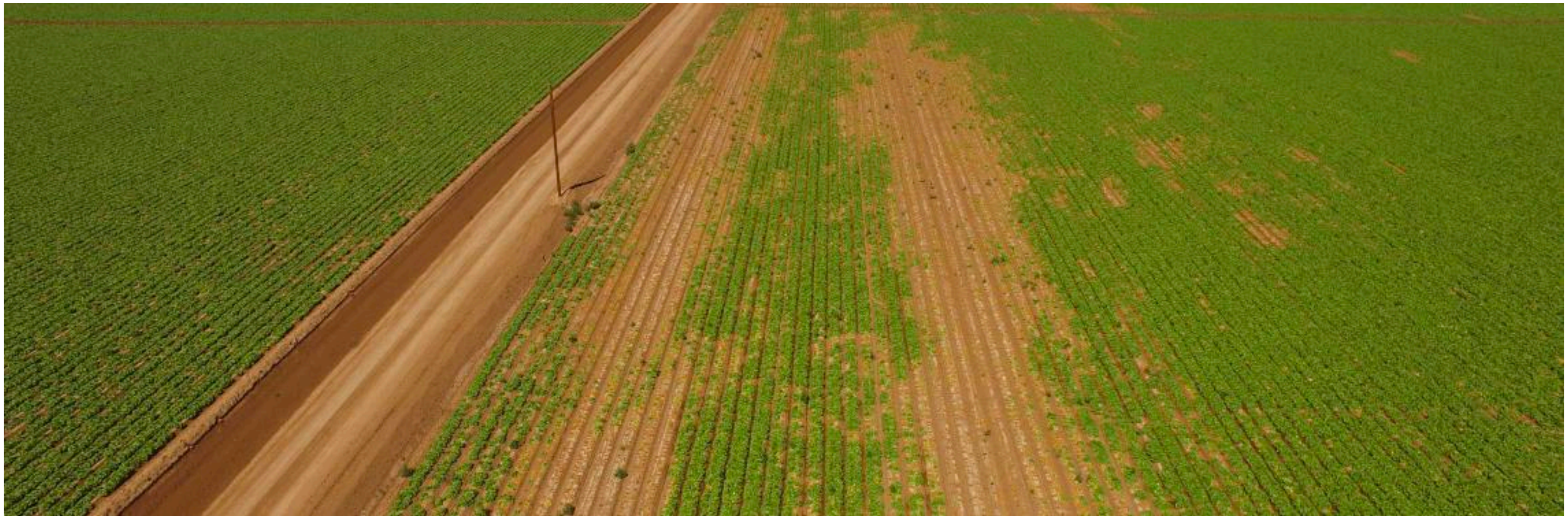
- No relationship among data points
- Graph network
- Spatially ordered data
- Temporally ordered data points

Is 3 an anomaly?

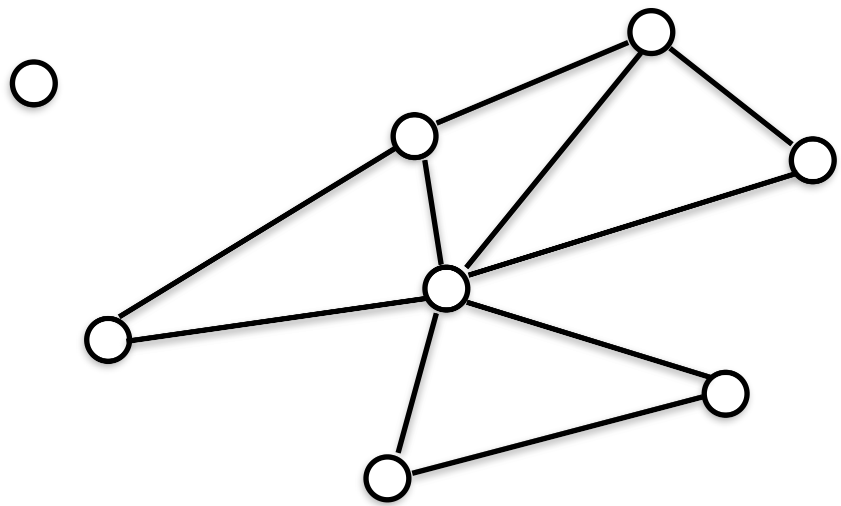
3, 2, 3, 2, 3, 87, 86, 85 87, 89, 86, 3, 84, 91, 86, 91, 88

time

Spatial Anomaly Example



Network/Graph Anomaly



Related Data

- The relationship may provide anomaly detection criteria
- Such anomalies are also called contextual anomalies

Univariate and Multivariate Outliers

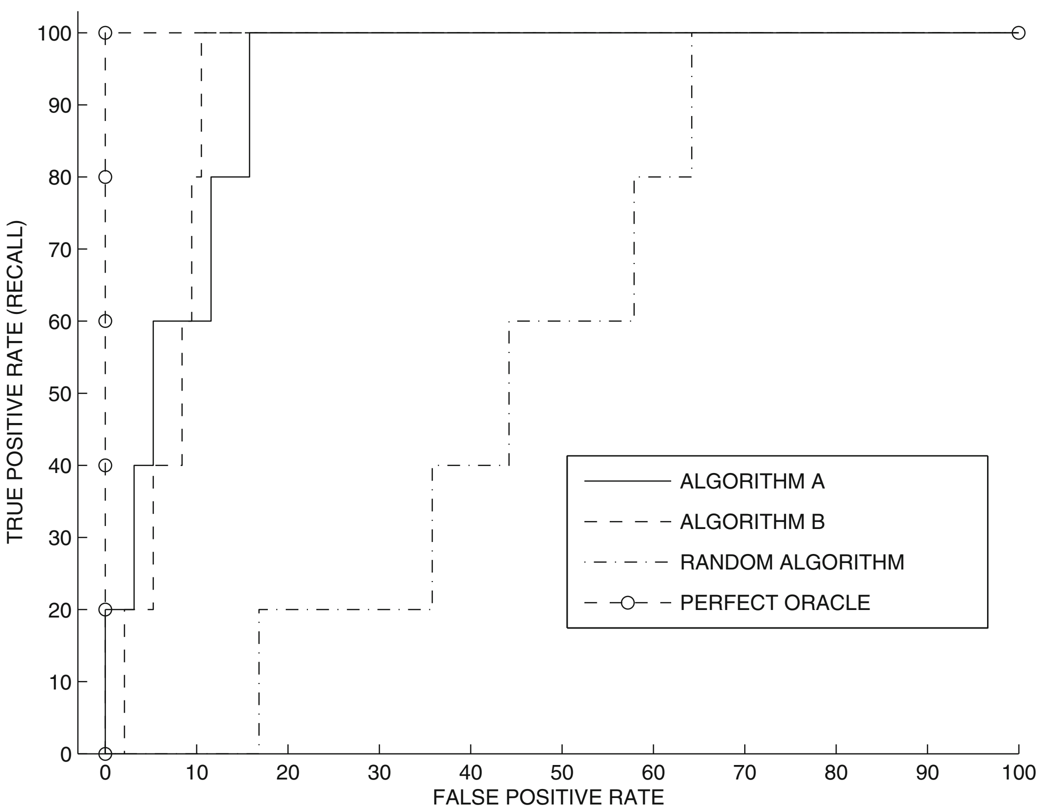
- Univariate: Data point consists of one variable
- Multivariate: Data point consists of at least two variables

Outlier Evaluation Technique

$$\text{Precision} = \frac{|S(\theta) \cap G|}{|S(\theta)|}$$

$$\text{Recall} = \frac{|S(\theta) \cap G|}{|G|}$$

Receiver Operating Characteristic Curve (ROC)



$$TPR = 100\frac{|S(\theta) \cap G|}{|G|}$$

$$FPR = 100\frac{|S(\theta) - G|}{|D - G|}$$

Z-Value Test Limitations

- Data may not be Gaussian distributed
- Sufficient samples may not be available to robustly estimate mean and standard deviation
- Applies to only univariate data points

Nearest Neighbour-based Anomaly Detection

- Need a similarity measure defined between two data points!
- For continuous attributes, Euclidean distance is popular!
- For categorical data, matching techniques are used, e.g., hamming distance
- The distance measure should be symmetric

Assumption: Normal data instances occur in dense neighbourhood!

Two Approaches

- Distance of nth nearest neighbour as anomaly score
- Relative density based anomaly score

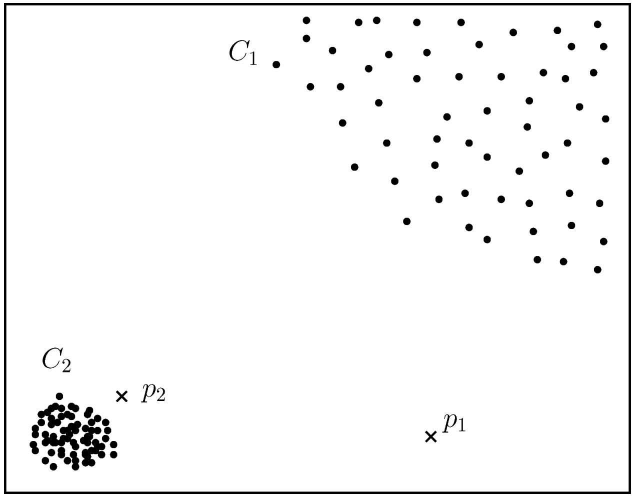
K-NN Distance-based Anomaly

- A non-parametric model
- For each data point, find kth nearest neighbour
- K is generally a small number
- A large distance means anomaly

Density-based Anomaly

- Calculate density of neighbourhood of each data point
- Low density indicates anomaly
- How to calculate density?

Using Inverse of KNN distance as density indicator!



Prob: Many points in C1 will have lower density than point p2!

Soln: calculate density relative to its neighbours!

Ref: Anomaly Detection: A Survey. Chandola et al.

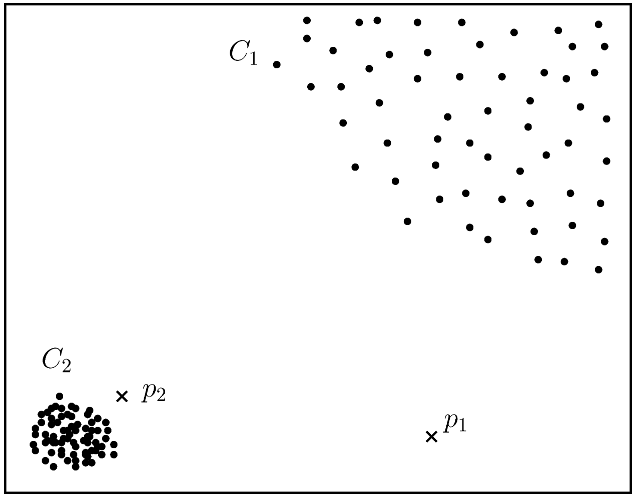
Relative Local Density

- Calculate the distance d of kth nearest neighbour
- Calculate the volume v of the hypersphere with radius d
- The local density at that point is calculated as k/v

Local Outlier Factor (LOF)

- Find the local density of k nearest neighbours
- Ratio of average local density of k nearest neighbours and the given point is LOF score of the point
- Anomaly will have higher LOF score

P2 will have high LOF score in comparison to points in C1.



Ref: Anomaly Detection: A Survey. Chandola et al.

Pros and Cons

- Pros
  - Unsupervised
  - Data driven, no assumption about distribution
- Cons
  - If normal instances do not have enough neighbours, the method will fail

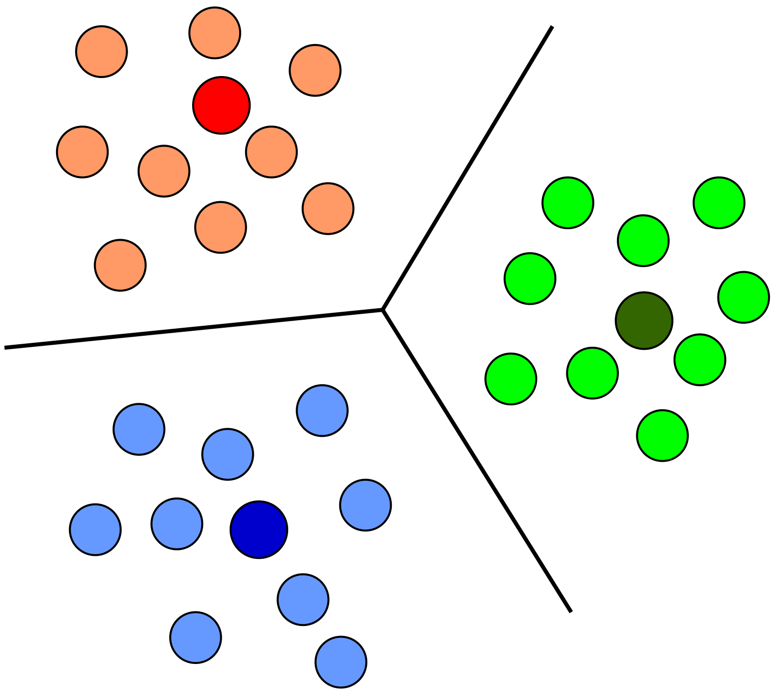
Clustering-based Anomaly Detection Methods

- Group similar data instances into clusters
- Analyse the clustered data to find anomalies

Case I: Normal data instances lie close to their nearest cluster centroid, while anomalies are far away from their closest cluster centroid.

- Consists of two steps
- First step is to find clusters using any standard algorithm
- Anomaly score is the distance from the nearest centroid

How to find the clusters?



Linde–Buzo–Gray Algorithm for k-Means Clustering

1. Guess the cluster centroids $C=\{c_1,c_2,\ldots,c_K\}$;
2. REPEAT

   - For each training vector x_j, find the nearest cluster centroid: $q_j = \arg\min_k ||x_j - c_k||$

   - For each cluster k, re-calculate the cluster centroid from the vectors assigned to the cluster: $c_k = \text{mean } \{x_j | q_j=k\}$

   - UNTIL convergence

Obtaining Cluster Centroids

Input vectors: $S = \{X_i \in R^d \mid \text{i=1, 2, ..., n}\}$

Initial centroids: $C = \{C_j \in R^d \mid \text{j=1, 2, ..., k}\}$

Obtain clusters: $X_i \in S_q$ if $||x_i - C_q||_p \leq ||X_i - C_j||_p$

Update centroids: $C_j = \dfrac{1}{|S_i|} \sum_{X_i \in S_j} X_i$

Calculate distortion: $D_k = \sum_{j=1}^{K} \sum_{X_i \in S_i} ||X_i - C_j||_p$

$\boxed{\text{Repeat until distortion < threshold}}$

The codebook: $C = \{C_j \in R^d \mid \text{j=1, 2, ..., k}\}$

Take the distance from nearest centroid as the anomaly score!

Limitations

- Only works with spherical clusters
- Difficult to know k in advance
- To find k, use hierarchical or agglomerative clustering

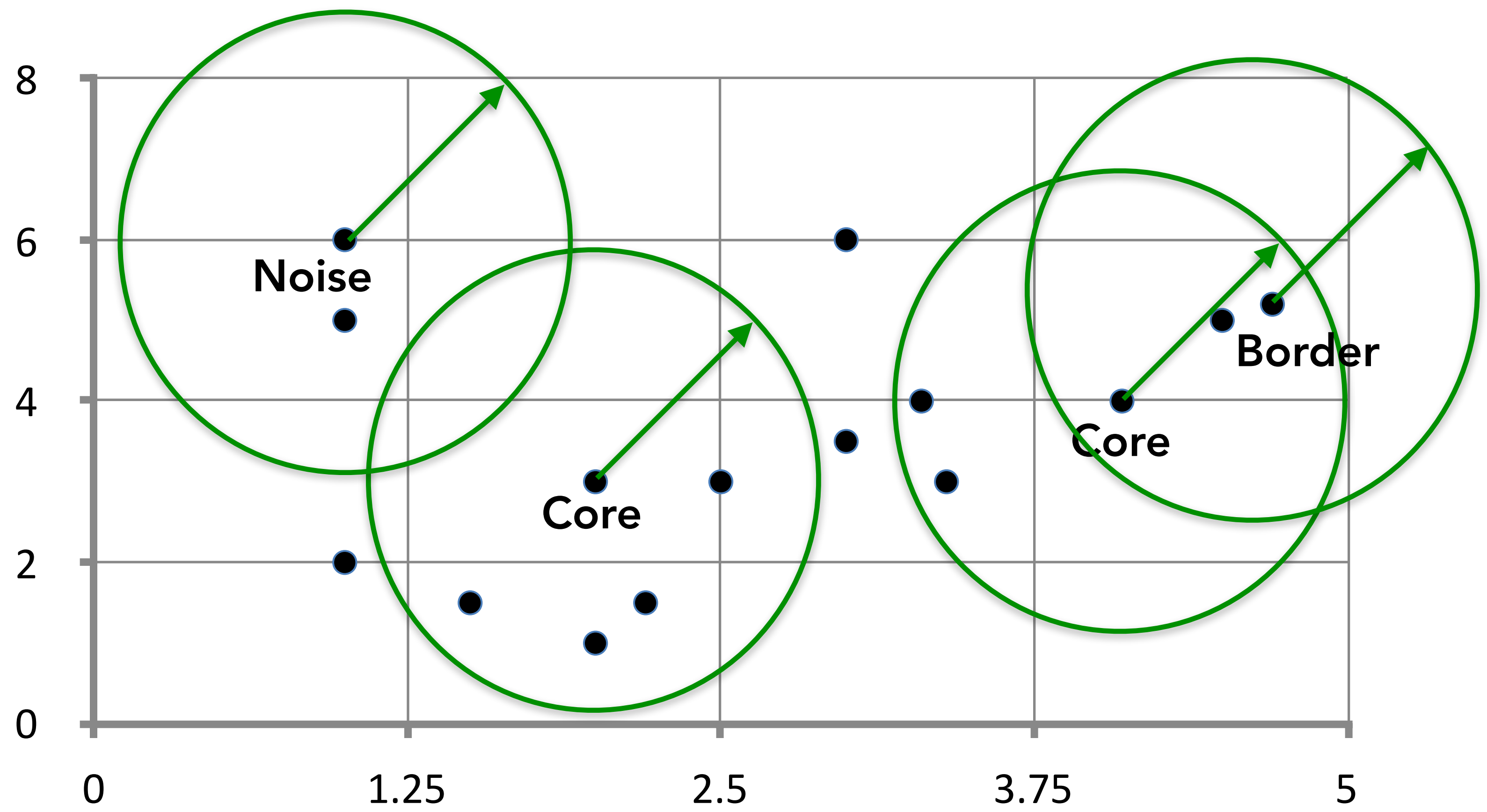
Case II: Normal data instances belong to a cluster in the data, while anomalies do not belong to any cluster.

- Use clustering algorithms that do not force each data point to be associated with a cluster
- Data points not associated with any cluster are anomalies
- Example: DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN)

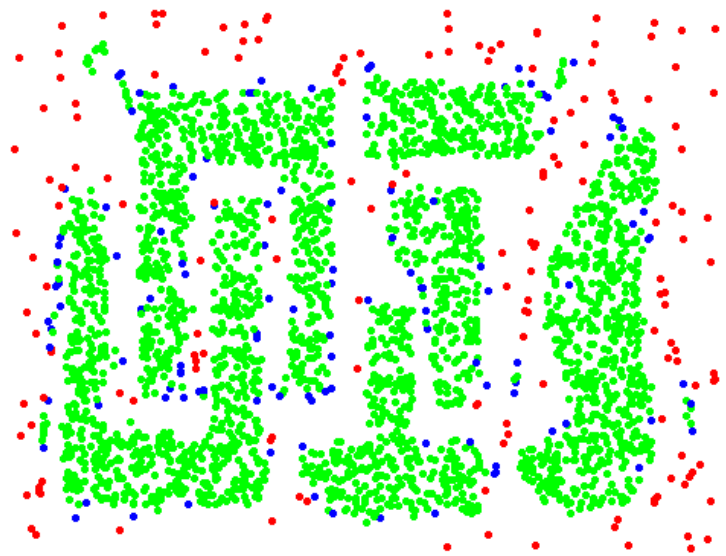
- Divides the points into three types: **core points**, **border points**, **and noise**
- If there are more than MinPts around a point within eps distance, it's a **core point**
- If a point is not a core point, but within eps distance from a core point, it is a **border point**
- Else, it is a **noise, outlier, or anomaly**

Example



Another Example

Original Points

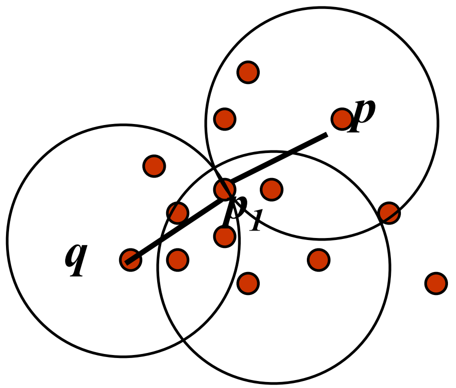


Point types: core, border and noise

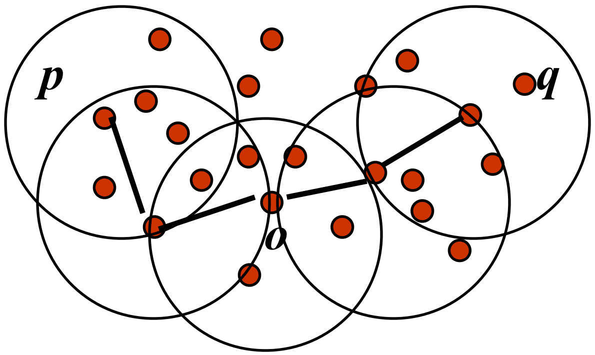
Eps = 10, MinPts = 4

Density-Connected points

- Density edge
  - We place an edge between two core points q and p if they are within distance Eps.



- Density-connected
  - A point p is density-connected to a point q if there is a path of edges from p to q



DBSCAN Algorithm

- Label points as core, border and noise
- For every core point p that has not been assigned to a cluster
  - Create a new cluster with the point p and all the points that are density-connected to p.
- Repeat until all points are visited.
- Points not assigned to any cluster are anomalies.

Benefits of DBSCAN

- Can find arbitrary shape clusters, while k-means (and most other) can only find spherical clusters
- It is effective in handling noise as it does not forces cluster association to each data point

The previous two techniques will not work if the anomalies also form a cluster!

Case III: Normal data instances belong to large and dense clusters, while anomalies either belong to small or sparse clusters.

- The goal is to tag the clusters as anomalous
- Anomaly clusters are generally small and sparse
- A possible metric is size/distortion or size/variance of each cluster.