Data Mining

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

1

1

Data Mining

Anomaly Detection

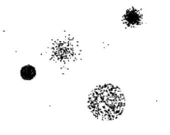
- Outline
  - Characteristics of Anomaly Detection Problems
  - Characteristics of Anomaly Detection Methods
  - Statistical Approaches
  - Proximity-based Approaches
  - Clustering-based Approaches
  - Reconstruction-based Approaches
  - One-class Classification
  - Information Theoretic Approaches
  - Evaluation of Anomaly Detection

2

2

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data



- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July
- Can be important or a nuisance
  - Unusually high blood pressure
  - 100 kg, 2 year old

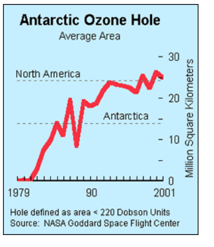
3

3

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!



Source:

http://www.epa.gov/ozone/science/hole/size.html

4

4

Causes of Anomalies

- Data from different classes
  - Measuring the weights of oranges, but a few grapefruit are mixed in
- Natural variation
  - Unusually tall people
- Data errors
  - 100 kg 2 year old

5

5

Distinction Between Noise and Anomalies

- Noise doesn't necessarily produce unusual values or objects
- Noise is not interesting
- Noise and anomalies are related but distinct concepts

6

6

Model-based vs Model-free

- Model-based Approaches
  - Model can be parametric or non-parametric
  - Anomalies are those points that don't fit well
  - Anomalies are those points that distort the model
- Model-free Approaches
  - Anomalies are identified directly from the data without building a model
- Often the underlying assumption is that most of the points in the data are normal

7

General Issues

- Global vs. Local Perspective
  - An instance can be identified as an anomaly by
    - building a model over all normal instances and using this global model for anomaly detection
    - by considering the local perspective of every data instance
      - an anomaly detection approach is termed local if its output on a given instance does not change if instances outside its local neighborhood are modified or removed
- Label vs Score
  - Some anomaly detection techniques provide only a binary categorization (anomali or normal)
  - Other approaches measure the degree to which an object is an anomaly
    - This allows objects to be ranked
    - Scores can also have associated meaning (e.g., statistical significance)

8

Anomaly Detection Techniques

- Statistical Approaches
- Proximity-based
  - Anomalies are points far away from other points
- Clustering-based
  - Points far away from cluster centers are outliers
  - Small clusters are outliers
- Reconstruction Based
  - rely on the assumption that the normal class resides in a space of lower dimensionality than the original space of attributes

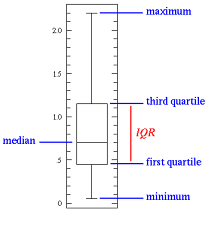
9

Statistical Approaches

- **Probabilistic definition of an outlier**:
  - An outlier is an object that has a low probability with respect to a probability distribution model of the data.
- Usually assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on
  - Data distribution
  - Parameters of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)
- Issues
  - Identifying the distribution of a data set
    - Heavy tailed distribution
  - Number of attributes
  - Is the data a mixture of distributions?

10

Boxplot



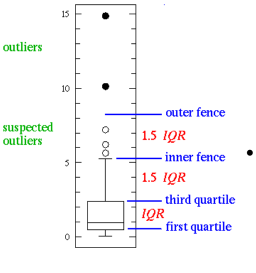
This simplest possible box plot displays the full range of variation (from min to max), the likely range of variation (the IQR), and a typical value (the median). Not uncommonly real datasets will display surprisingly high maximums or surprisingly low minimums called outliers.

John Tukey has provided a precise definition for two types of outliers:

- Outliers are either 3×IQR or more above the third quartile or 3×IQR or more below the first quartile.
- Suspected outliers are slightly more central versions of outliers:
  - either 1.5×IQR or more above the third quartile
    - (Q3 + 1.5 x IQR)
  - or 1.5×IQR or more below the first quartile
    - (Q1-1.5 x IQR)

11

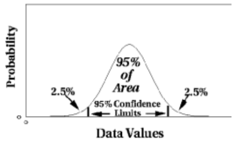
Boxplot



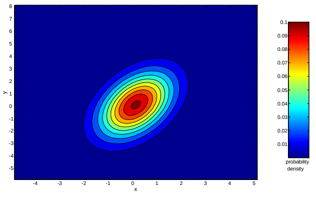
- If either type of outlier is present
  - the whisker on the appropriate side is taken to 1.5×IQR from the quartile (the "inner fence") rather than the max or min,
- individual outlying data points are displayed as
  - unfilled circles for suspected outliers
  - or filled circles for outliers.
- The "outer fence" is 3×IQR from the quartile.

12

Normal Distributions



One-dimensional Gaussian



Two-dimensional Gaussian

13

13

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: $G = \dfrac{\max|X - \bar{X}|}{s}$
- Reject H_0 if: $G > \dfrac{(N-1)}{\sqrt{N}}\sqrt{\dfrac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$

14

14

Statistically-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 time 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

15

15

Statistically-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 initially assumed to be uniform distribution
- Likelihood at time t:

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

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

16

16

Strengths/Weaknesses of Statistical Approaches

- Firm mathematical foundation
- Can be very efficient
- Good results if distribution is known
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution
- Anomalies can distort the parameters of the distribution

17

17

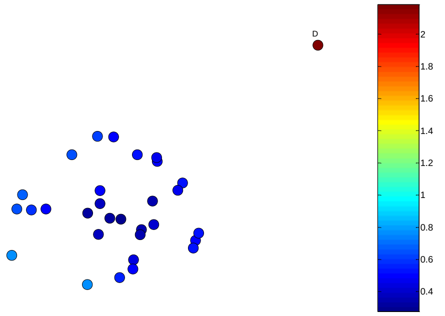
Distance-Based Approaches

- The outlier score of an object is the distance to its kth nearest neighbor

18

18

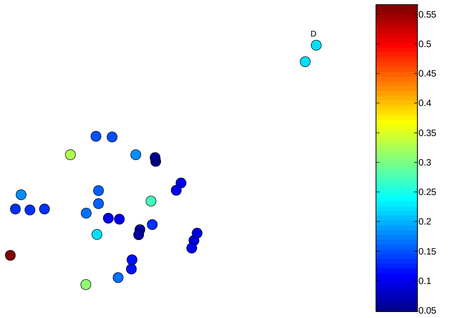
One Nearest Neighbor - One Outlier



19

19

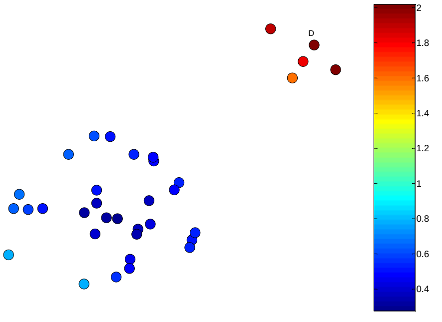
One Nearest Neighbor - Two Outliers



20

20

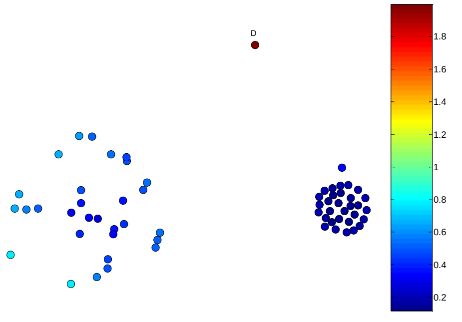
Five Nearest Neighbors - Small Cluster



21

21

Five Nearest Neighbors - Differing Density



22

22

Strengths/Weaknesses of Distance-Based Approaches

- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Sensitive to variations in density
- Distance becomes less meaningful in high-dimensional space

23

23

Density-Based Approaches

- **Density-based Outlier:**
  - The outlier score of an object is the inverse of the density around the object.
  - Can be defined in terms of the k nearest neighbors
  - One definition:
    - Inverse of distance to kth neighbor
  - Another definition:
    - Inverse of the average distance to k neighbors
  - DBSCAN definition
    - Density-Based Spatial Clustering of Applications with Noise
- If there are regions of different density, this approach can have problems

24

24

Relative Density

- Consider the density of a point relative to that of its k nearest neighbors
- Let $y_1, \dots, y_k$ be the k nearest neighbors of $\boldsymbol{x}$

$$density(\boldsymbol{x}, k) = \frac{1}{dist(\boldsymbol{x},k)} = \frac{1}{dist(\boldsymbol{x},\boldsymbol{y}_k)}$$

$$relative\ density(\boldsymbol{x}, k) = \frac{\sum_{i=1}^{k} density(\boldsymbol{y}_i,k)/k}{density(\boldsymbol{x},k)}$$

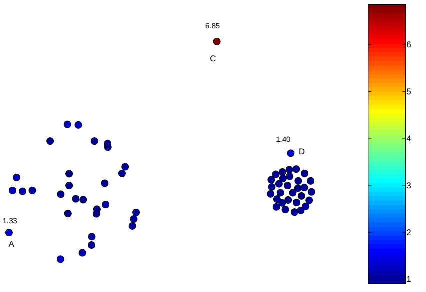
$$= \frac{dist(\boldsymbol{x},k)}{\sum_{i=1}^{k} dist(\boldsymbol{y}_i,k)/k} = \frac{dist(\boldsymbol{x},\boldsymbol{y})}{\sum_{i=1}^{k} dist(\boldsymbol{y}_i,k)/k}$$

- Can use average distance instead

25

25

Relative Density Outlier Scores

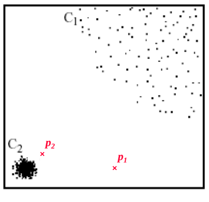


26

26

Relative 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

27

27

Strengths/Weaknesses of Density-Based Approaches

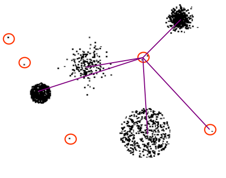
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

28

28

Clustering-Based Approaches

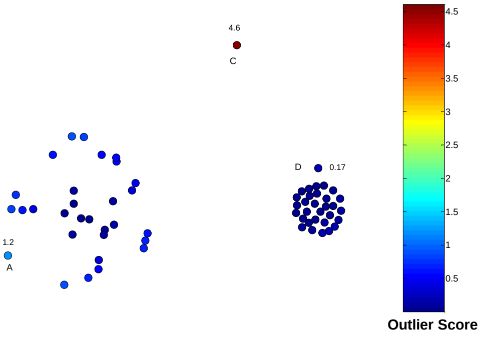
- An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
    - Outliers can impact the clustering produced
  - For density-based clusters, an objec is an outlier if its density is too low
    - Can't distinguish between noise and outliers
  - For graph-based clusters, an object is an outlier if it is not well connected



29

29

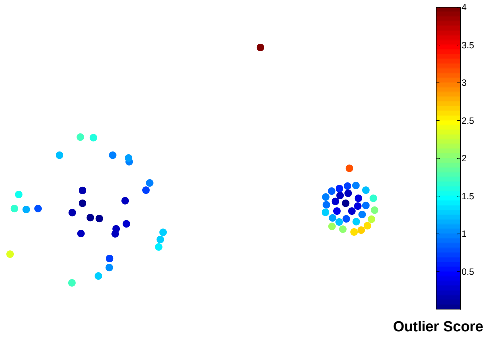
Distance of Points from Closest Centroids



30

30

Relative Distance of Points from Closest Centroid



31

31

Strengths/Weaknesses of Clustering-Based Approaches

- Simple
- Many clustering techniques can be used
- Can be difficult to decide on a clustering technique
- Can be difficult to decide on number of clusters
- Outliers can distort the clusters

32

32

Reconstruction-Based Approaches

- Based on assumptions there are patterns in the distribution of the normal class that can be captured using lower-dimensional representations
- Reduce data to lower dimensional data
  - E.g. Use Principal Components Analysis (PCA) or Auto-encoders
- Measure the reconstruction error for each object
  - The difference between original and reduced dimensionality version

33

33

Reconstruction Error

- Let $\mathbf{x}$ be the original data object
- Find the representation of the object in a lower dimensional space
- Project the object back to the original space
- Call this object $\hat{\mathbf{x}}$

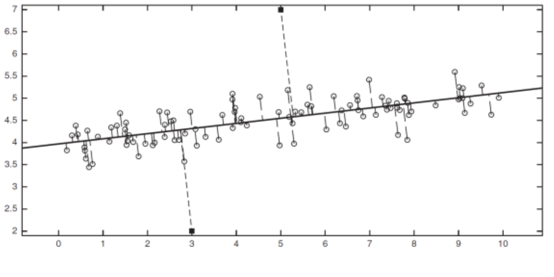
$$\text{Reconstruction Error}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|$$

- Objects with large reconstruction errors are anomalies

34

34

Reconstruction of two-dimensional data

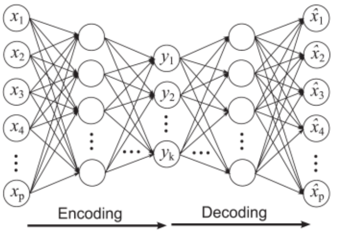


35

35

Basic Architecture of an Autoencoder

- An autoencoder is a multi-layer neural network
- The number of input and output neurons is equal to the number of original attributes.



36

36

Strengths and Weaknesses

- Does not require assumptions about distribution of normal class
- Can use many dimensionality reduction approaches
- The reconstruction error is computed in the original space
  - This can be a problem if dimensionality is high

37

37

One Class SVM

- Uses an SVM approach to classify normal objects
- Uses the given data to construct such a model
- This data may contain outliers
- But the data does not contain class labels
- How to build a classifier given one class?

38

38

How Does One-Class SVM Work?

- Uses the “origin” trick
- Use a Gaussian kernel $\kappa(\mathbf{x},\mathbf{y}) = \exp(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2\sigma^2})$
  - Every point mapped to a unit hypersphere

$$\kappa(\mathbf{x},\mathbf{x}) = \langle\phi(\mathbf{x}),\phi(\mathbf{x})\rangle = ||\phi(\mathbf{x})||^2 = 1$$

  - Every point in the same orthant (quadrant)

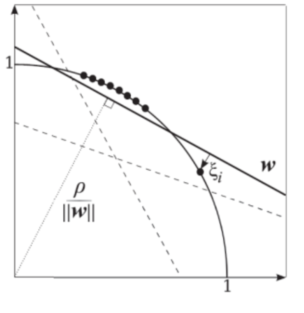
$$\kappa(\mathbf{x},\mathbf{y}) = \langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle \geq 0$$

- Aim to maximize the distance of the separating plane from the origin

39

39

Two-dimensional One Class SVM



40

40

Equations for One-Class SVM

- Equation of hyperplane $\langle\mathbf{w},\phi(\mathbf{x})\rangle = \rho$
- ϕ is the mapping to high dimensional space
- Weight vector is $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x_i})$
- ν is fraction of outliers
- Optimization condition is the following

$$\min_{\mathbf{w},\ \rho,\ \xi} \frac{1}{2}||\mathbf{w}||^2 - \rho + \frac{1}{n\nu}\sum_{i=1}^{n}\xi_i,$$

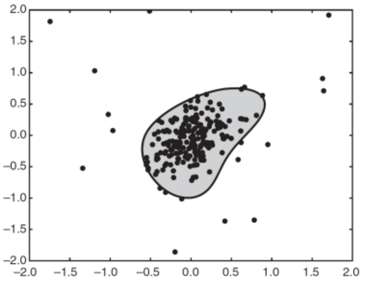
$$\text{subject to: } \langle\mathbf{w},\phi(\mathbf{x_i})\rangle \geq \rho - \xi_i,\ \ \xi_i \geq 0$$

41

41

Finding Outliers with a One-Class SVM

- Decision boundary with $\nu = 0.1$

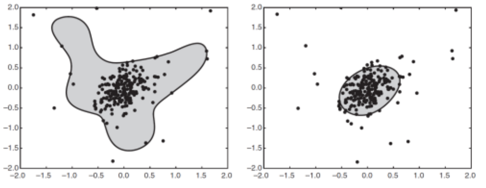


42

42

Finding Outliers with a One-Class SVM

- Decision boundary with $\nu = 0.05$ and $\nu = 0.2$



(a) $\nu = 0.05$. (b) $\nu = 0.2$.

43

43

Strengths and Weaknesses

- Strong theoretical foundation
- Choice of ν is difficult
- Computationally expensive

44

44

Information Theoretic Approaches

- Key idea is to measure how much information decreases when you delete an observation

$$Gain(x) = Info(D) - Info(D \setminus x)$$

- Anomalies should show higher gain
- Normal points should have less gain

45

45

Information Theoretic Example

- Survey of height and weight for 100 participants

weight	height	Frequency
low	low	20
low	medium	15
medium	medium	40
high	high	20
high	low	5

- Eliminating last group give a gain of $2.08 - 1.89 = 0.19$

46

46

Strengths and Weaknesses

- Solid theoretical foundation
- Theoretically applicable to all kinds of data
- Difficult and computationally expensive to implement in practice

47

47

Evaluation of Anomaly Detection

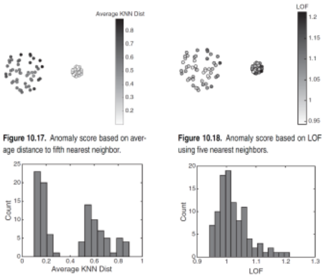
- If class labels are present, then use standard evaluation approaches for rare class such as precision, recall, or false positive rate
  – FPR is also know as false alarm rate
- For unsupervised anomaly detection use measures provided by the anomaly method
  – E.g. reconstruction error or gain
- Can also look at histograms of anomaly scores.

48

48

Distribution of Anomaly Scores

- Anomaly scores should show a tail



49

49

50

50