Artificial Intelligence for Medicine II

Spring 2025

Lecture 10: Anomaly Detection

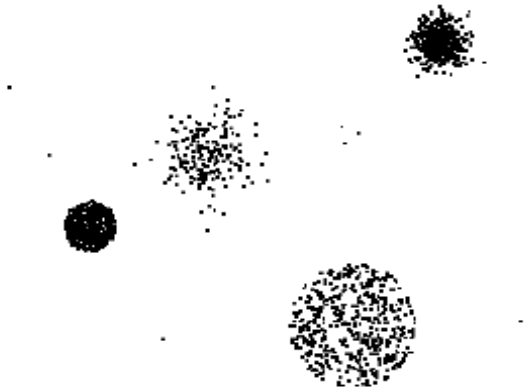
(Many slides adapted from Bing Liu, Han, Kamber & Pei; Tan, Steinbach, Kumar and the web)

Anomaly Detection

- *In anomaly detection, the goal is to find objects that do **not** conform to **normal patterns or behavior**.*
- *Often, anomalous objects are known as **outliers**, since, on a scatter plot of the data, they lie far away from other data points.*
- *Anomaly detection is also known as **deviation detection**, because anomalous objects have attribute values that deviate significantly from the expected or typical attribute values, or as **exception mining**.*
- Applications:
  - *Fraud Detection*
  - *Intrusion Detection*
  - *Ecosystem Disturbances*
  - *Medicine and Public Health*
    - For a particular patient, unusual symptoms or test results, such as an anomalous MRI scan, may indicate potential health problems.

Anomaly/Outlier Detection

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



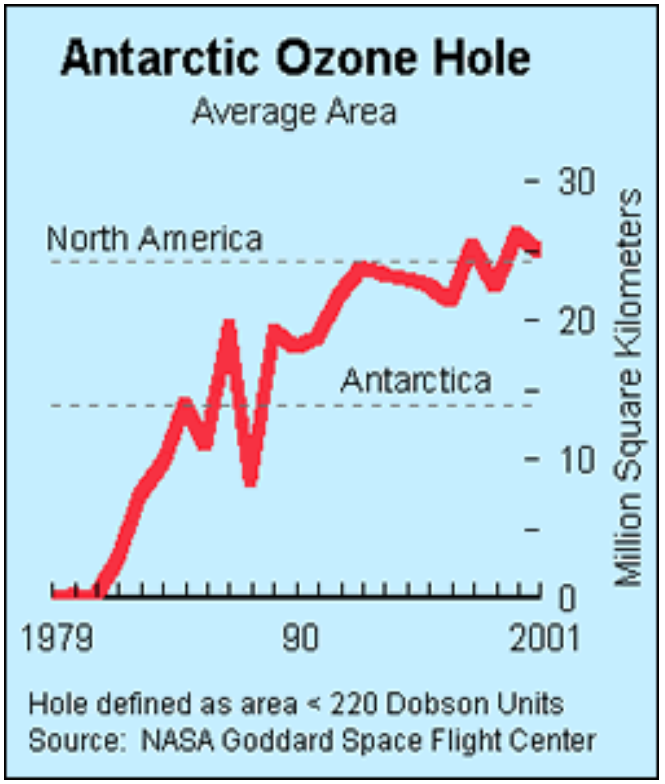
- Natural implication is that anomalies are relatively rare
  - One in a thousand occurs often if you have lots of data
  - Context is important, e.g., freezing temps in July

- Can be important or a nuisance
  - Unusually high blood pressure
  - 200 pound, 2 year old

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!
  - Since the Antarctic ozone depletion was far beyond anticipated values, the system ignored the most critical data.



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

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
  - 200 pound 2 year old

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

Model-based vs Model-free

- Model-based Approaches
  - Build models that can be used to identify whether a test instance is anomalous or not.
  - Build a model of the normal class and identify anomalies that do not fit this model.
  - 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 the most of the points in the data are normal

General Issues: Label vs Score

- Some anomaly detection techniques provide only a binary categorization

- 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)

Anomaly Detection Techniques

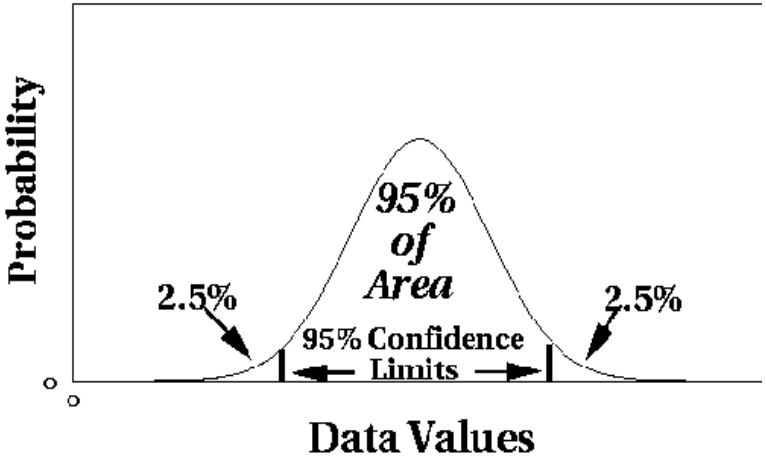
- Statistical Approaches
- Proximity-based
  - Anomalies are points far away from other points
- Clustering-based
  - Points far away from cluster centers are outliers
  - Small clusters are outliers
- Reconstruction Based

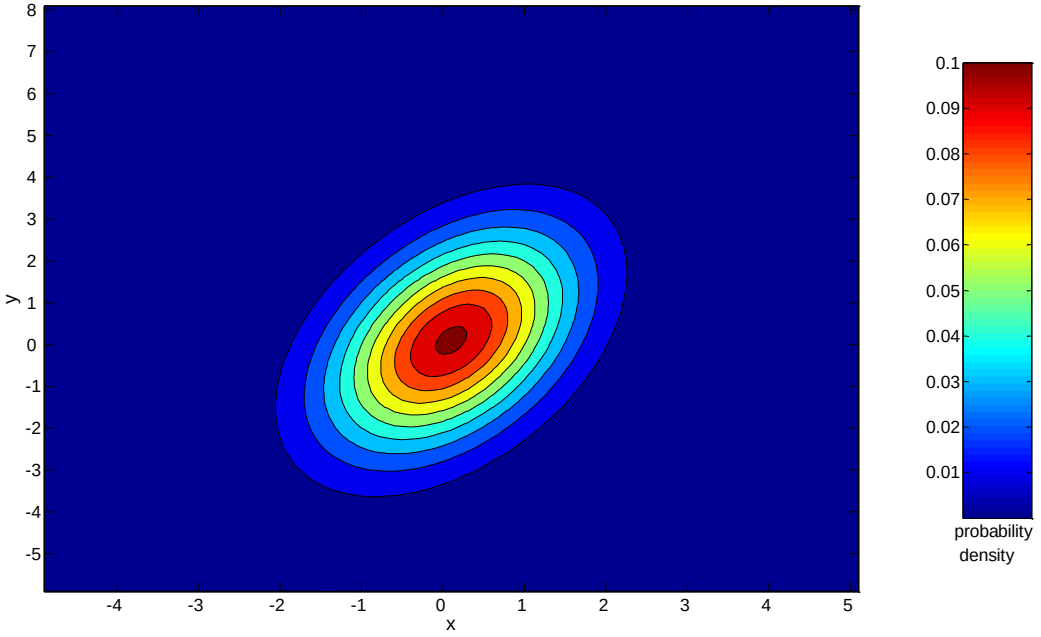
Statistical Approaches

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?

Normal Distributions





One-dimensional Gaussian

The Gaussian distribution has two parameters, μ and σ, which are the mean and standard deviation, respectively, and is represented using the notation $N(\mu,\sigma)$.

Two-dimensional Gaussian

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:

$$G = \frac{\max\left|X - \overline{X}\right|}{s}$$

  - s: standard deviation, $\overline{X}$: mean

- Reject H_0 if:

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

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

Statistically-based – Likelihood Approach

- Data distribution, D = (1 – λ) M + λ A

  where λ is a number between 0 and 1 that gives the expected fraction of outliers.

- 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
- The likelihood and log likelihood of the entire data set D at time t:

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

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

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

Likelihood-based outlier detection

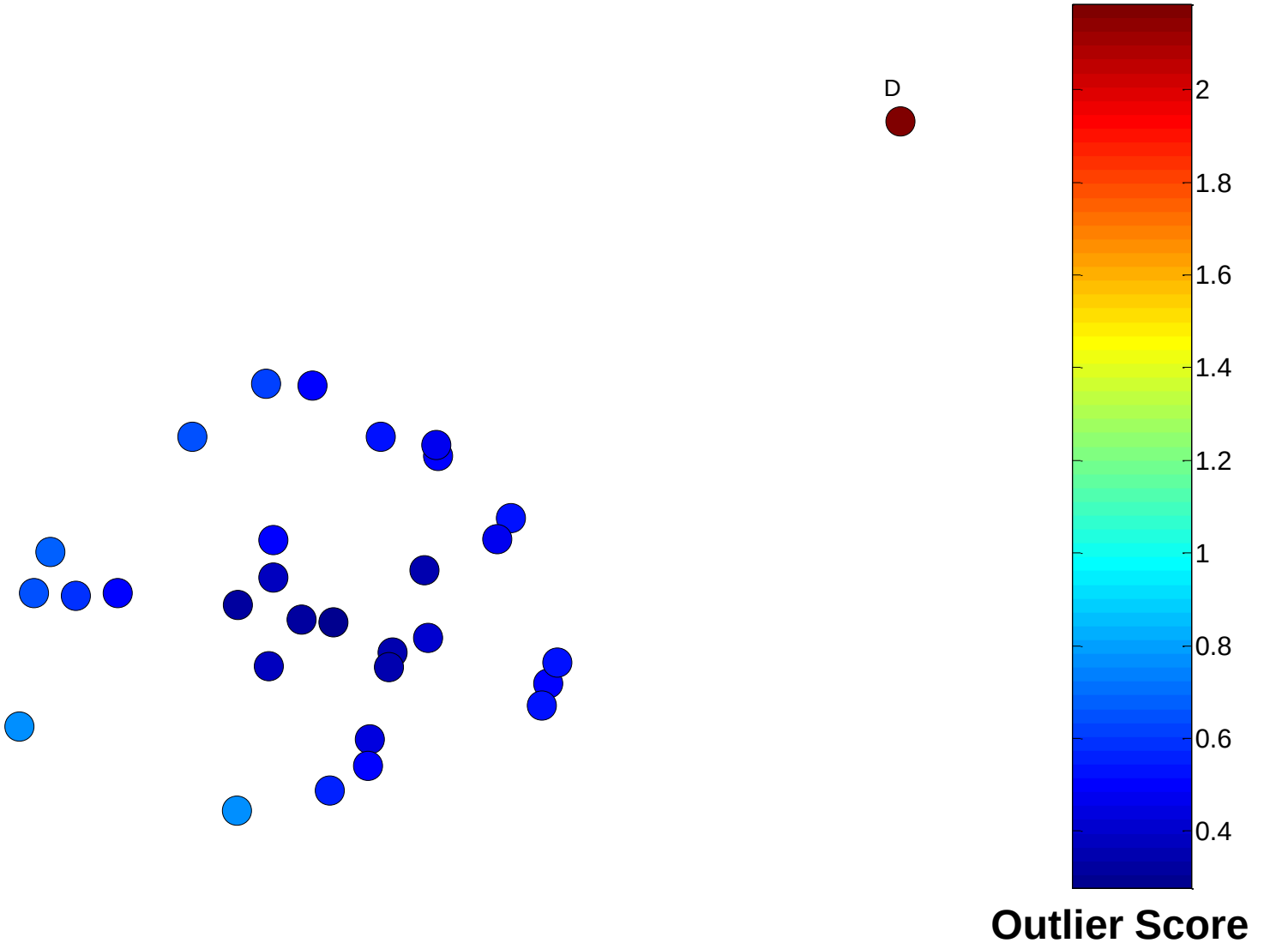
Algorithm 10.1 Likelihood-based outlier detection.

1: Initialization: At time $t = 0$, let M_t contain all the objects, while A_t is empty. Let $LL_t(D) = LL(M_t) + LL(A_t)$ be the log likelihood of all the data.
2: **for** each point **x** that belongs to M_t **do**
3: Move **x** from M_t to A_t to produce the new data sets A_{t+1} and M_{t+1}.
4: Compute the new log likelihood of D, $LL_{t+1}(D) = LL(M_{t+1}) + LL(A_{t+1})$
5: Compute the difference, $\Delta = LL_t(D) - LL_{t+1}(D)$
6: **if** $\Delta > c$, where c is some threshold **then**
7: **x** is classified as an anomaly, i.e., M_{t+1} and A_{t+1} are left unchanged and become the current normal and anomaly sets.
8: **end if**
9: **end for**

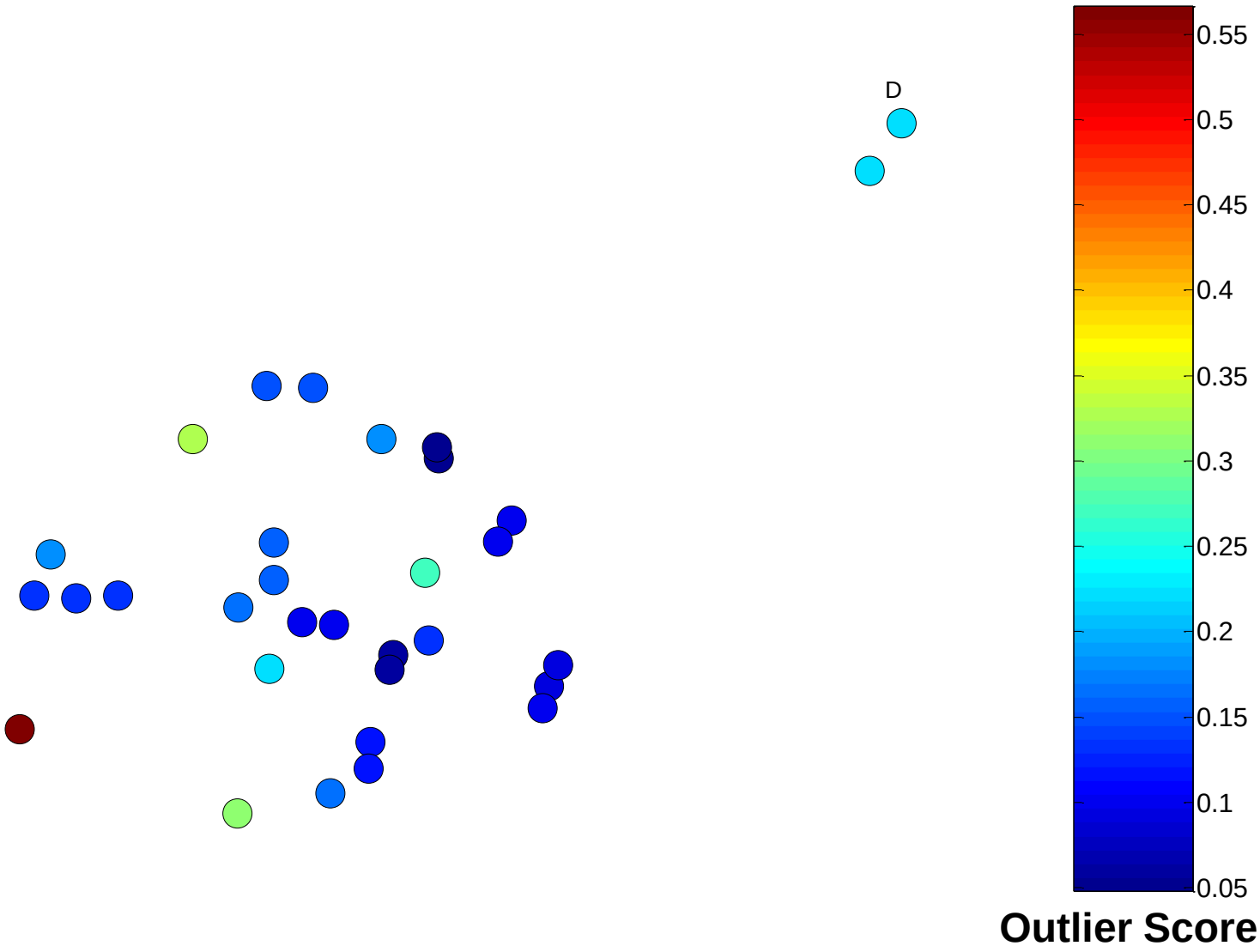
Distance-Based Approaches

- The outlier score of an object is the distance to its kth nearest neighbor
- The next figures show a set of two-dimensional points.
- The shading of each point indicates its outlier score using a value of k.
- The outlier score can be highly sensitive to the value of k. If k is too small, e.g., 1, then a small number of nearby outliers can cause a low outlier score.

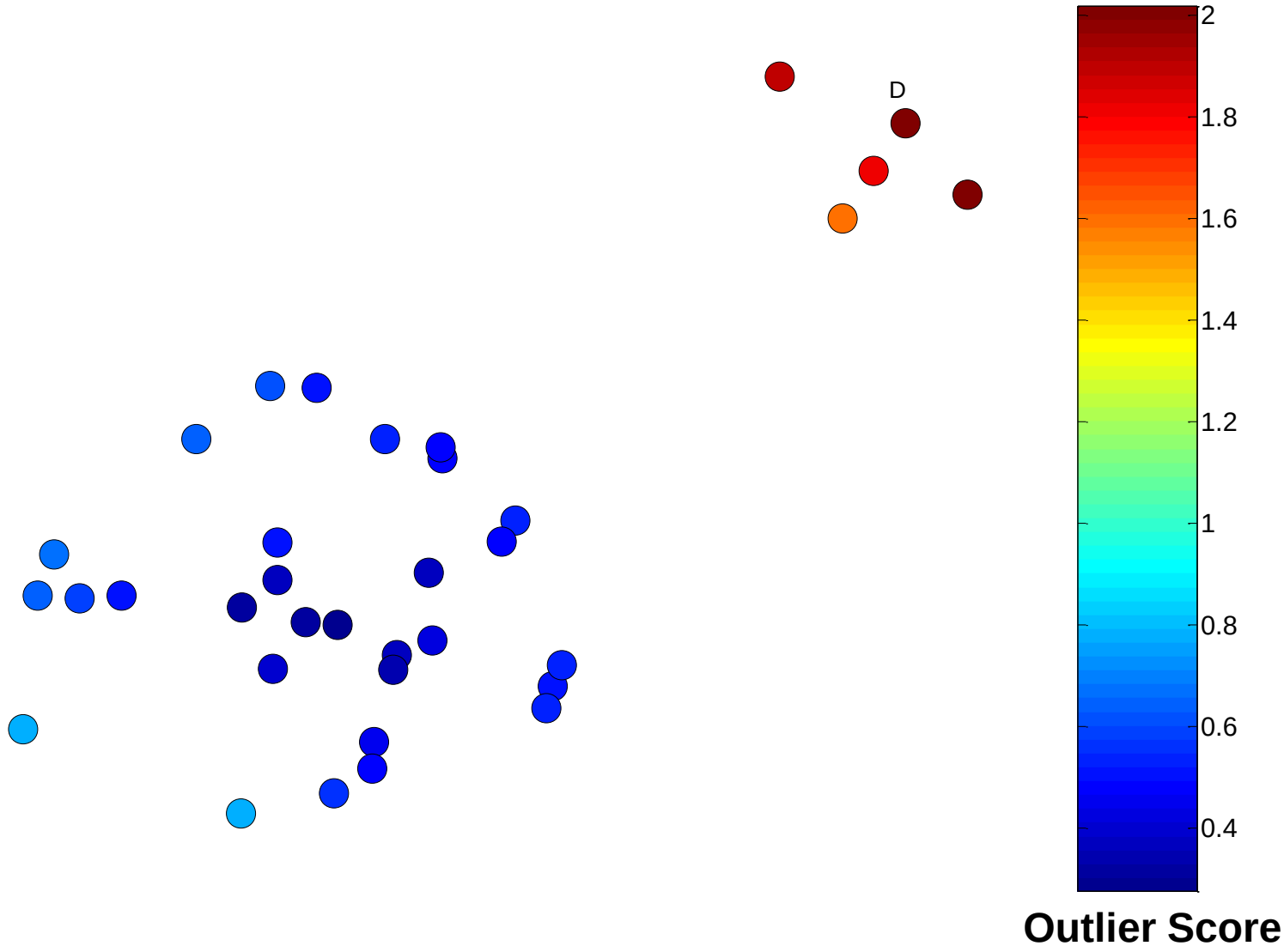
One Nearest Neighbor - One Outlier



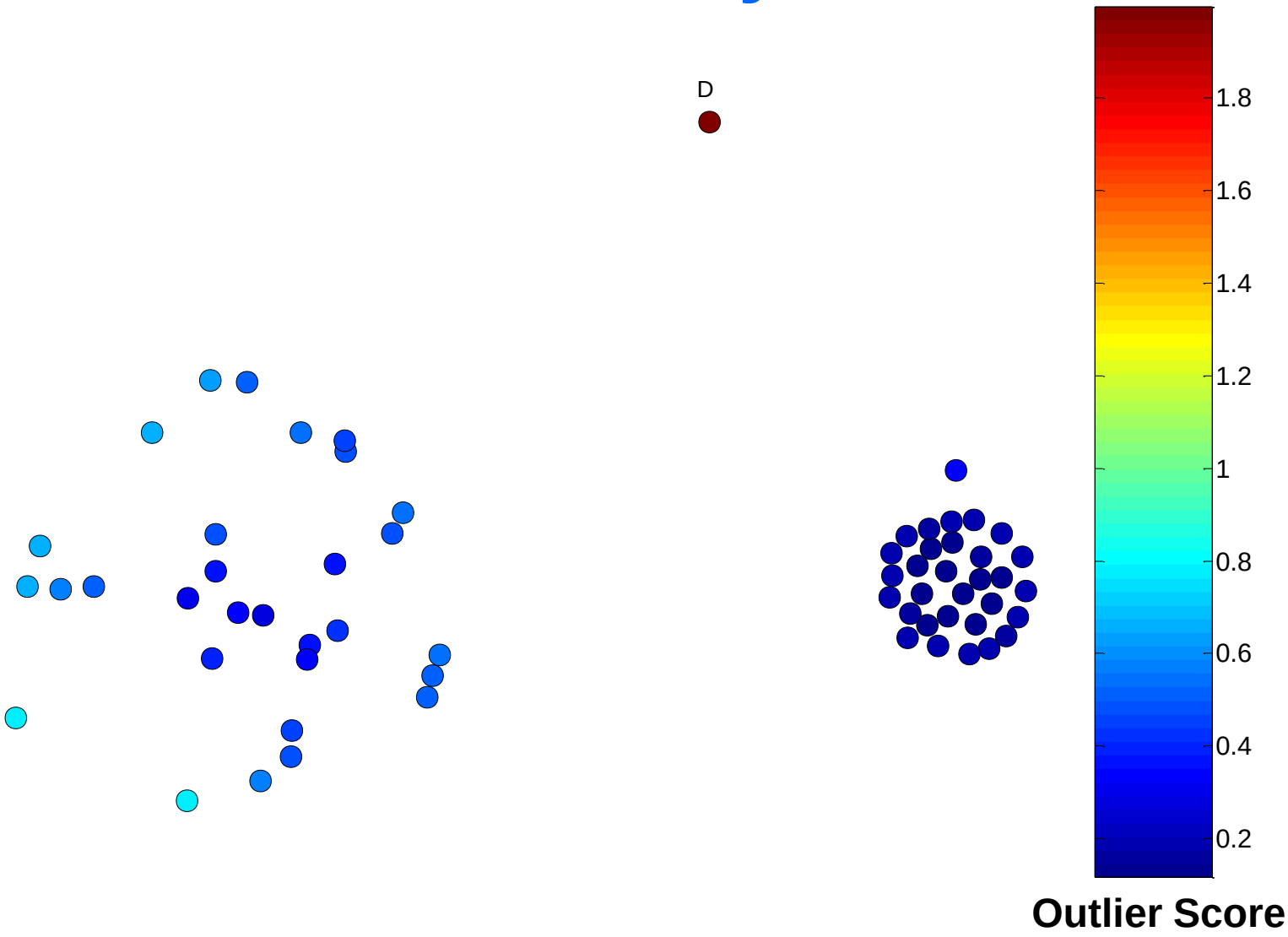
One Nearest Neighbor - Two Outliers



Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



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

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

- If there are regions of different density, this approach can have problems

Relative Density

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

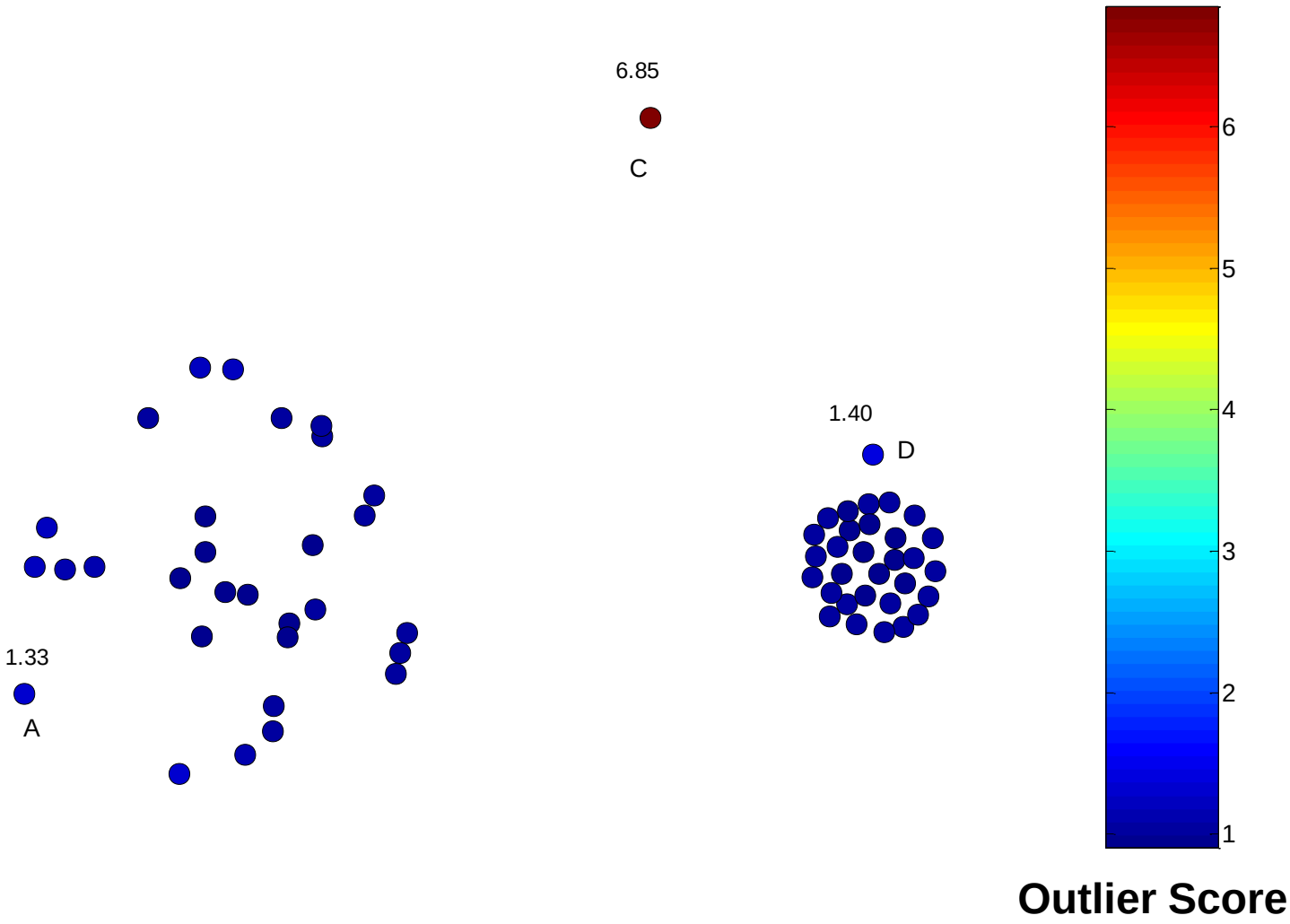
$$density(\boldsymbol{x}, k) = \frac{1}{dist(\boldsymbol{x}, k)} = \frac{1}{dist(\boldsymbol{x}, \boldsymbol{y}_k)}$$

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

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

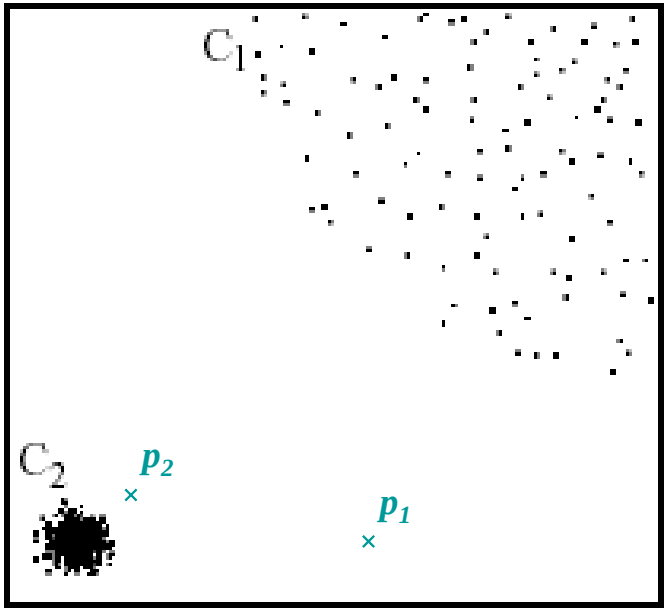
- Can use average distance instead

Relative Density Outlier Scores



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



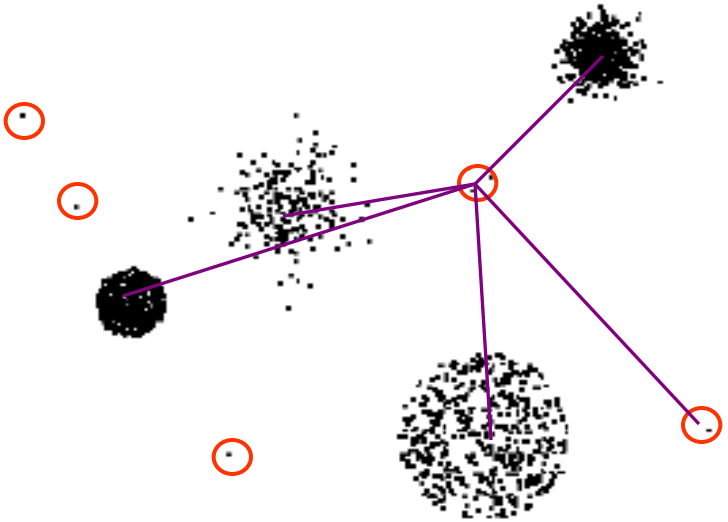
In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Strengths/Weaknesses of Density-Based Approaches

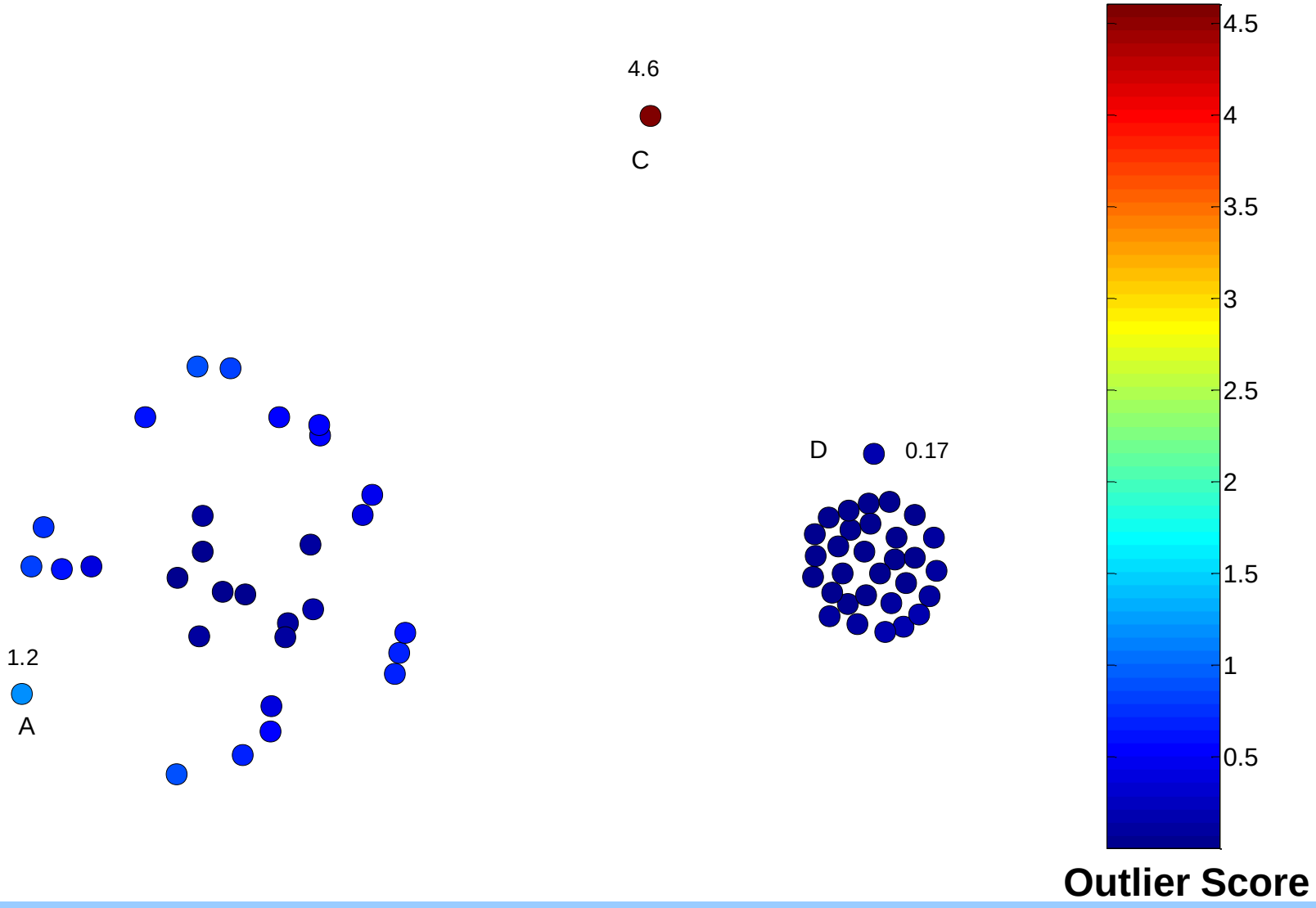
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

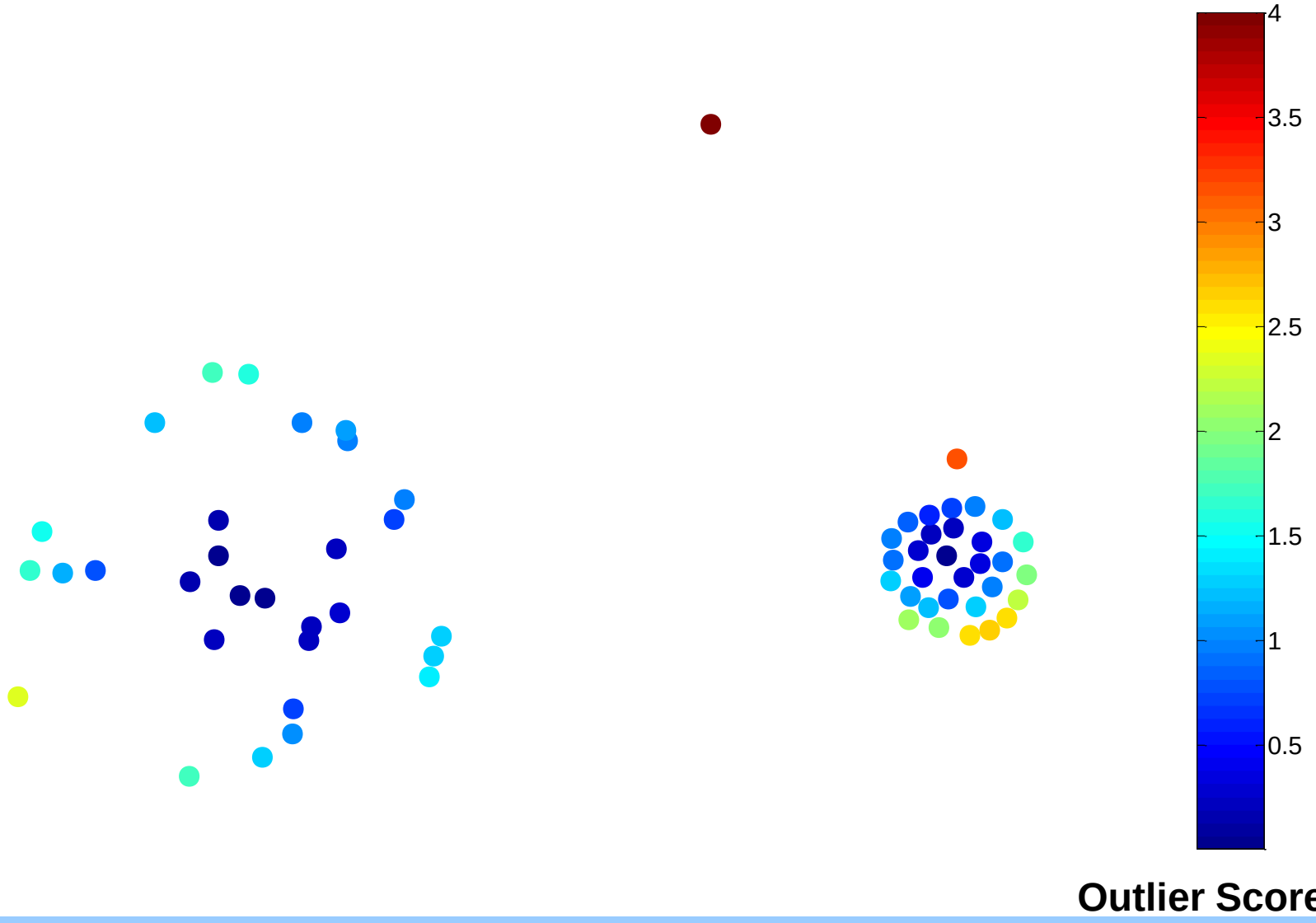
- Cluster analysis finds groups of strongly related objects, while anomaly detection finds objects that are not strongly related to other objects
- An object is a cluster-based outlier if it does not strongly belong to any cluster
  - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
    - Outliers can impact the clustering produced
  - For density-based clusters, an object 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



Distance of Points from Closest Centroids



Relative Distance of Points from Closest Centroid



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

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

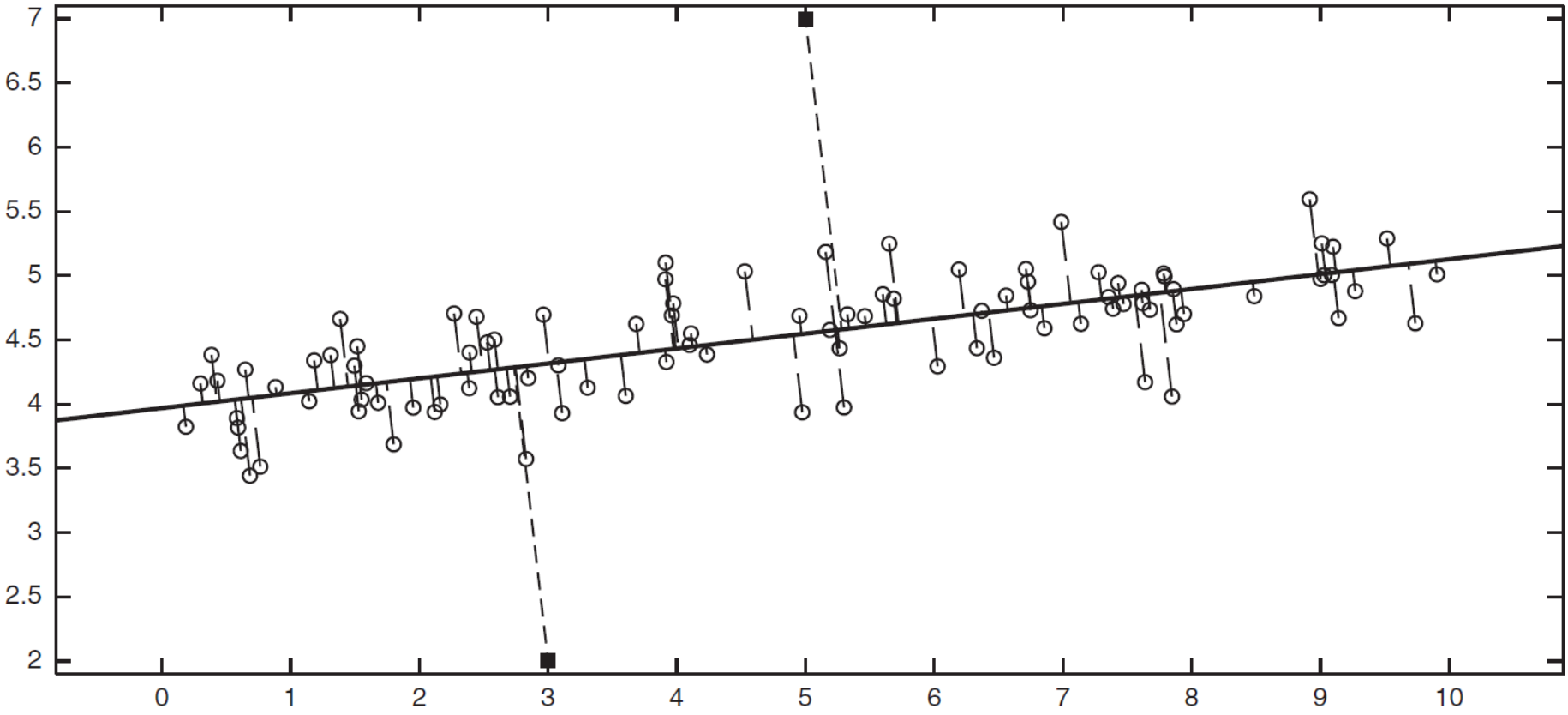
Reconstruction Error

- 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

Reconstruction of two-dimensional data



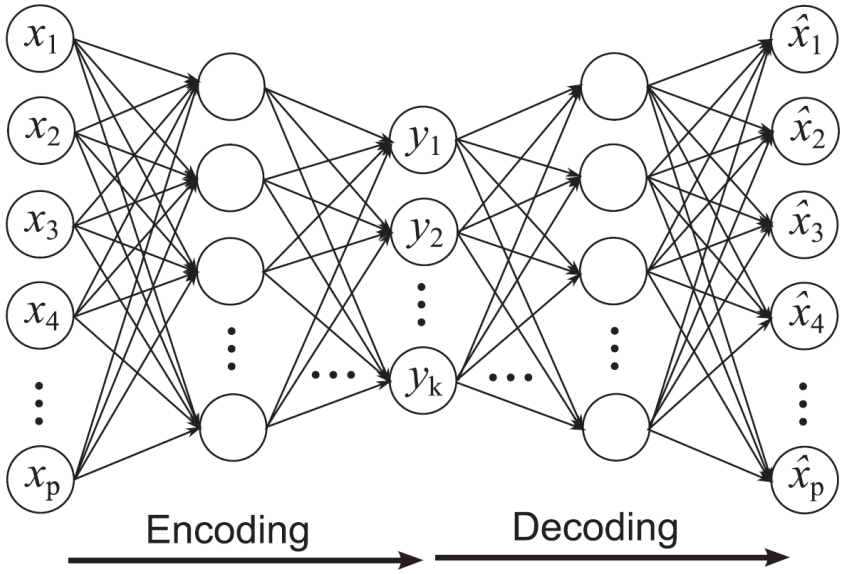
The black squares are anomalous instances. The solid black line shows the first principal component learned from this data, which corresponds to the direction of maximum variance of normal instances

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

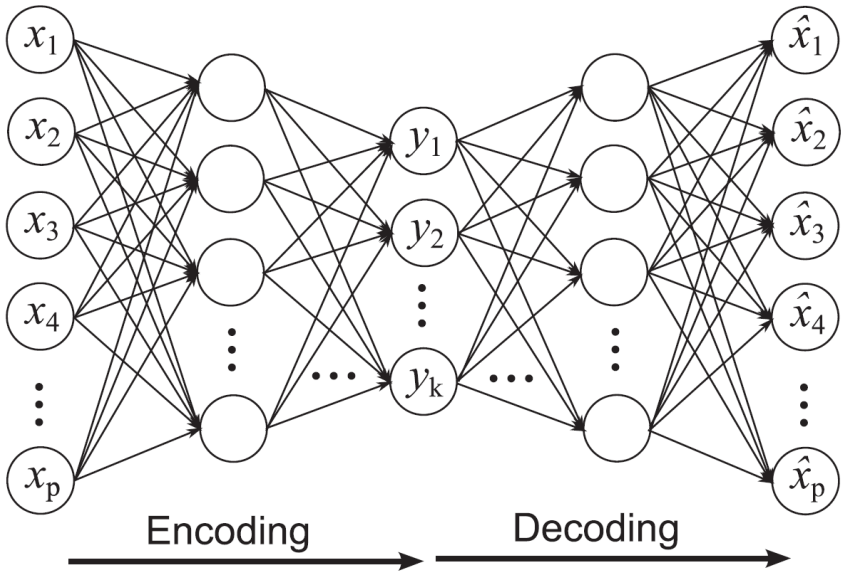
Basic Architecture of an Autoencoder

- An autoencoder is a multi-layer neural network.
- The number of input and output neurons is equal to the number of original attributes.
- An Autoencoder involves two basic steps, encoding and decoding.
- During encoding, a data instance x is transformed to a low dimensional representation y, using a number of nonlinear transformations in the encoding layers.
- The autoencoder scheme provides a powerful approach for learning complex and nonlinear representations of the normal class.
- In order to learn an autoencoder from an input data set comprising primarily of normal instances, we can use the backpropagation techniques introduced in the context of artificial neural networks.



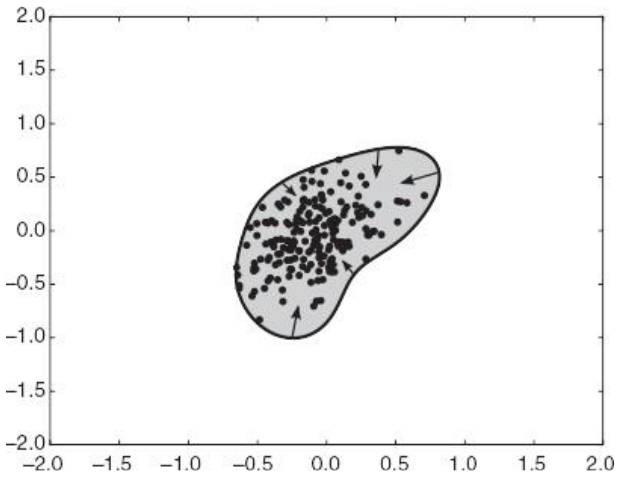
Basic Architecture of an Autoencoder

- Notice that the number of neurons reduces at every encoding layer, so as to learn low-dimensional representations from the original data.
- The learned representation y is then mapped back to the original space of attributes using the decoding layers, resulting in a reconstruction of x, denoted by $\hat{x}$.
- The distance between x and $\hat{x}$. (the reconstruction error) is then used as a measure of an anomaly score.



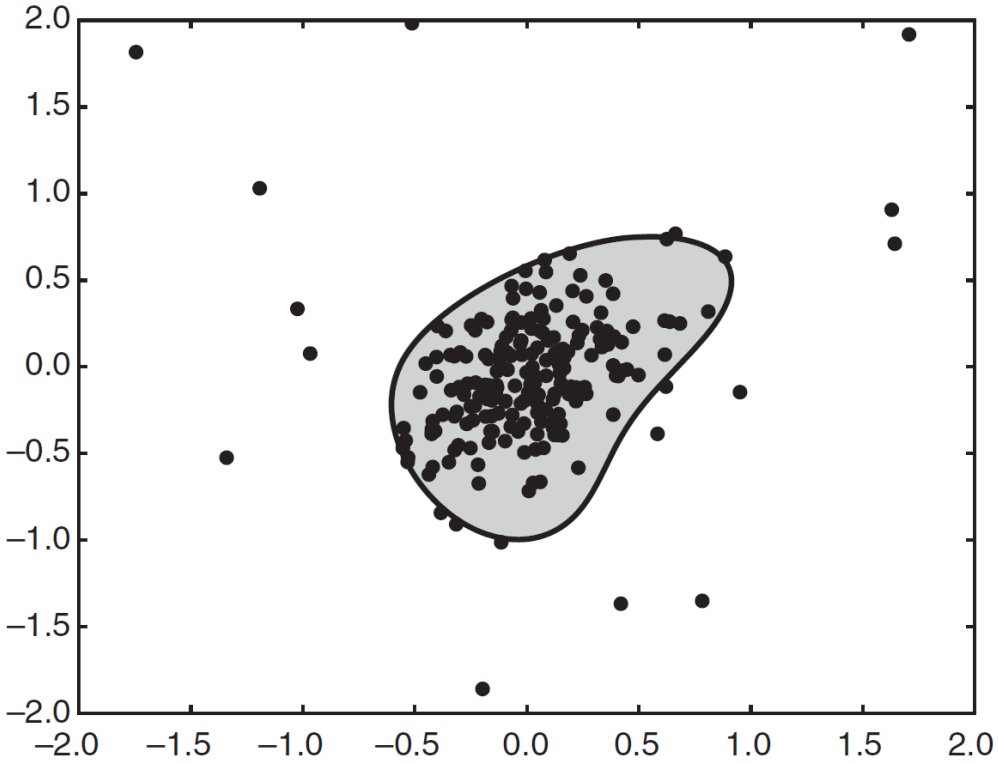
One Class SVM

- One-class classification approaches learn a decision boundary in the attribute space that encloses all normal objects on one side of the boundary.
- One-class SVM uses an SVM approach to classify normal objects
- It only uses training instances from the normal class to learn its decision boundary.



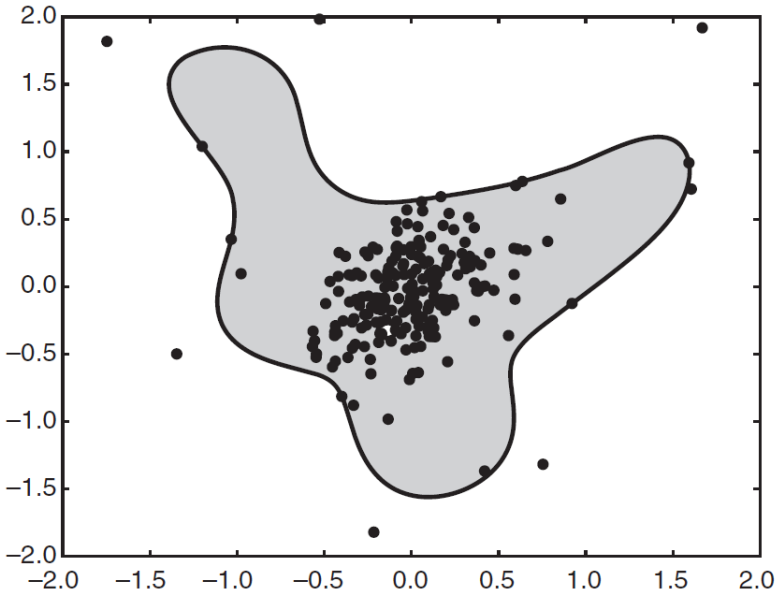
Finding Outliers with a One-Class SVM

- The hyper-parameter v of one-class SVM has a special interpretation. It represents an upper bound on the fraction of training instances that can be tolerated as anomalies while learning the hyperplane
- Decision boundary with $\nu = 0.1$

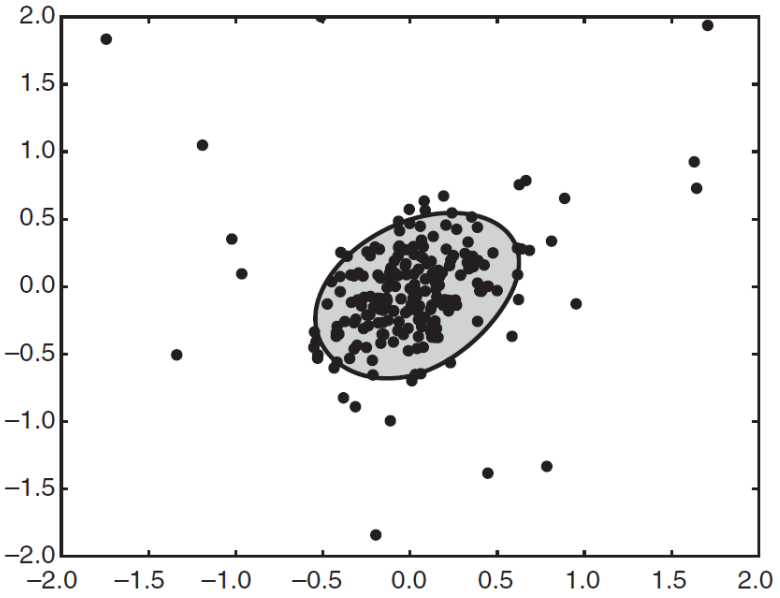


Finding Outliers with a One-Class SVM

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



(a) $\nu = 0.05$.



(b) $\nu = 0.2$.

Strengths and Weaknesses

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

Information Theoretic Approaches

- The focus of information theoretic approaches is to quantify the amount of information required for encoding them. If the normal class shows some structure or pattern, we can expect to encode it using a small number of bits.
- Anomalies can then be identified as instances that introduce irregularities in the data, which increase the overall information content of the data set.
- There are a number of approaches for quantifying the information content (also referred to as complexity) of a data set like:
  - Entropy measure
  - Kolmogorov complexity

Information Theoretic Approaches

- Let us denote the information content of a data set D as Info(D). Consider computing the anomaly score of a data instance x in D.
- If we remove x from D, we can measure the information content of the remaining data as Info(D \ x).
- Key idea is to measure how much information decreases when you delete an observation

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

- This happens because anomalies are expected to be surprising, and thus, their elimination should result in a substantial reduction in the information content. We can thus use as a measure of anomaly score.

- Anomalies should show higher gain

- Normal points should have less gain

Information Theoretic Example

- Survey of height and weight for 100 participants. The weight and height information of 100 participants, which has an entropy of 2.08.

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

- We can see that there is a pattern in the height and weight distribution of normal participants, since most participants that have a high value of weight also have a high value of height and vice-versa. However, there are 5 participants that have a high weight value but low height value, which is quite unusual.

- By eliminating these 5 instances, the entropy of the resulting data set becomes 1.89, resulting in a gain of

$$2.08 - 1.89 = 0.19$$

Strengths and Weaknesses

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

Evaluation of Anomaly Detection

- When class labels are available to distinguish between anomalies and normal data, then the effectiveness of an anomaly detection scheme can be evaluated by using measures of classification performance.
- 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.

Distribution of Anomaly Scores

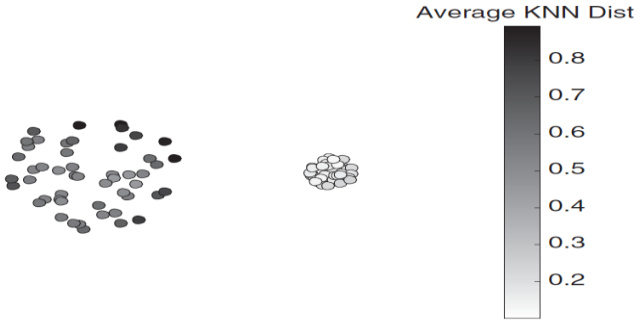


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

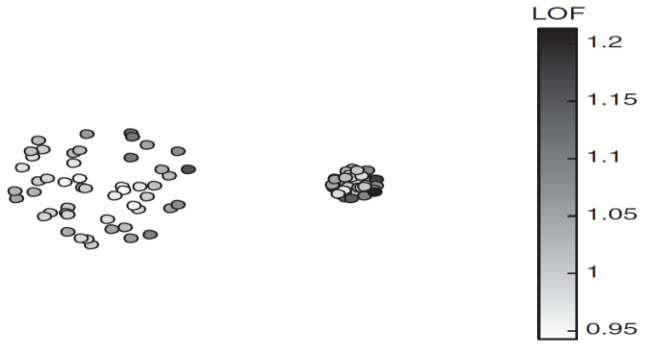
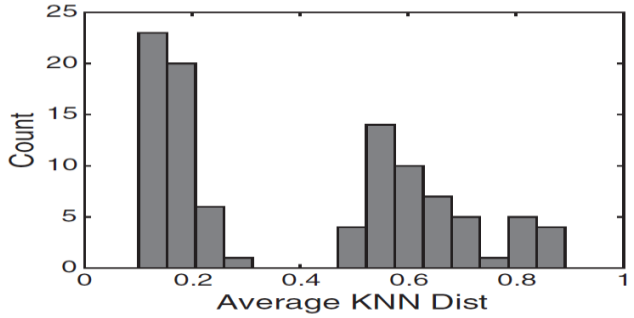
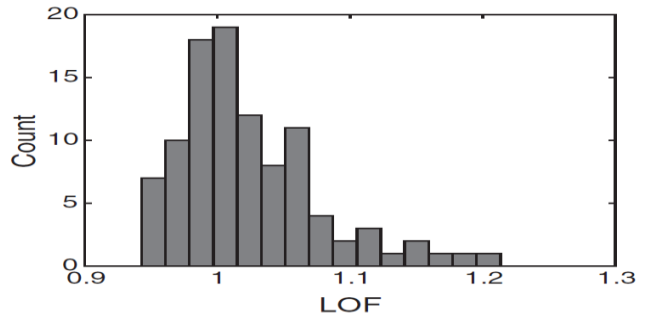


Figure 10.18. Anomaly score based on LOF using five nearest neighbors.





- By looking at the distribution of the scores via a histogram or density plot, we can assess whether the approach we are using generates scores that behave in a reasonable manner.
- The histogram of the average KNN dist shows a bimodal distribution.
- The key point is that the distribution of anomaly scores should look similar to that of the LOF scores in this example.
- There may be one or more secondary peaks in the distribution as one moves to the right, but these secondary peaks should only contain a relatively small fraction of the points, and not a large fraction of the points as with the average KNN dist approach.