

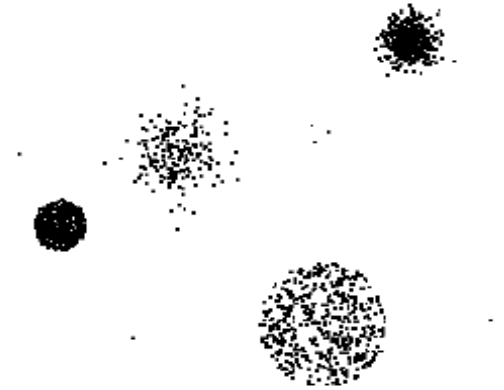
Anomaly Detection

INFO 523 – Lecture 9

Dr. Greg Chism

Anomaly/Outlier Detection

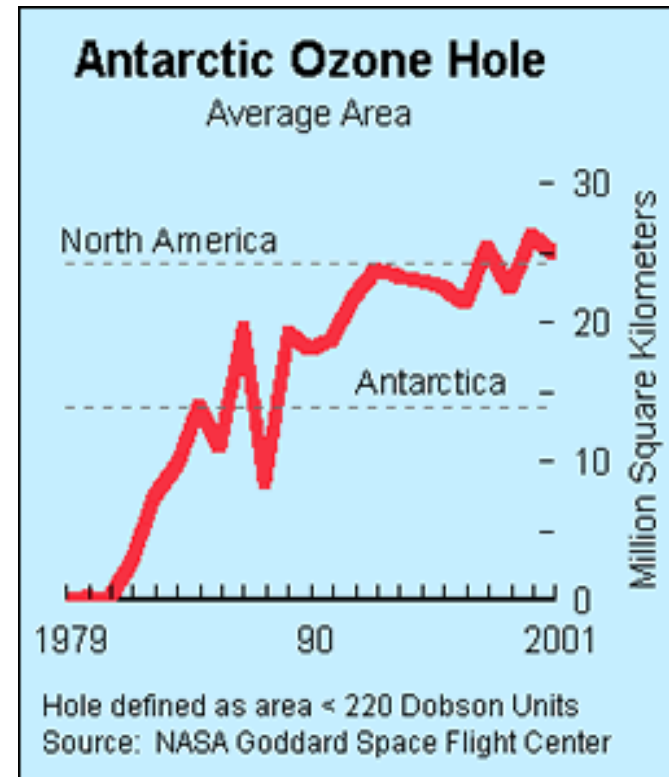
- What are anomalies/outliers?
 - The set of data points that are considerably different than the remainder of the data
- Natural implication is that anomalies are relatively rare
 - One in a thousand occurs often if you have lots of data
 - Context is important, e.g., freezing temps in July
- Can be 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!



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
 - 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

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

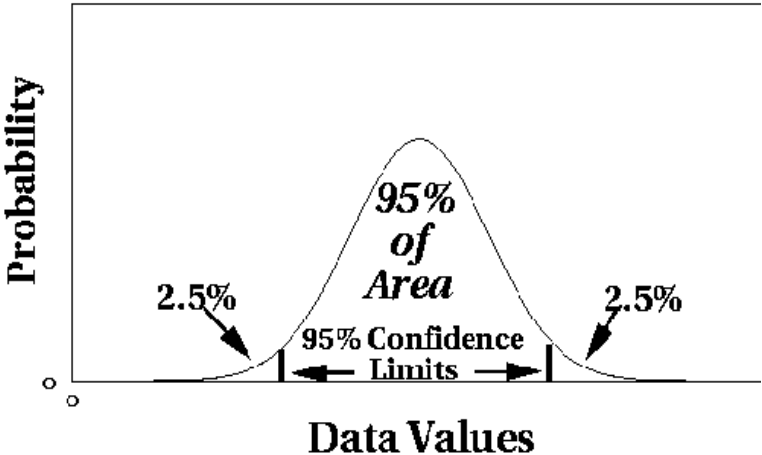
- Statistical Approaches
- Proximity-based
 - Anomalies are points far away from other points
- Clustering-based
 - Points far away from cluster centers are outliers
 - Small clusters are outliers
- Reconstruction Based

Statistical Approaches

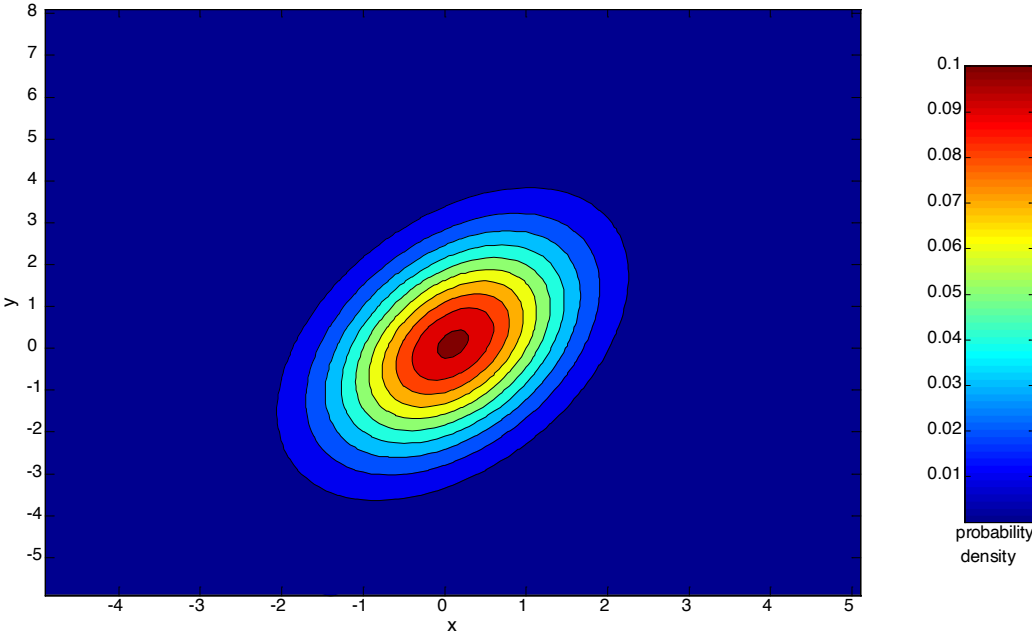
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

- 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**



**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:
$$G = \frac{\max |X - \bar{X}|}{s}$$

- Reject H_0 if:

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

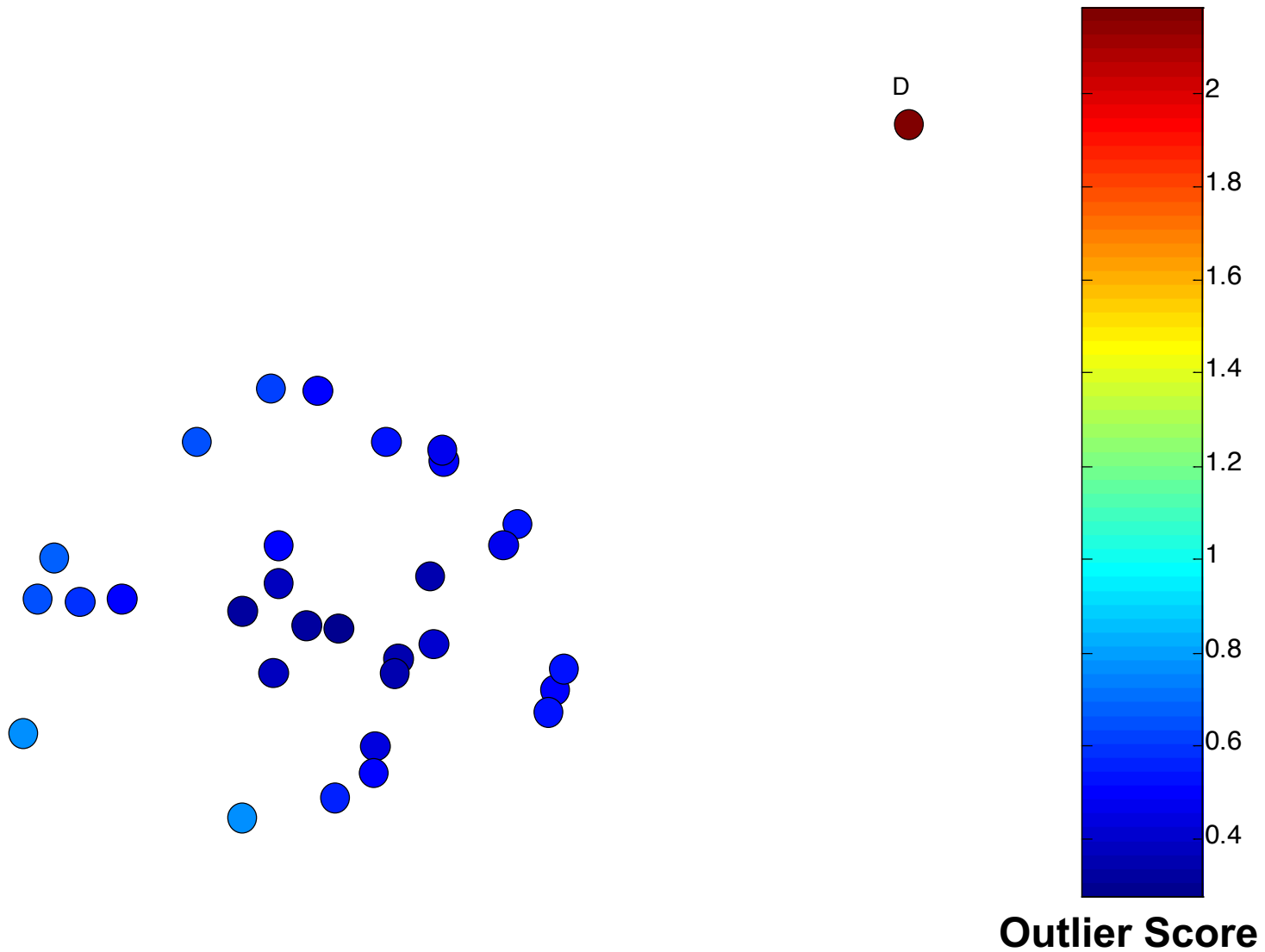
Strengths/Weaknesses of Statistical Approaches

- 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

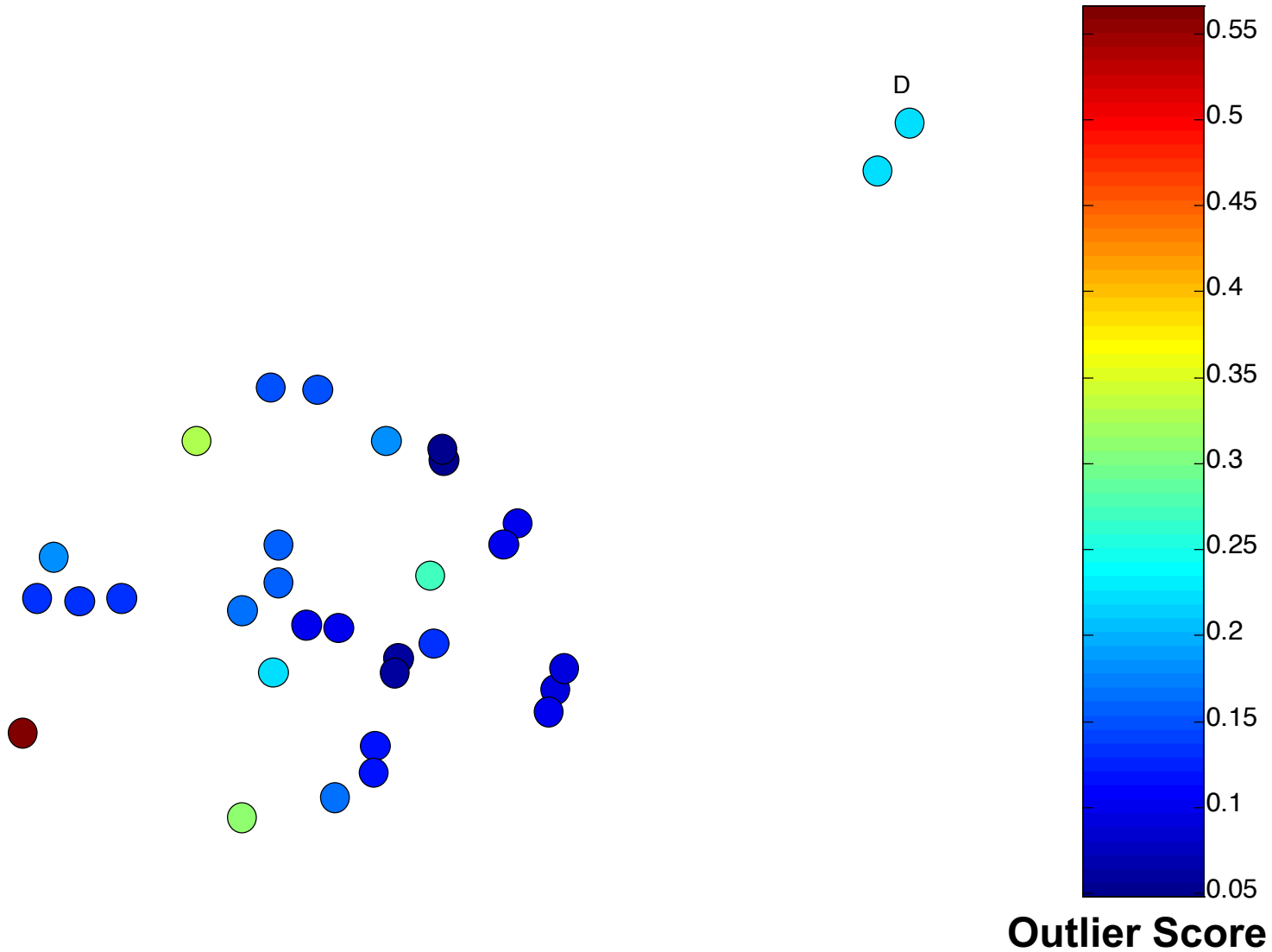
Distance-Based Approaches

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

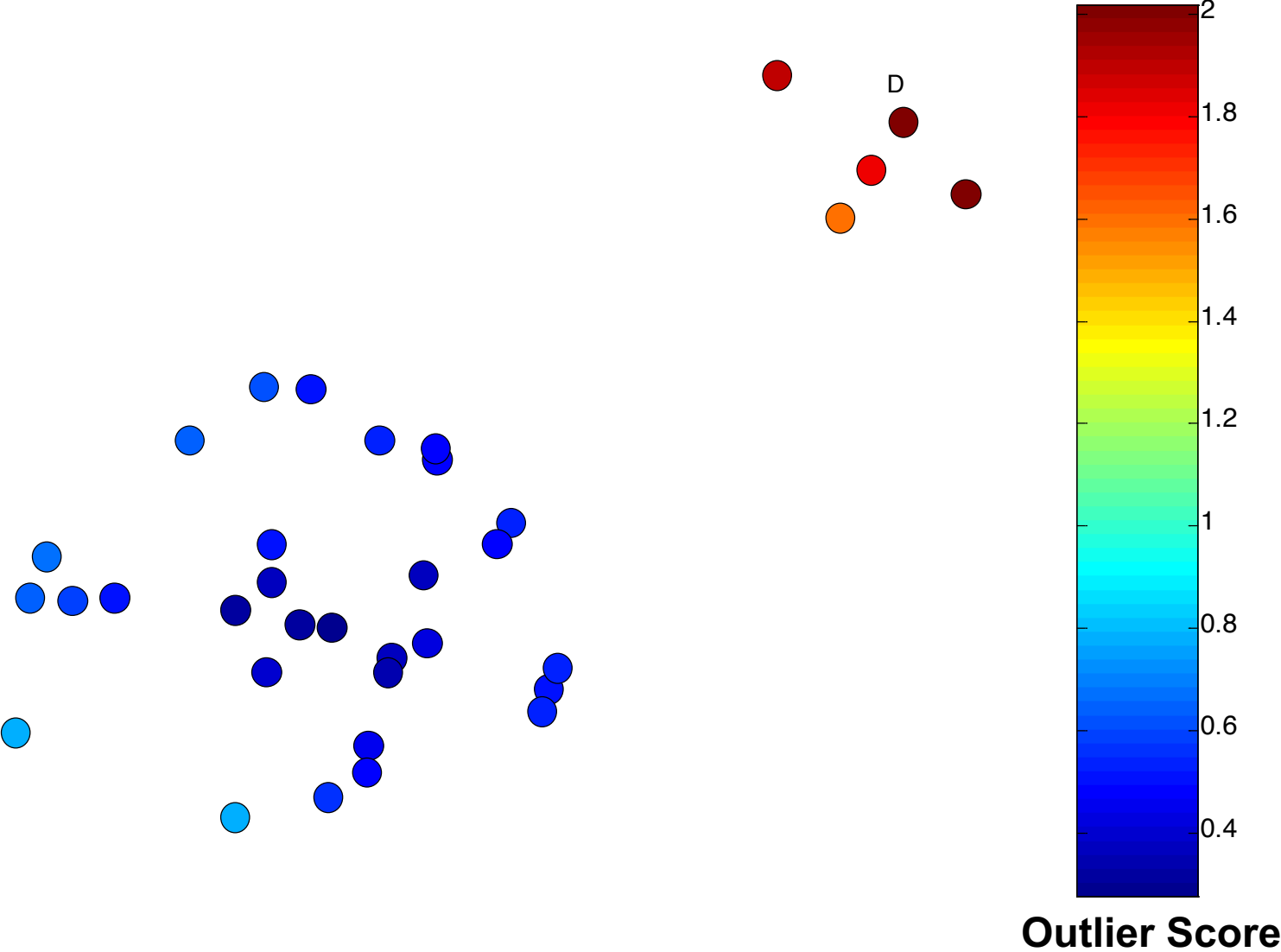
One Nearest Neighbor - One Outlier



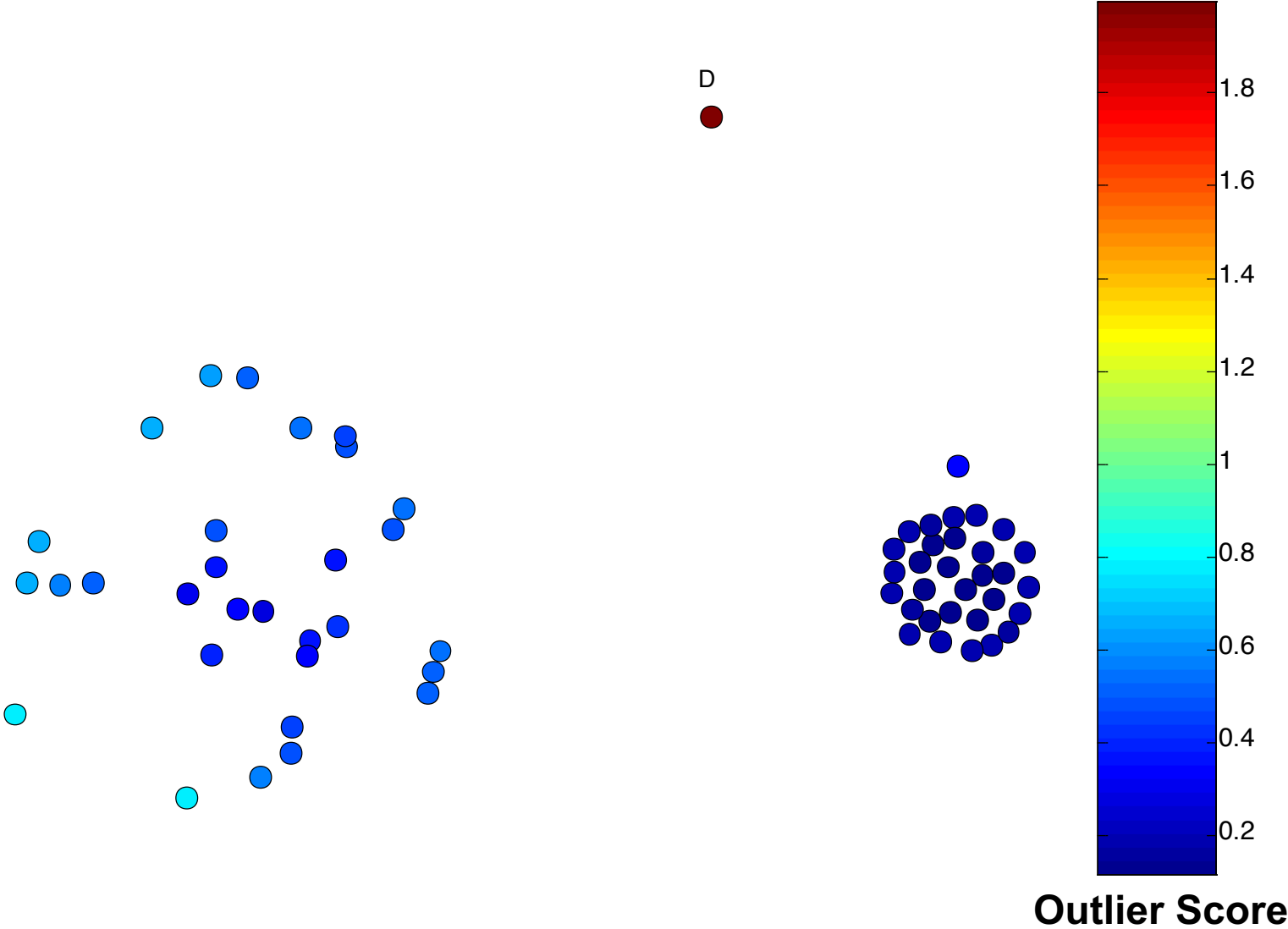
One Nearest Neighbor - Two Outliers



Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



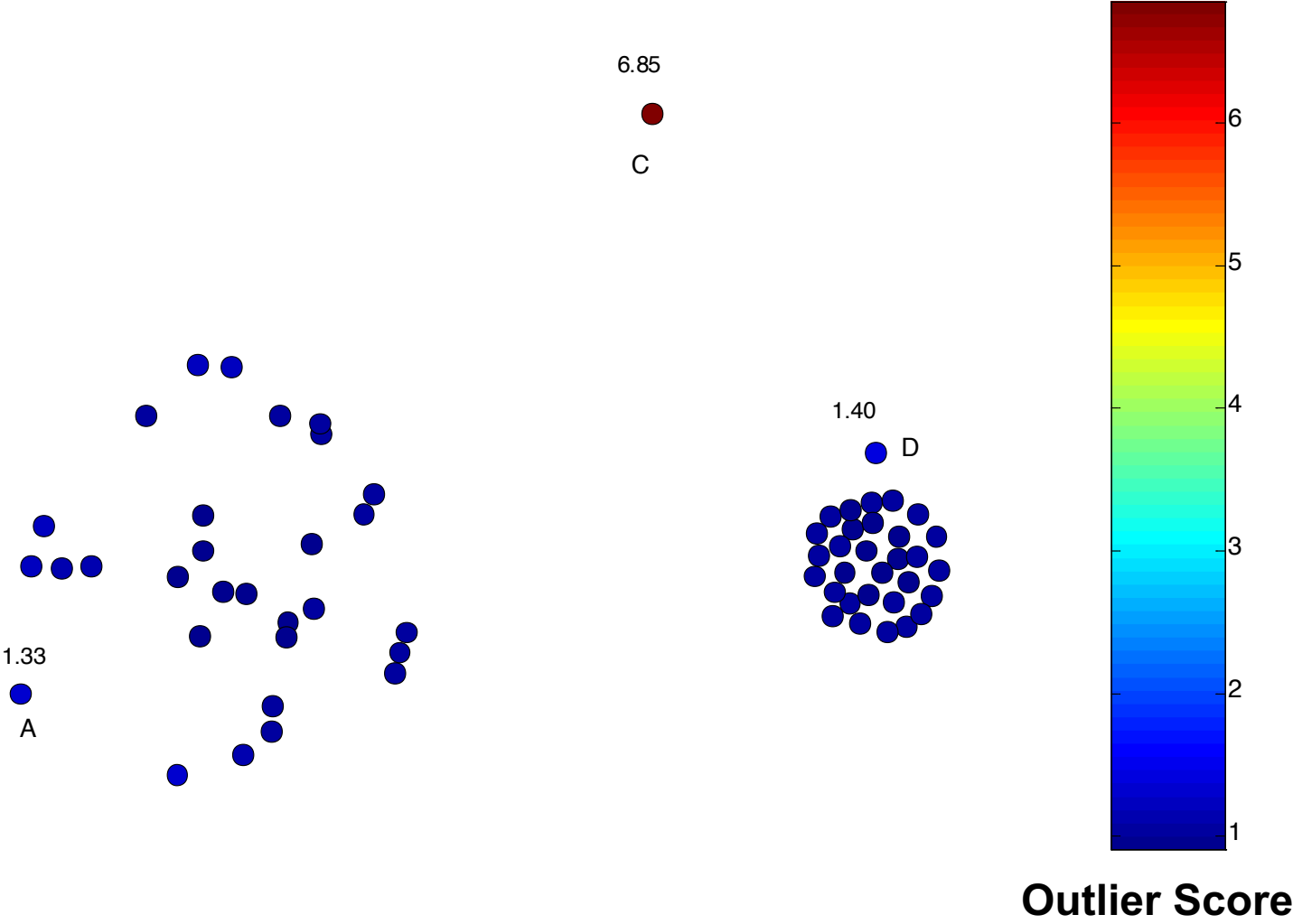
Strengths/Weaknesses of Distance-Based Approaches

- 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 k th 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 Outlier Scores

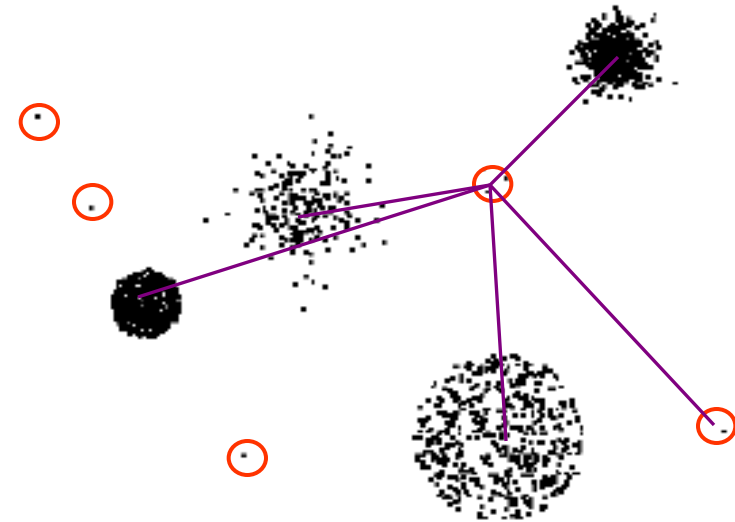


Strengths/Weaknesses of Density-Based Approaches

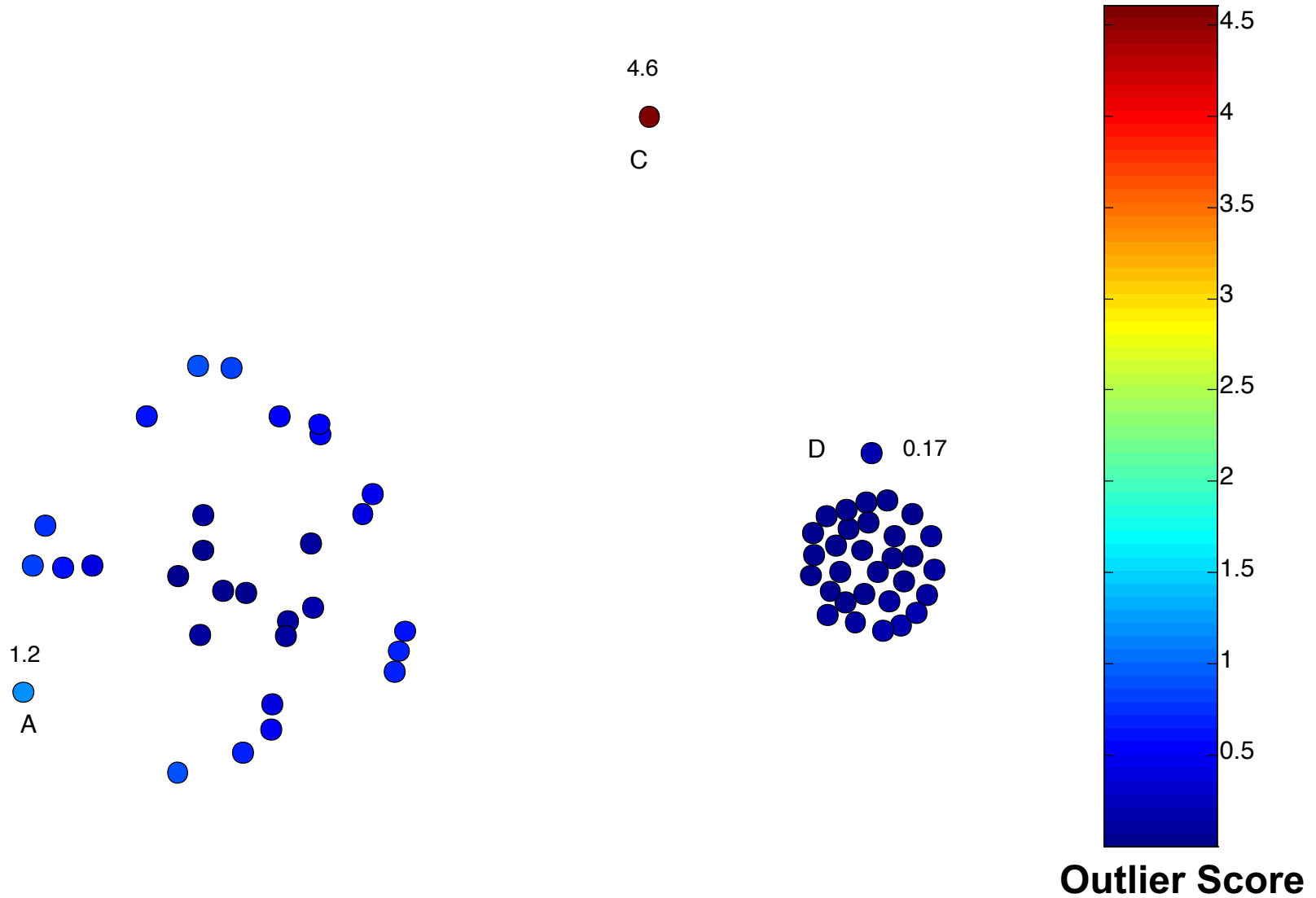
- Simple
- Expensive – $O(n^2)$
- Sensitive to parameters
- Density becomes less meaningful in high-dimensional space

Clustering-Based Approaches

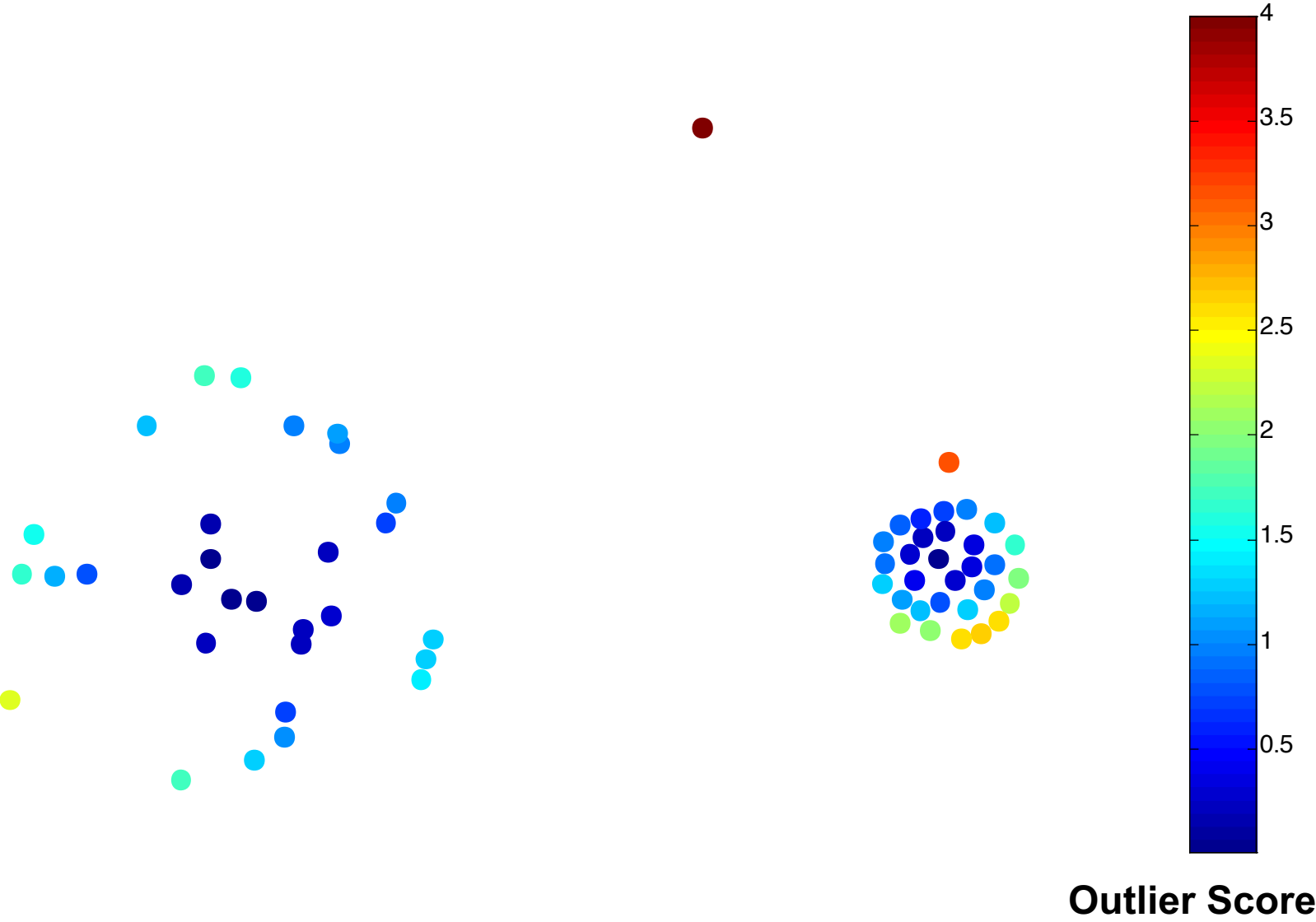
- An object is a cluster-based outlier if it does not strongly belong to any cluster
 - For prototype-based clusters, an object is an outlier if it is not close enough to a cluster center
 - Outliers can impact the clustering produced
 - For density-based clusters, an object is an outlier if its density 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

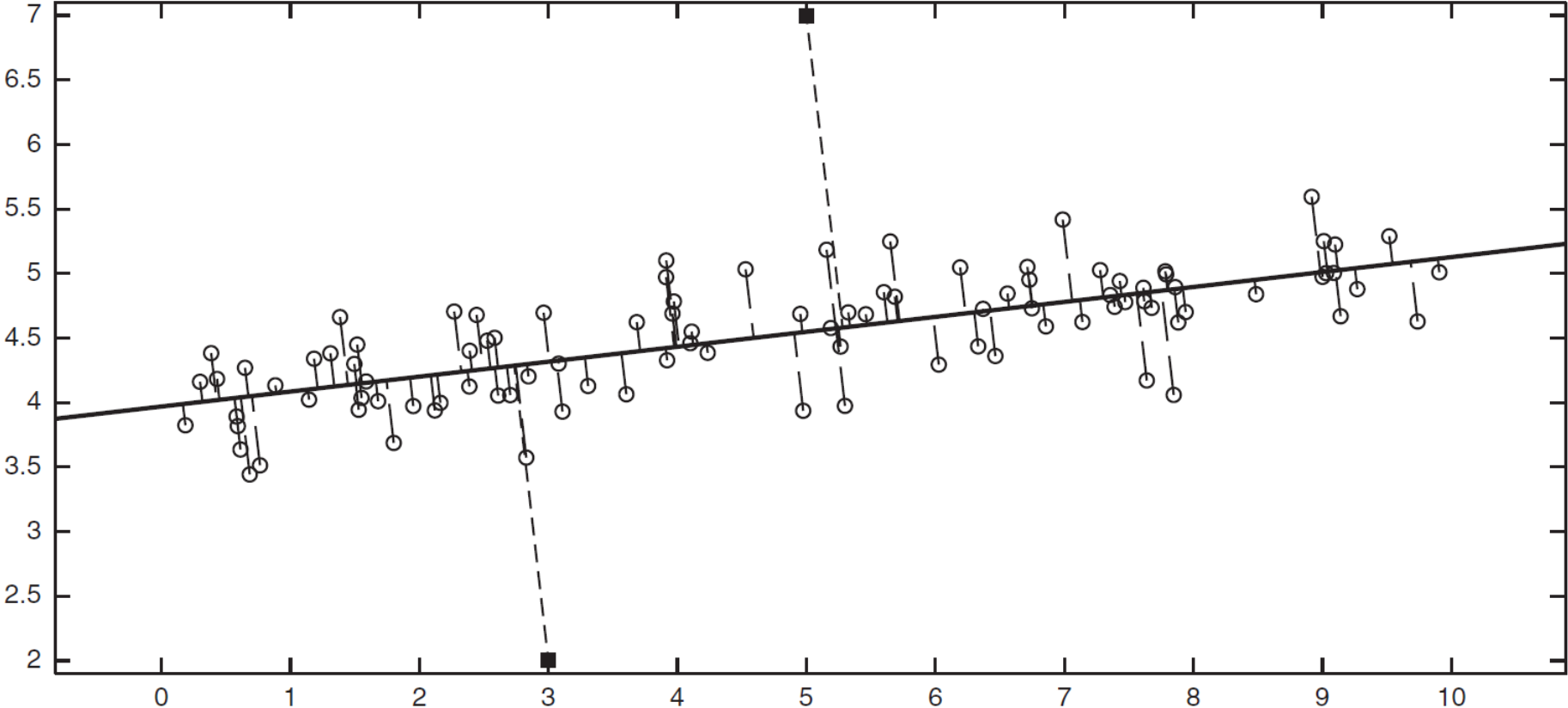
Reconstruction Error

- 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 $\hat{\mathbf{x}}$

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

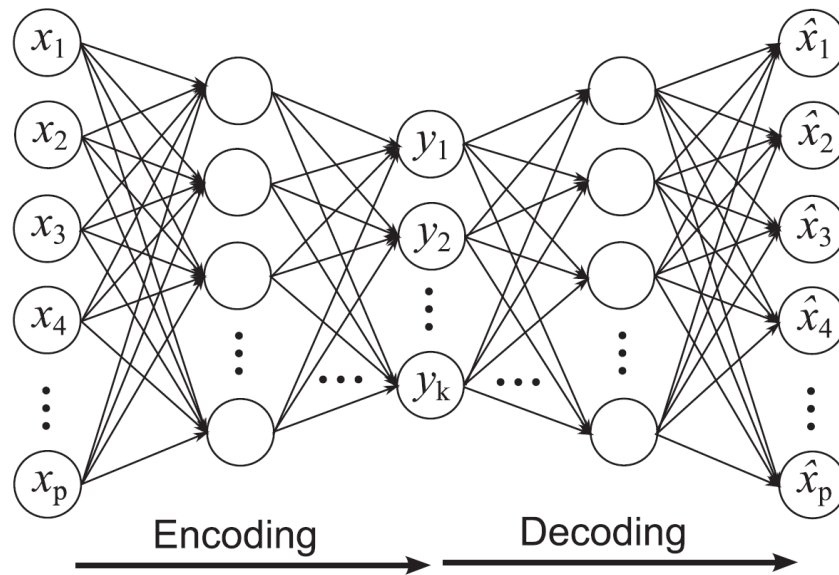
- Objects with large reconstruction errors are anomalies

Reconstruction of two-dimensional data



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.



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

One Class SVM

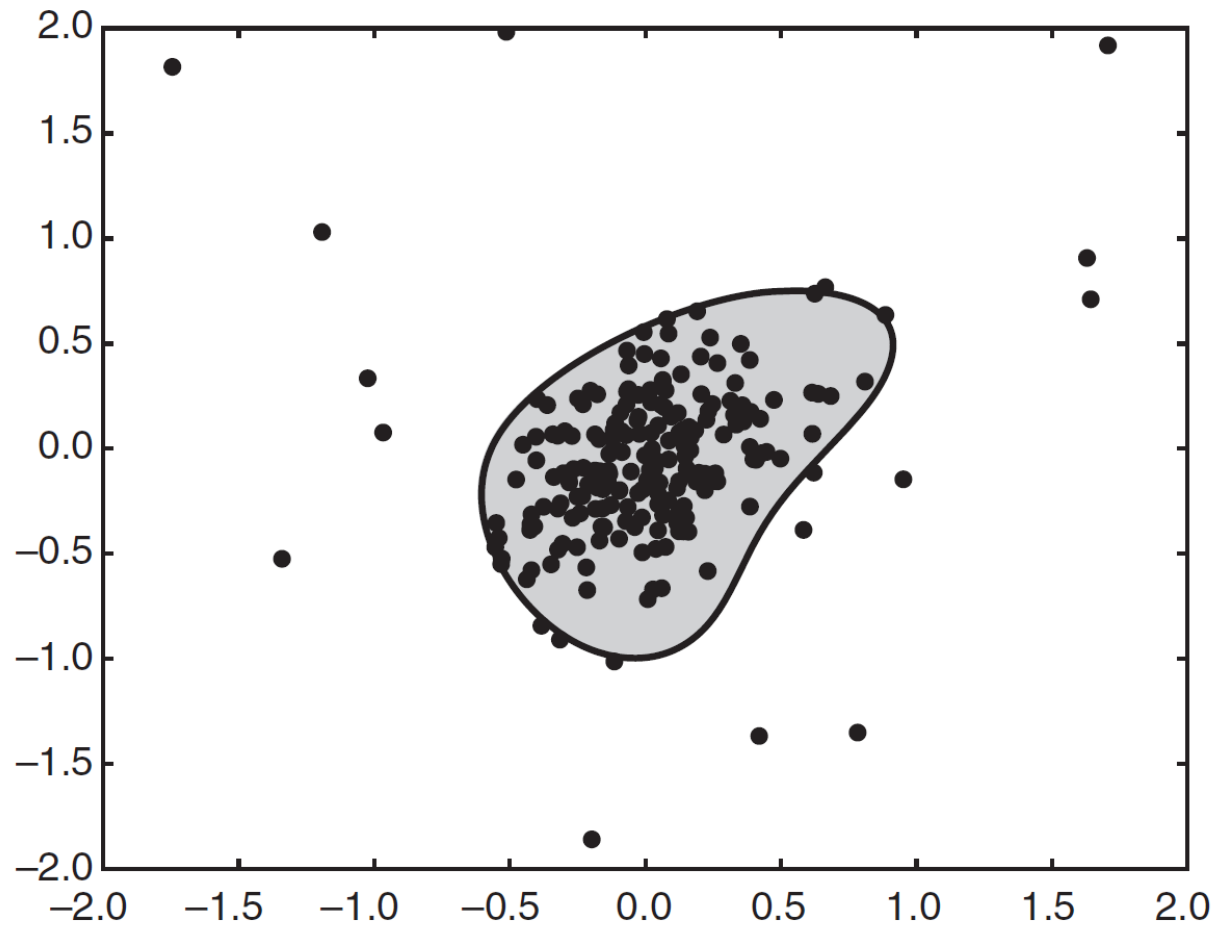
- Uses an SVM approach to classify normal objects
- Uses the given data to construct such a model
- This data may contain outliers
- But the data does not contain class labels
- How to build a classifier given one class?

How Does One-Class SVM Work?

- Uses the “origin” trick
- Use a Gaussian kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma^2}\right)$
 - 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

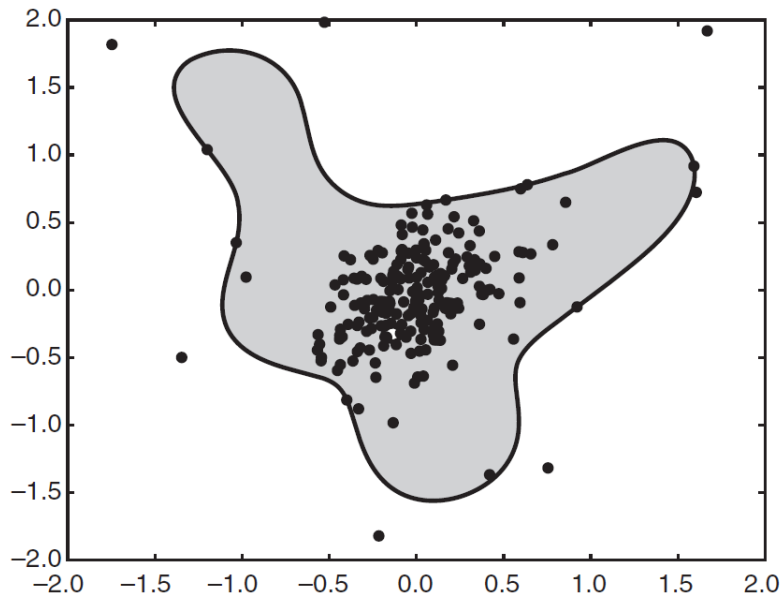
Finding Outliers with a One-Class SVM

- Decision boundary with $\nu = 0.1$

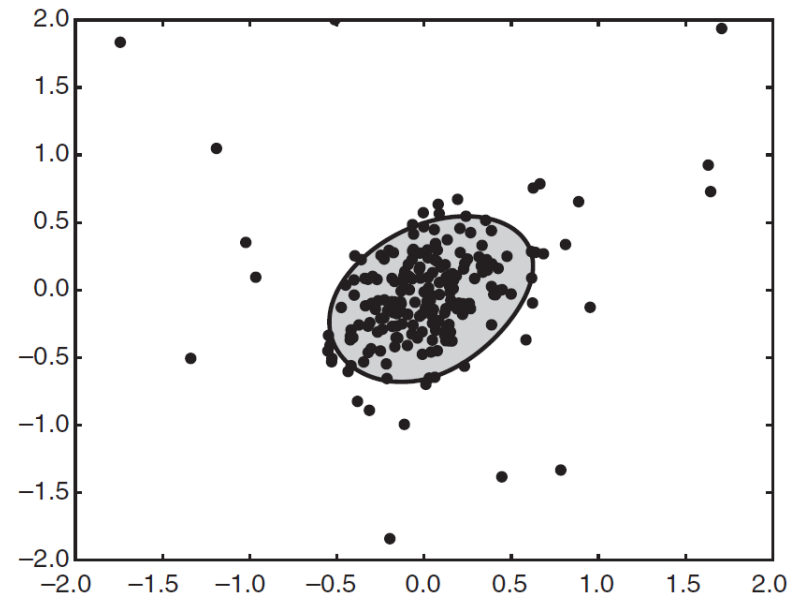


Finding Outliers with a One-Class SVM

- 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

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 known 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

- Anomaly scores should show a tail

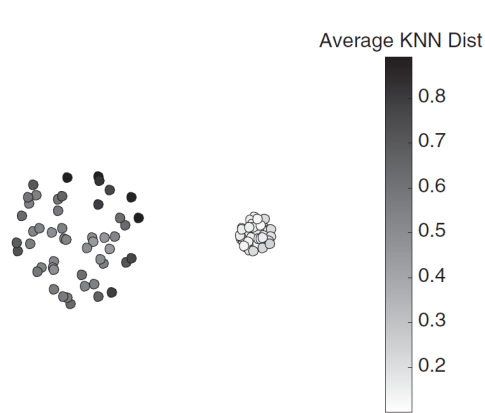


Figure 10.17. Anomaly score based on average distance to fifth nearest neighbor.

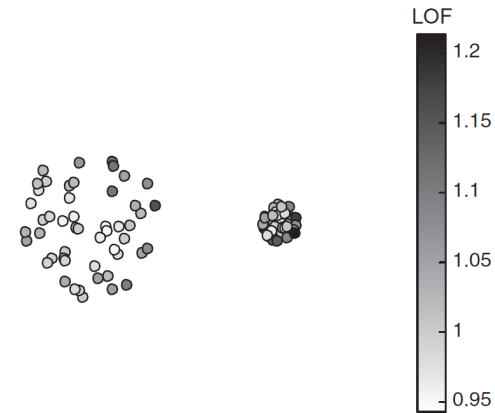
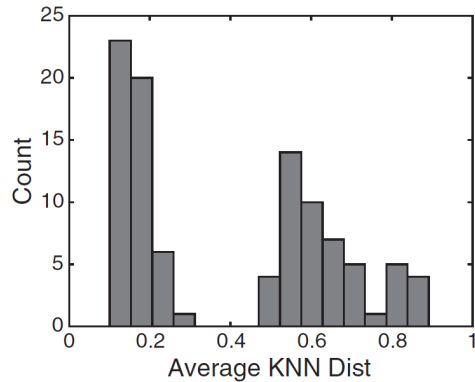


Figure 10.18. Anomaly score based on LOF using five nearest neighbors.

