Anomaly Detection

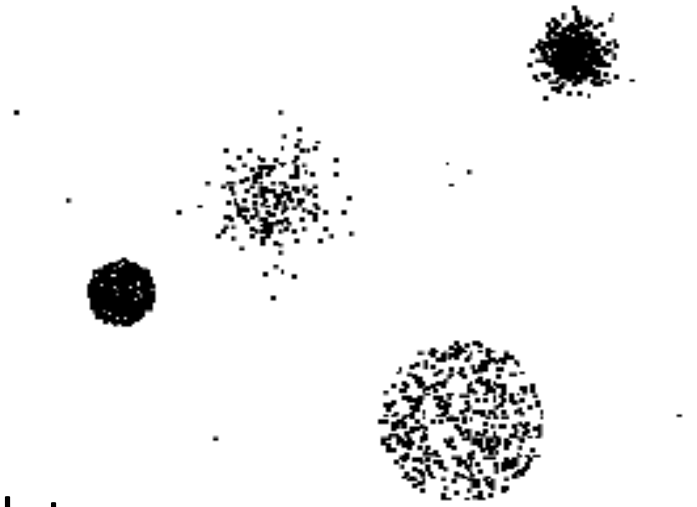
SED690 Selected Topic in Software Engineering for Data Science (Data Mining)

Instructor: Niwan Wattanakitrungroj

References: Jiawei Han, Jian Pei and Hanghang Tong, Data Mining: Concepts and Techniques, 4th ed. Morgan Kaufmann Publishers 2023
P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, Addison Wesley, 2nd ed., 2018.

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

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 the most of the points in the data are normal

Anomaly Detection Techniques

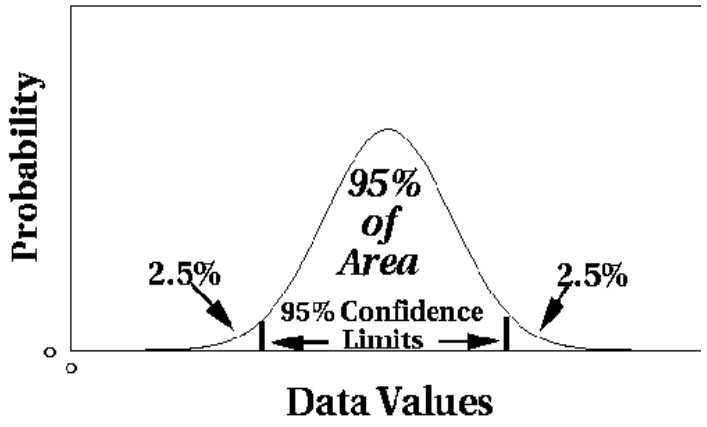
- Statistical Approaches
- Proximity-based
  - Anomalies are points far away from other points
- Clustering-based
  - Points far away from cluster centers are outliers
  - Small clusters are outliers
- One Class SVM

Statistical Approaches

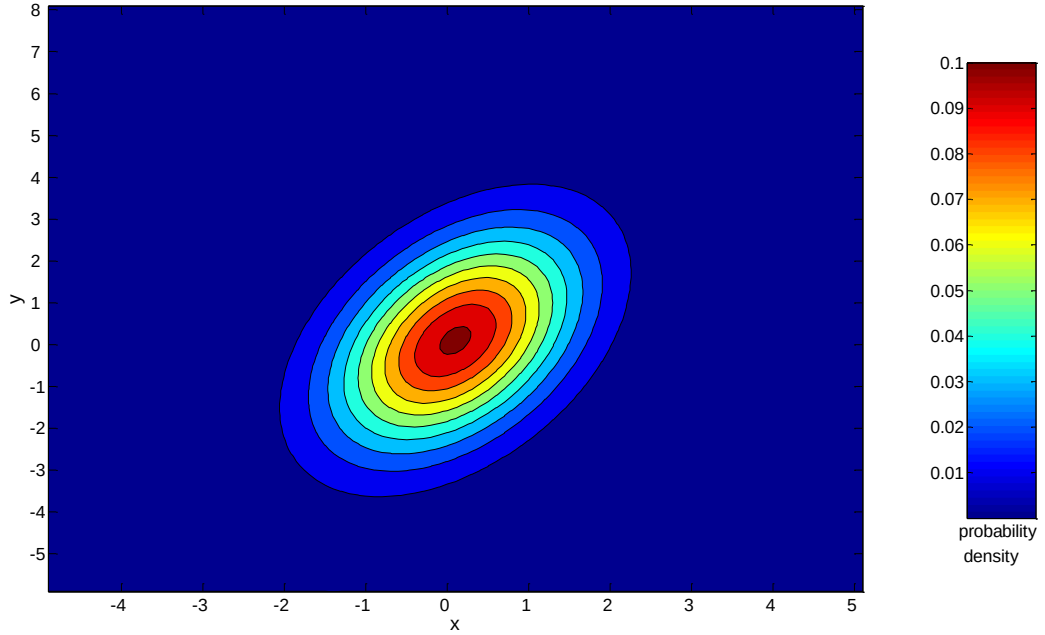
Probabilistic definition of an outlier: An outlier is an object that has a low probability with respect to a probability distribution model of the data.

- 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:

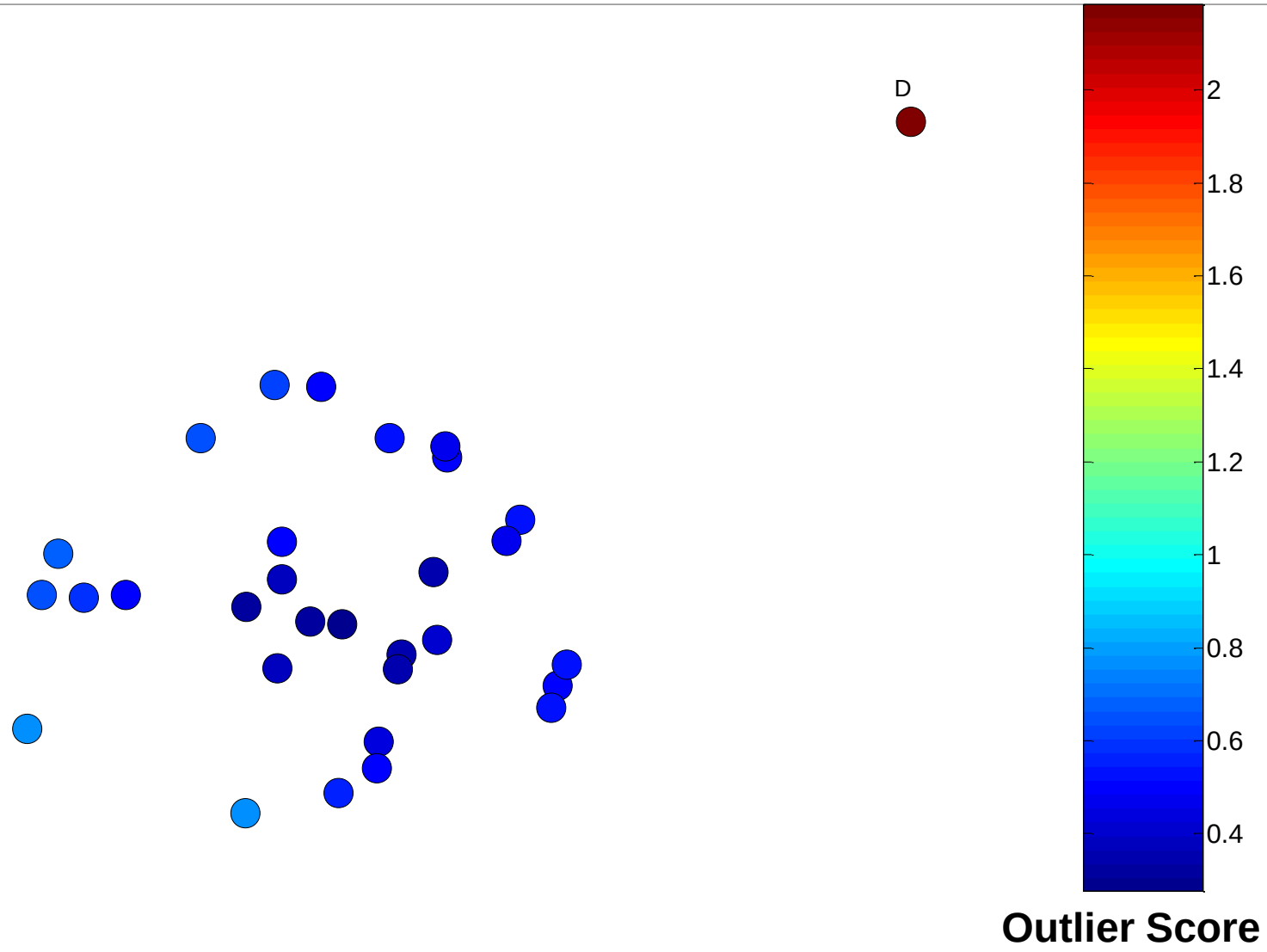
$$G = \frac{\max|X - \overline{X}|}{s}$$

- Reject H_0 if: $G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N,N-2)}}{N-2+t^2_{(\alpha/N,N-2)}}}$

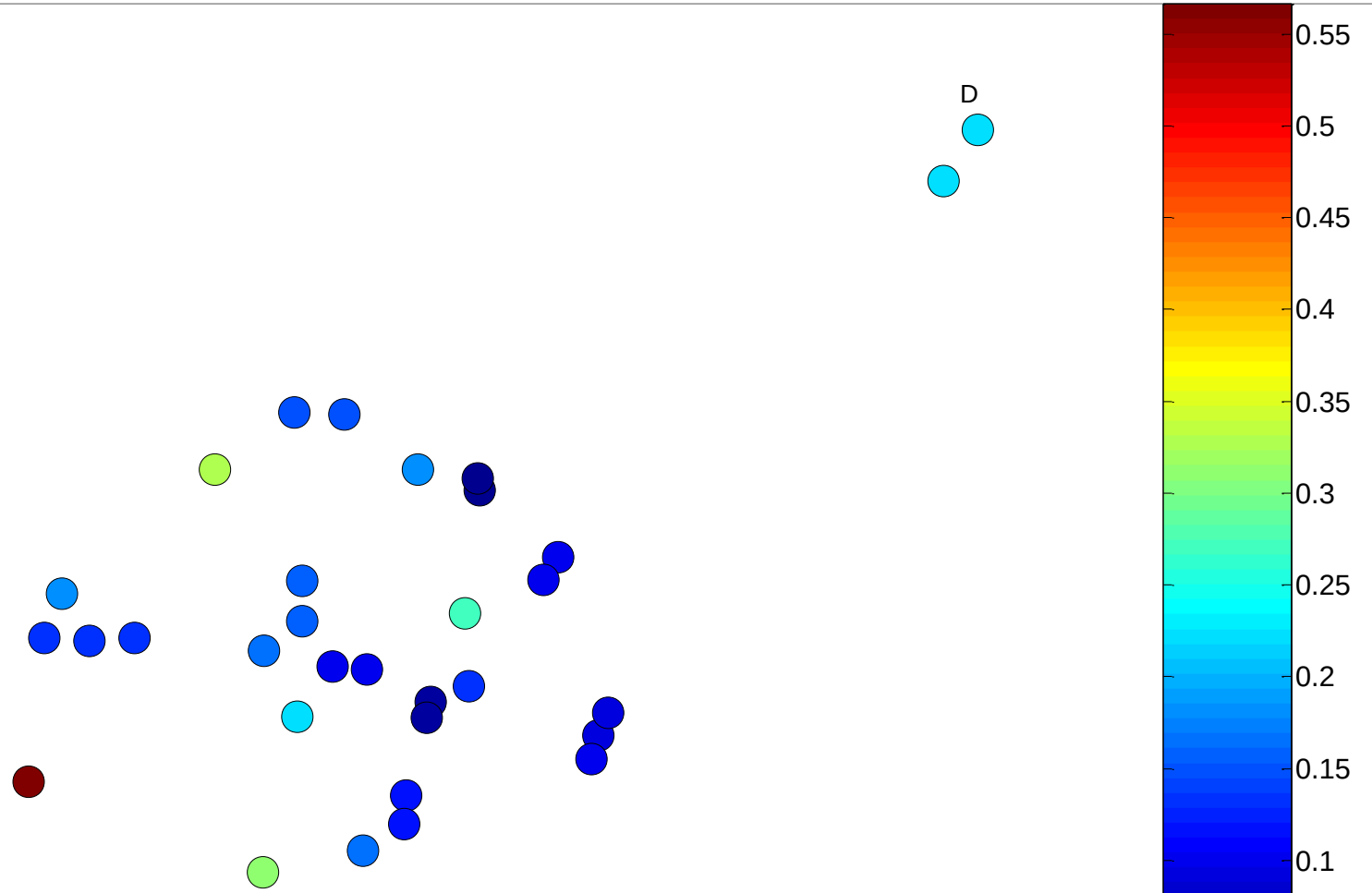
Distance-Based Approaches

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

One Nearest Neighbor - One Outlier

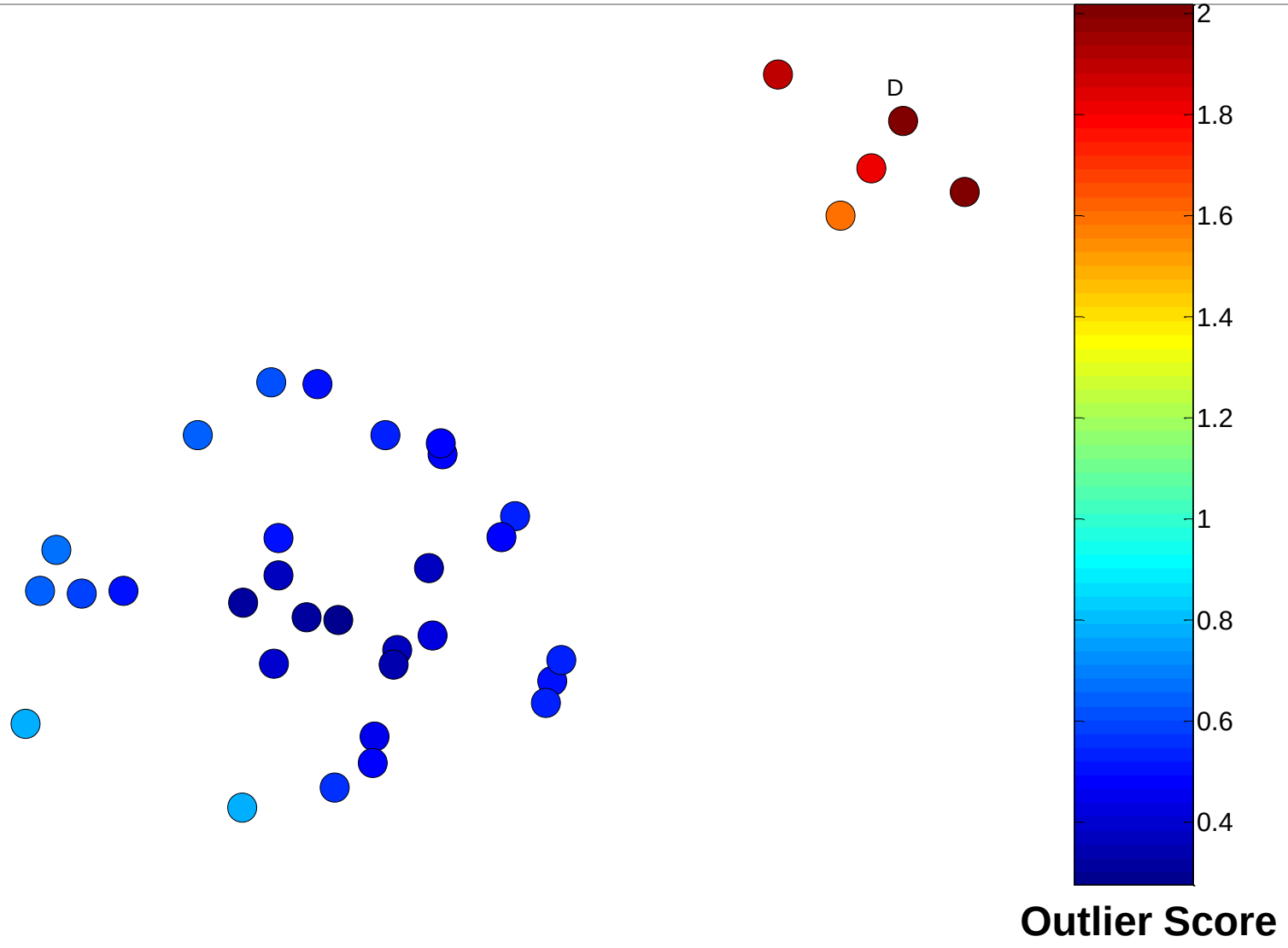


One Nearest Neighbor - Two Outliers

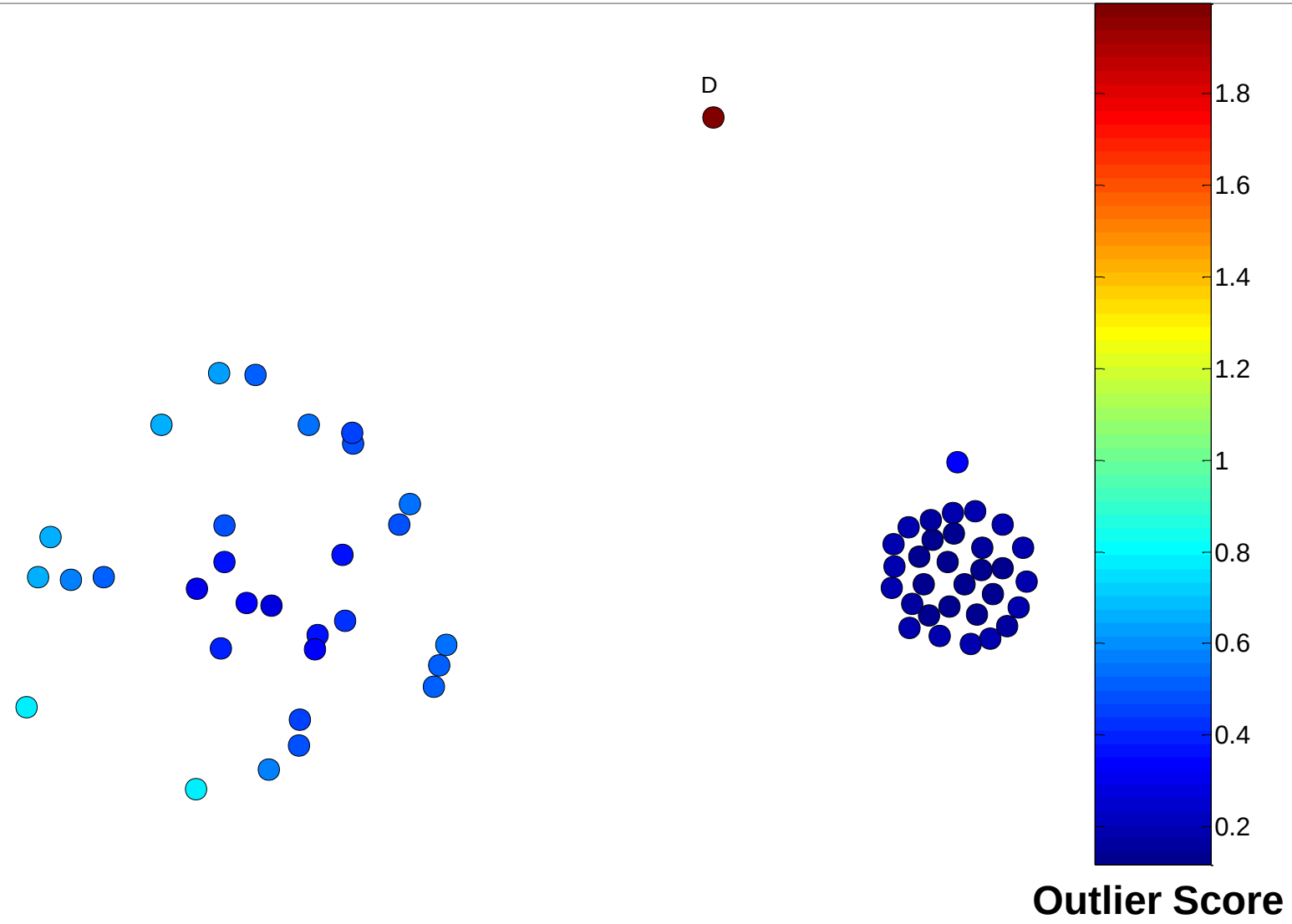


Outlier Score

Five Nearest Neighbors - Small Cluster



Five Nearest Neighbors - Differing Density



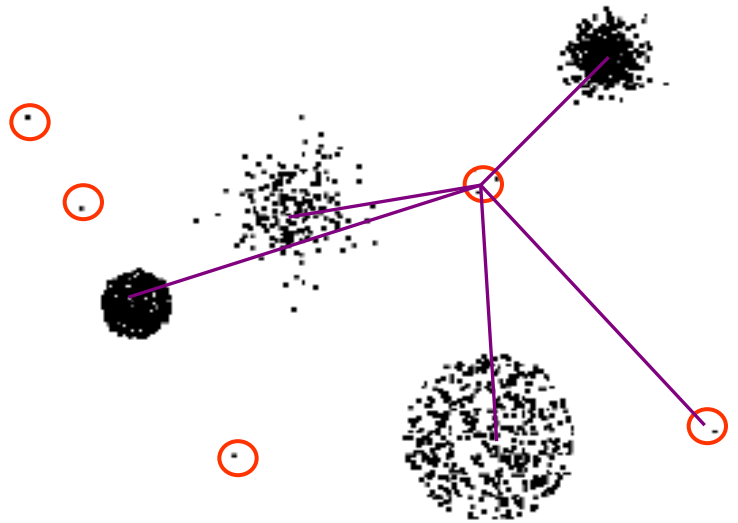
Density-Based Approaches

- **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

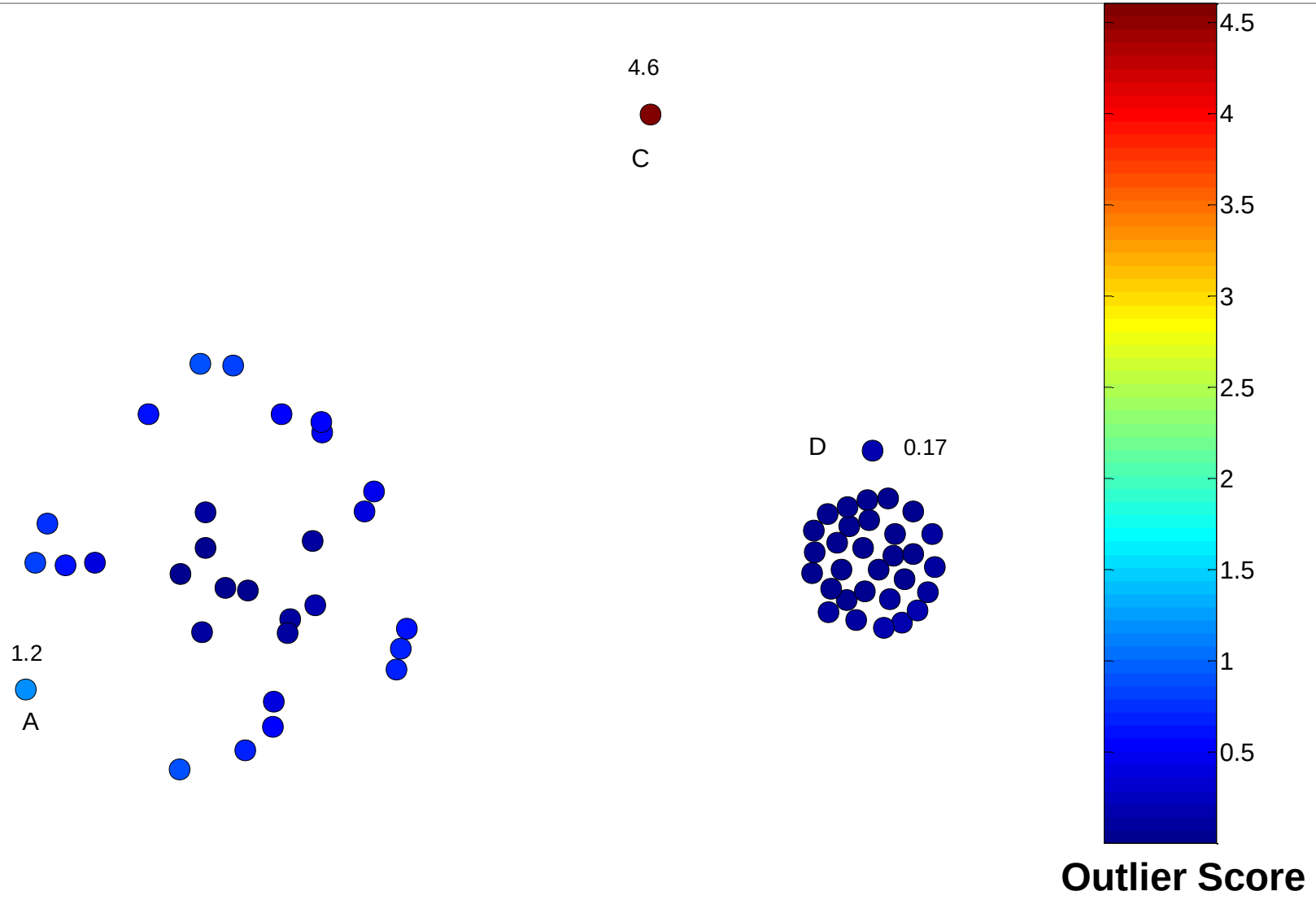
- If there are regions of different density, this approach can have problems

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



Distance of Points from Closest Centroids



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 $\kappa(\mathbf{x}, \mathbf{y}) = \exp(-\frac{||\mathbf{x} - \mathbf{y}||^2}{2\sigma^2})$
  - Every point mapped to a unit hypersphere

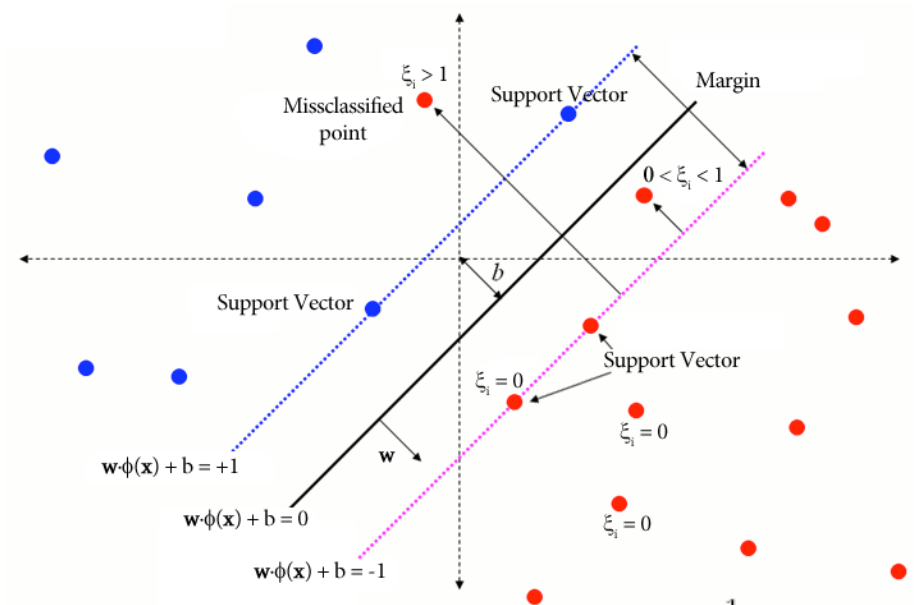
$$\kappa(\mathbf{x}, \mathbf{x}) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}) \rangle = ||\phi(\mathbf{x})||^2 = 1$$

  - 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

Traditional SVM vs One Class SVM

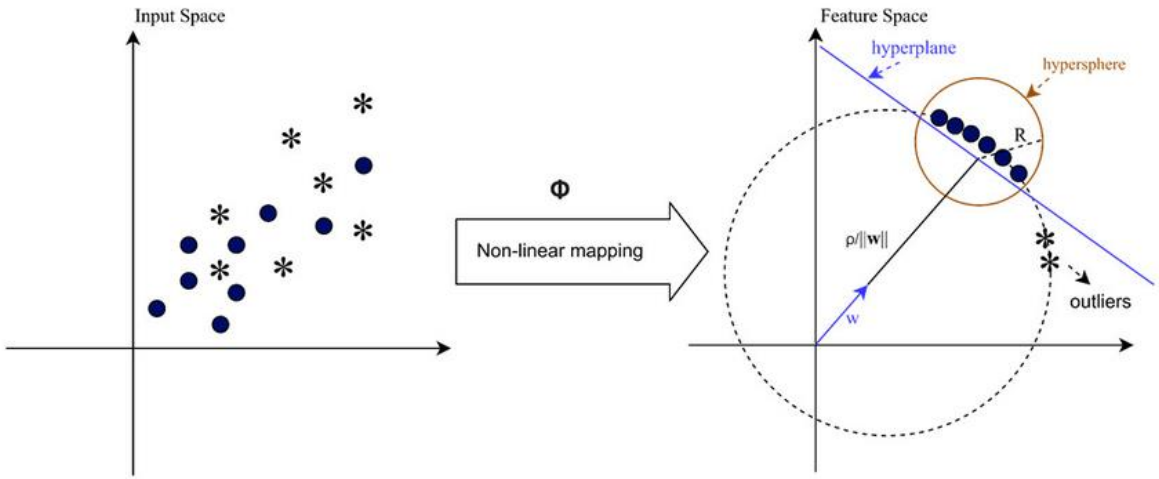


$$\text{Minimize:} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{Subject to:} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

SKernel Functions in One-Class SVM



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

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

https://www.analyticsvidhya.com/blog/2024/03/one-class-svm-for-anomaly-detection/

Equations for One-Class SVM

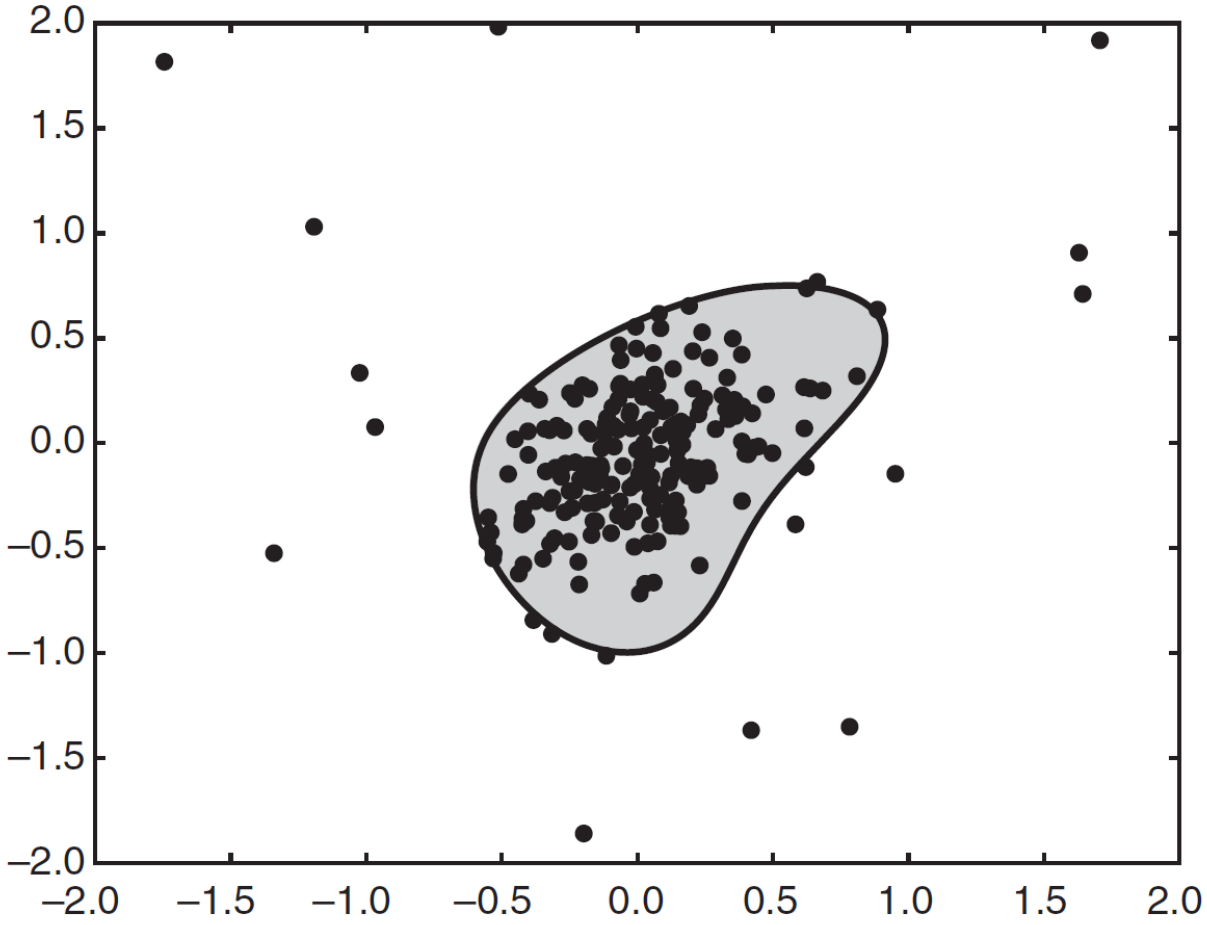
- Equation of hyperplane $\langle \mathbf{w}, \phi(\mathbf{x}) \rangle = \rho$
- ϕ is the mapping to high dimensional space
- Weight vector is $\mathbf{w} = \sum_{i=1}^{n} \alpha_i \phi(\mathbf{x_i})$
- ν is fraction of outliers
- Optimization condition is the following

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

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

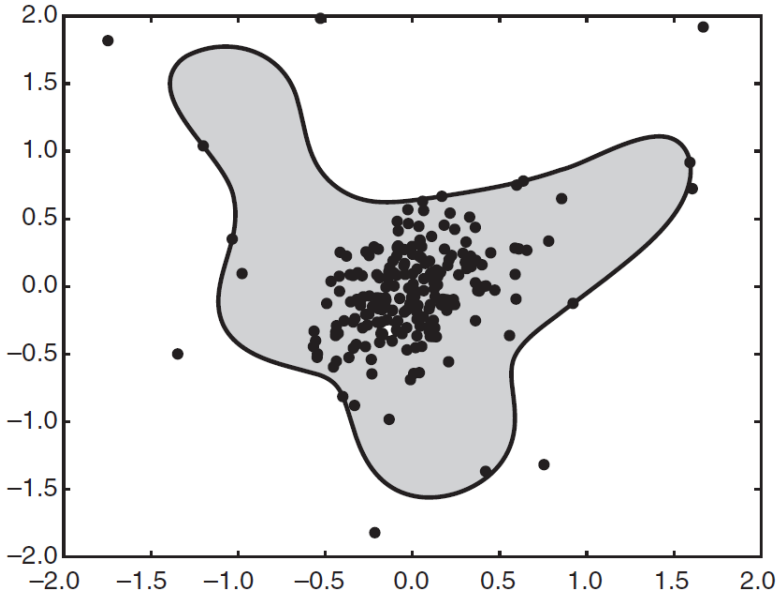
Finding Outliers with a One-Class SVM

- Decision boundary with $\nu = 0.1$

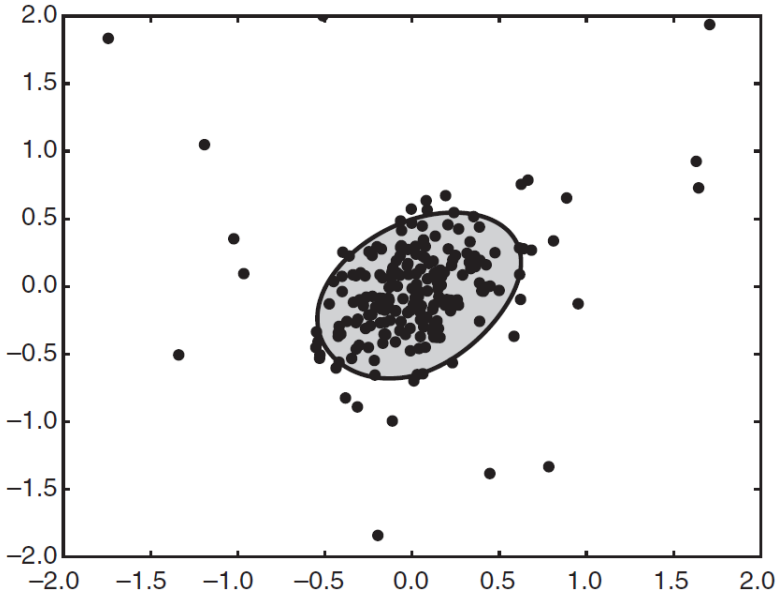


Finding Outliers with a One-Class SVM

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

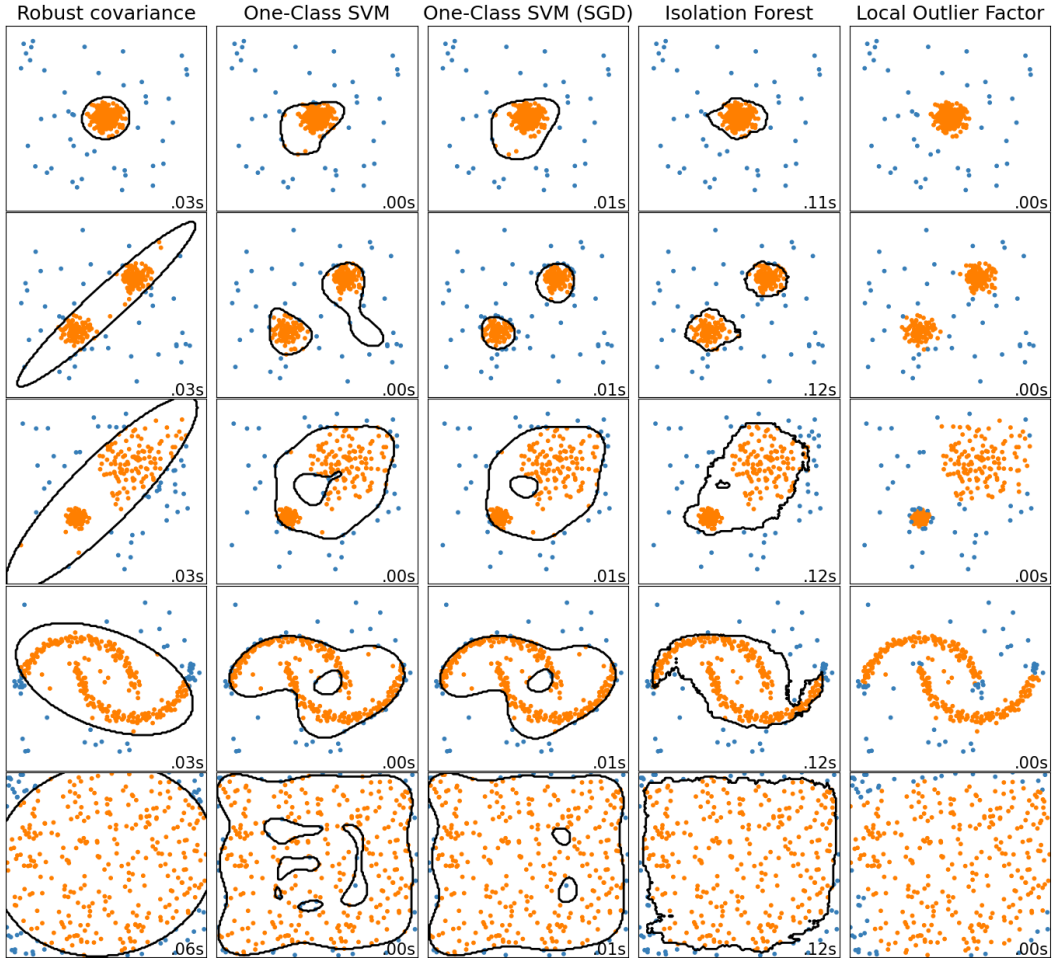


(a) $\nu = 0.05$.



(b) $\nu = 0.2$.

Comparing anomaly detection algorithms



https://scikit-learn.org/1.5/auto_examples/miscellaneous/plot_anomaly_comparison.html