JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Algorithms for Data Science

Unsupervised Learning: Anomaly Detection

Anomaly Detection via Clustering

Anomaly Detection

- Identifies data points significantly deviating from normal patterns.
- Relies on detecting areas of sparse density or outliers in clusters.

Clustering Approach

- Clustering algorithms group similar data points, enabling identification of outliers.
- Anomalies are often located outside defined clusters or in sparse regions.

Applications

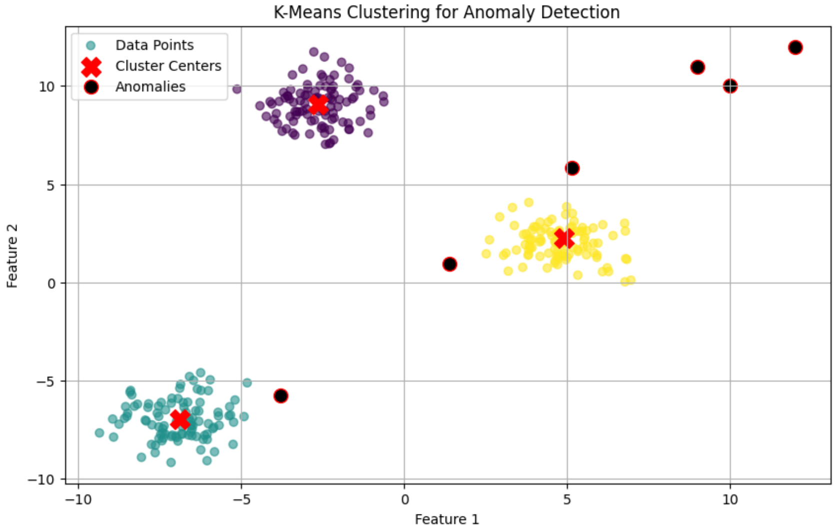
Fraud Detection

Equipment Failure

Network Intrusion

K-Means for Anomaly Detection

- **Approach:**
  - **Perform clustering** with K-means.
  - **Calculate distance** of points to their assigned centroids.
  - Define anomalies as points with distance exceeding a **threshold**.



Density-Based Anomaly Detection

DBSCAN algorithm is used to identify clusters and noise points, where the noise points are candidate anomalies.

Steps:

1. Define ϵ (neighborhood radius) and MinPts (minimum points in a cluster).

2. Identify core points, border points, and noise points.

3. Label noise points as anomalies.

Handles irregular cluster shapes.

Automatic detection of noise without predefined thresholds.

DBSCAN: Mathematical Formulation

- Neighborhood Definition:

$$N_{\varepsilon}(p) = \{ q \in D \mid \text{distance}(p, q) \leq \varepsilon \}$$

- Where:
  - ϵ: Maximum radius of the neighborhood.
  - D: Dataset

- Point Classes:
  - **Core Point:** A point is a core point if $|N_{\epsilon}(p)| \geq MinPts$.
  - **Border Point:** A point within ϵ of a core point but with fewer than MinPts neighbors.
  - **Noise Point:** A point that is neither a core nor a border point.

DBSCAN Algorithm Analysis

1. Neighborhood Computation
   i. For each point p, find all points within ϵ.

Naïve Approach: $O(N^2)$

With spatial indexing: $O(N\log(N))$

2. Cluster Expansion:
   i. Traverse N points, processing each neighbor.

Traversal: $O(N)$

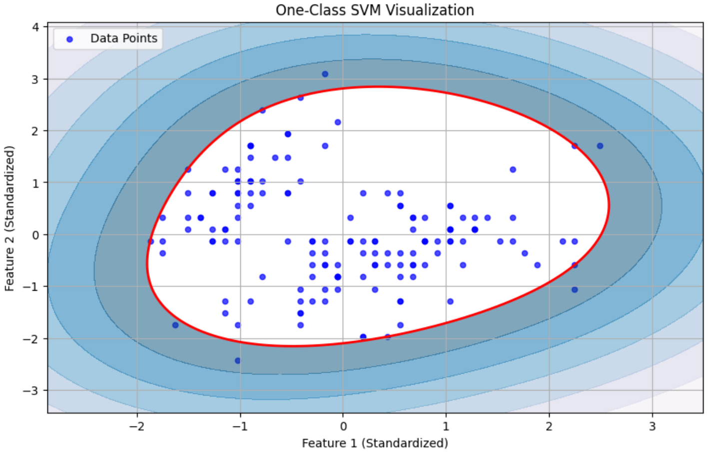
Total Complexity: $O(N^2)$

Optimized Complexity: $O(N\log(N))$

One-Class SVM for Anomaly Detection

A specialized Support Vector Machine (SVM) used to identify outliers by learning a decision boundary around normal data.

- **Purpose:** Distinguish normal data points from anomalies in high-dimensional, complex datasets.
- **Applications:** Fraud Detection, Equipment Monitoring, Rare Event Detection.
- **Key Features:**
  - Nonlinear boundaries via kernel functions.
  - Unsupervised approach.



One-Class SVM: Mathematical Formulation

- Objective Function:

$$\min_{w,\rho} \frac{1}{2}\left\|w\right\|^2 + \frac{1}{\nu N}\sum_{i=1}^{N} \max\left(0, 1 - \left(w \cdot x_i + b\right)\right) - \rho$$

- Where:
  - w: Hyperplane normal vector.
  - ρ: Decision threshold.
  - ν: Trade-off parameter (controls fraction of outliers and support vectors).
  - N: Number of data points.

One-Class SVM: Mathematical Formulation

- Constraints:

$$w \cdot x_i + b \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad \forall i$$

- Where ξ_i are the slack variables for a soft margin.

- Kernel Function:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

DBSCAN Algorithm Analysis

1. Training Complexity:
   i. Quadratic in the number of samples, required by the kernel computation. → Training Complexity: $O(N^2)$

2. Prediction Complexity:
   i. Linear in the number of support vectors. → Prediction Complexity: $O(N_s \cdot d)$

One-Class SVM: Correctness

Theorem: The One-Class SVM algorithm correctly identifies anomalies by finding a decision boundary that encloses the majority of the data points.

Proof:

1. **Optimization Problem:**
   i. The algorithm solves an optimization problem to minimize the complexity of the decision boundary while enclosing most of the data.
2. **Decision Boundary:**
   i. Decision function, $s_i = \left(w \cdot \phi\left(x_i \right) \right) - \rho,$ ensures that:
      - Normal data points within the boundary have $s_i > 0.$
      - Anomalies outside the boundary have $s_i < 0.$
3. **Guaranteed Support:**
   i. The parameter v ensures that a proportion v of the data points will be classified as anomalies.

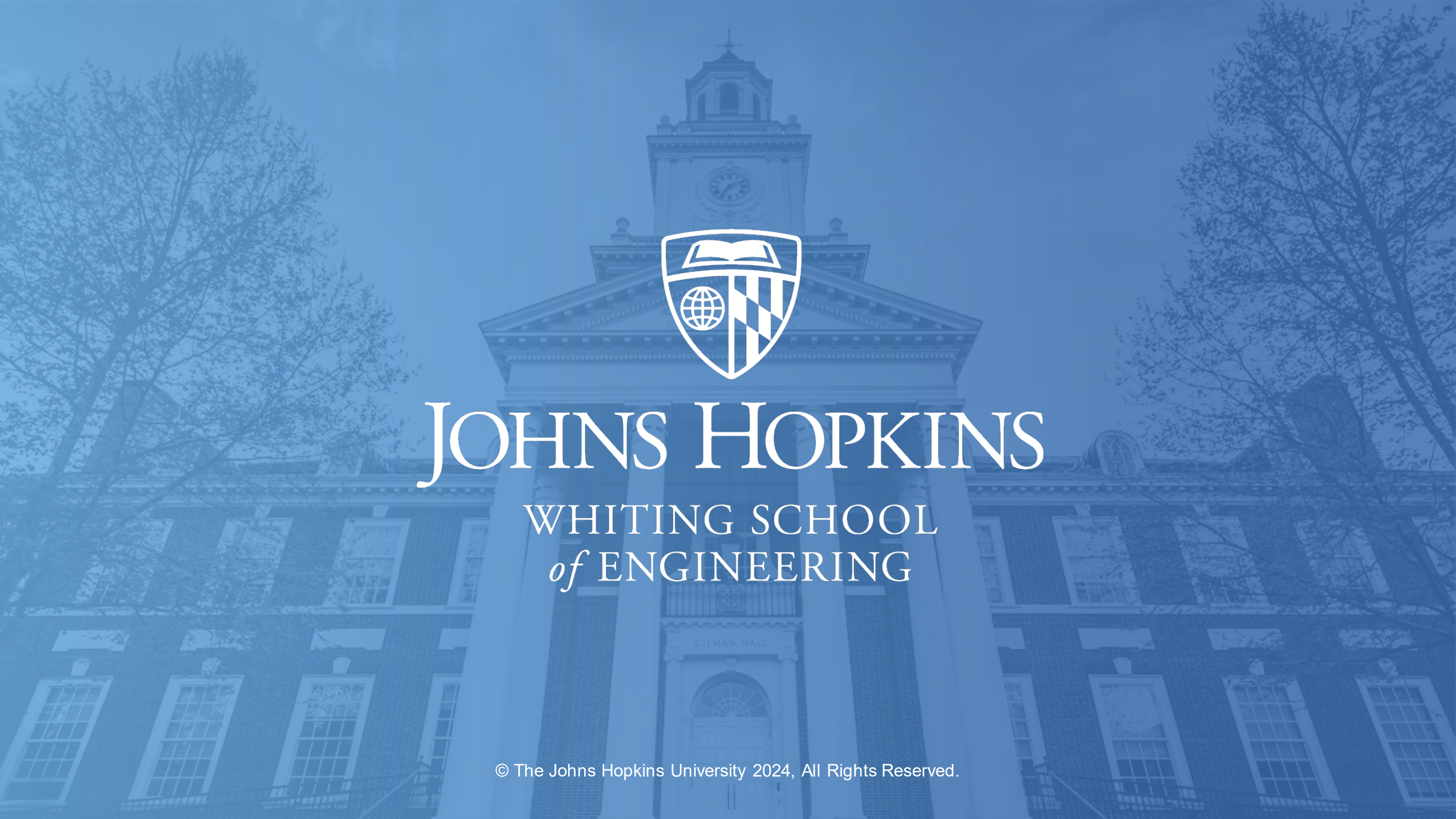
Advantages and Limitations

Advantages

- **Applicability:** Effective for high-dimensional data.
- **Kernel Function:** Captures nonlinear boundaries.
- **Requirements:** Requires only data for training, no labels.

Limitations

- **Runtime Complexity:** Expensive for large datasets.
- **Sensitivity:** Highly sensitive to parameter tuning.
- **Overfitting:** Can overfit when anomalies are similar to train data.



JOHNS HOPKINS

WHITING SCHOOL of ENGINEERING