2nd May Assignment

May 9, 2023

1 Assignment 84

Q1. What is anomaly detection and what is its purpose?

Ans.Anomaly detection is the process of identifying patterns or data points that deviate significantly from the normal behavior of a system or dataset. The purpose of anomaly detection is to detect unusual or suspicious behavior that may indicate a potential threat, fraud, or errors in the system. Anomaly detection is widely used in various fields, including cybersecurity, finance, healthcare, and manufacturing, to detect and prevent anomalies that could cause harm or damage to the system or the organization.

Q2. What are the key challenges in anomaly detection?

Ans. The key challenges in anamoly detection are:

- Lack of labeled data: In many cases, anomalies are rare and difficult to label, making it challenging to train accurate models.
- Imbalanced data: Anomalies often represent a small portion of the overall data, which can lead to imbalanced datasets that are difficult to model.
- Dynamic environments: Real-world systems are often dynamic and can change rapidly, making it challenging to adapt models to changing conditions.
- Feature engineering: Identifying the right features to use in anomaly detection models can be difficult, especially in complex systems with many variables.
- False positives: Anomaly detection models may generate false alarms, which can be costly and time-consuming to investigate.
- Interpretability: Understanding why a particular data point or pattern was flagged as an anomaly can be challenging, especially for complex models.

Q3. How does unsupervised anomaly detection differ from supervised anomaly detection?

Ans.Unsupervised anomaly detection: In this approach, the algorithm is not provided with labeled data that indicates which instances are anomalous. Instead, it relies on identifying patterns or data points that deviate significantly from the norm in

an unsupervised manner. This approach is useful when labeled data is scarce or unavailable.

Supervised anomaly detection: In this approach, the algorithm is trained on labeled data that indicates which instances are anomalous. The algorithm learns to differentiate between normal and anomalous instances based on the labeled data. This approach is useful when labeled data is available and the goal is to identify anomalies that are similar to those seen in the training data.

Q4. What are the main categories of anomaly detection algorithms?

Ans.The main categories of anamoly detection algorithms are:

1. Statistical-based algorithms: These algorithms use statistical methods to identify anomalies based on the assumption that anomalies have different statistical properties than normal data points.

2. Machine learning-based algorithms: These algorithms use machine learning techniques, such as clustering or classification, to identify anomalies. They learn from labeled or unlabeled data and can be either supervised or unsupervised.

3. Deep learning-based algorithms: These algorithms use deep neural networks to identify anomalies in complex datasets. They are particularly useful for identifying anomalies in image, video, and audio data.

4. Rule-based algorithms: These algorithms use a set of predefined rules to identify anomalies. They are often used in systems with specific rules and regulations that must be followed.

5. Hybrid algorithms: These algorithms combine multiple techniques from the above categories to improve anomaly detection performance. For example, a hybrid algorithm may use a statistical-based approach to preprocess the data and a machine learning-based approach to classify anomalies.

Q5. What are the main assumptions made by distance-based anomaly detection methods?

Ans.The main assumptions made by disatnce-based anomaly detction methods are:

1. Normal data points are dense and tightly clustered in the feature space, while anomalies are far away from the normal data points and less dense.

2. The distance between a data point and its k-nearest neighbors can be used to determine whether the data point is an anomaly or not. Anomalies are those data points that have a large distance to their k-nearest neighbors.

3. The distribution of the normal data points in the feature space is assumed to be smooth and continuous. Anomalies are those data points that are far away from this smooth distribution.

4. Distance-based methods assume that the feature space is Euclidean or can be transformed to be Euclidean. In other words, the distance between any two points can be computed using a distance metric such as Euclidean distance.

Q6. How does the LOF algorithm compute anomaly scores?

Ans.For each data point, the k-distance is computed, which is the distance to its k-th nearest neighbor.The k-distance is used to define a k-nearest neighbor (k-NN) region around each data point.

The reachability distance of each data point is computed as the maximum of the k-distance of the data point and the distance between the data point and its neighbor.

The local reachability density (LRD) of each data point is computed as the inverse of the average reachability distance of its k-nearest neighbors.

The local outlier factor (LOF) of each data point is computed as the ratio of the LRD of the data point and the LRD of its k-nearest neighbors. LOF values greater than 1 indicate that a data point is an outlier, while values less than 1 indicate that a data point is similar to its neighbors.

Q7. What are the key parameters of the Isolation Forest algorithm?

Ans.The Isolation Forest algorithm has two key parameters:

1. The number of trees (n_estimators): This parameter specifies the number of trees to be used in the ensemble. A larger number of trees increases the chances of detecting anomalies, but also increases the computational time.
2. The maximum depth of the trees (max_depth): This parameter specifies the maximum depth of each tree in the ensemble. A larger maximum depth allows the trees to better fit the data, but also increases the risk of overfitting.

The Isolation Forest algorithm also has optional parameters, such as the subsampling size (max_samples) and the contamination parameter, which controls the proportion of anomalies in the dataset. The subsampling size controls the number of data points to be randomly selected at each split of the tree, while the contamination parameter determines the expected proportion of anomalies in the dataset.

Q8. If a data point has only 2 neighbours of the same class within a radius of 0.5, what is its anomaly score using KNN with K=10?

Ans.The anomaly score of a data point with only 2 neighbors of the same class within a radius of 0.5 using KNN with K=10 cannot be determined without additional information about the distances to its other neighbors.

Q9. Using the Isolation Forest algorithm with 100 trees and a dataset of 3000 data points, what is the anomaly score for a data point that has an average path length of 5.0 compared to the average path length of the trees?

Ans.if the average path length of the trees in the forest is 10.0, and the data point has an average path length of 5.0, the anomaly score of the data point would be:

- Anomaly score = 5.0 / 10.0 = 0.5

A lower anomaly score indicates a higher likelihood of the data point being an anomaly.