**9**

**Anomaly Detection**

Contents

[**True-False and Short Answer Questions** 2](#_Toc45978459)

[**Proximity-and Density-based Approaches** 4](#_Toc45978460)

[**Clustering-based Approaches** 15](#_Toc45978461)

[**One-class Classification** 17](#_Toc45978462)

[Reconstruction-Based Approaches 18](#_Toc45978463)

[**Information Theoretic Approaches** 20](#_Toc45978464)

[**Comparison of Anomaly Detection Approaches** 21](#_Toc45978465)

# **True-False and Short Answer Questions**

1. Give a True or False answer and a short explanation.
2. When the data has regions of differing density, the LOF anomaly detection algorithm is more effective than the k-nearest neighbor-based outlier detection algorithm.

True, because LOF takes relative density into account.

1. Distance-based anomaly detection is more computationally efficient than statistical anomaly detection.

False. It is usually much easier to compute the parameters of a statistical distribution from data than to compute distances between pairs of objects that are needed for distance-based anomaly detection.

1. Density-based anomaly detection doesn’t work well when there are regions of different density in the data.

True for ordinary density-based approaches, as the outlier scores may be different for every one of the regions involved - some true data points may get high outlier scores. For relative density approaches, such as LOF, this is false.

1. Reconstruction-based anomaly detection is a supervised method.

False. No labeled training data is needed for such methods.

1. Anomalies can distort the model in the case of clustering-based anomaly detection

True, especially if algorithms such as K-means (that are sensitive to outliers) are used for clustering.

1. Statistical anomaly detection is an unsupervised method.

True, statistical anomaly detection does not need any labeled data.

1. Density-based clustering methods can be used for anomaly detection

False. They can’t distinguish between noise and anomalies

1. Data may be concentrated in certain subspaces of the data space. For this situation, which clustering approach that we studied would be appropriate for finding outliers using the cluster-based approach?

CLIQUE, because it takes subspaces into account for clustering.

1. Why does high-dimensionality make density-based outlier detection particularly challenging?

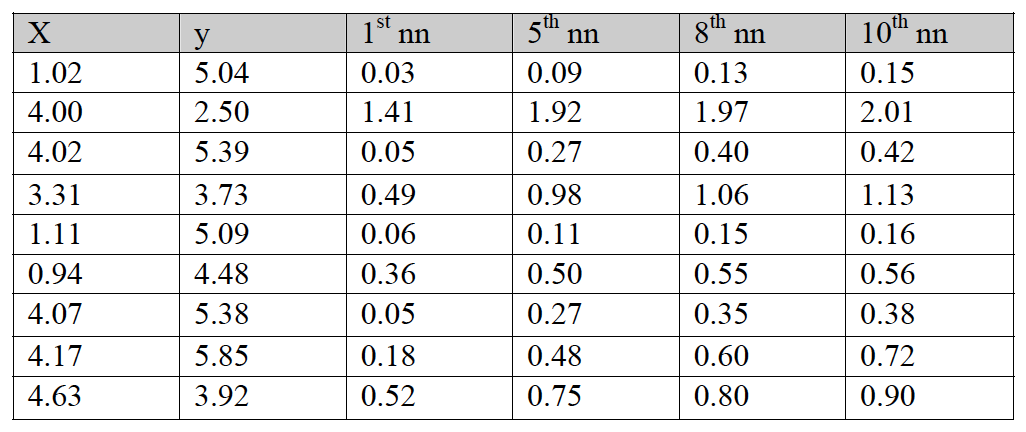
Density is difficult to define in high dimensions.

1. After performing anomaly detection, data miner A wants to find clusters of outliers. Data miner B claims that this does not make any sense and suggests that A re-read the definition of an anomaly. Do you think it is meaningful to cluster anomalies? Explain.

Yes. Although an anomaly is an object that is different from most other objects, it is common for anomalies to fall into groups. For example, anomalous readings on a medical test or a group of medical tests may be a result of various conditions and thus might fall into a set of well-defined groups.

# **Proximity-and Density-based Approaches**

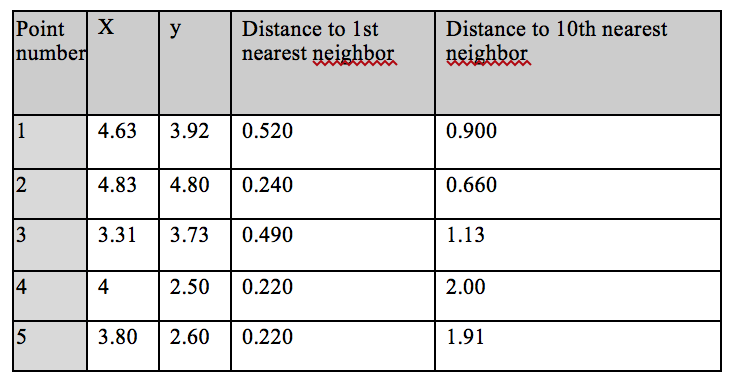
1. Consider the table below, which gives the distances of the points to their 1st, 5th, 8th and 10th nearest neighbors. Determine which points have the top three outlier scores with respect to the KNN method, and comment on the effect of different values of k on the distance to the kth-nearest neighbor algorithm. What would happen if k were increased to n-1, where n is the number of data points?



X=4, X = 3.31, and X= 4.63 have the highest anomaly scores across all *k*. Thus, the parameter *k* has little effect on discovering the top anomalies in this data set for the values of *k* given.

More generally, the effect of *k* is the following. If a few outliers are close to each other relative to the majority of the points, then small values of k will cause the algorithm to fail to discover these outliers. If *k* is too large, e.g., *k* = *n* – 1, then the method will also fail since the *k*th nearest neighbor distance, which is the distance of the farthest point in the data set, is not likely to be related to the local context.

1. Consider a 2-dimensional data with 90 points that belong to two well-defined clusters. The table below shows distances to nth nearest neighbor of the 5 points that are likely to be outliers.



(a) Using the table, state which two points would be considered outliers using the distance to the kth-nearest neighbor algorithm (use k=1 and k=10). Specify the point number.

1. k=1:

**Outliers (pick two points from the set {1,2,3,4,5}):** 1, 3

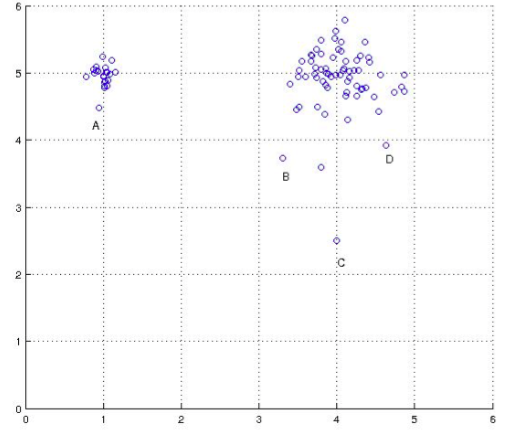
2. k=10:

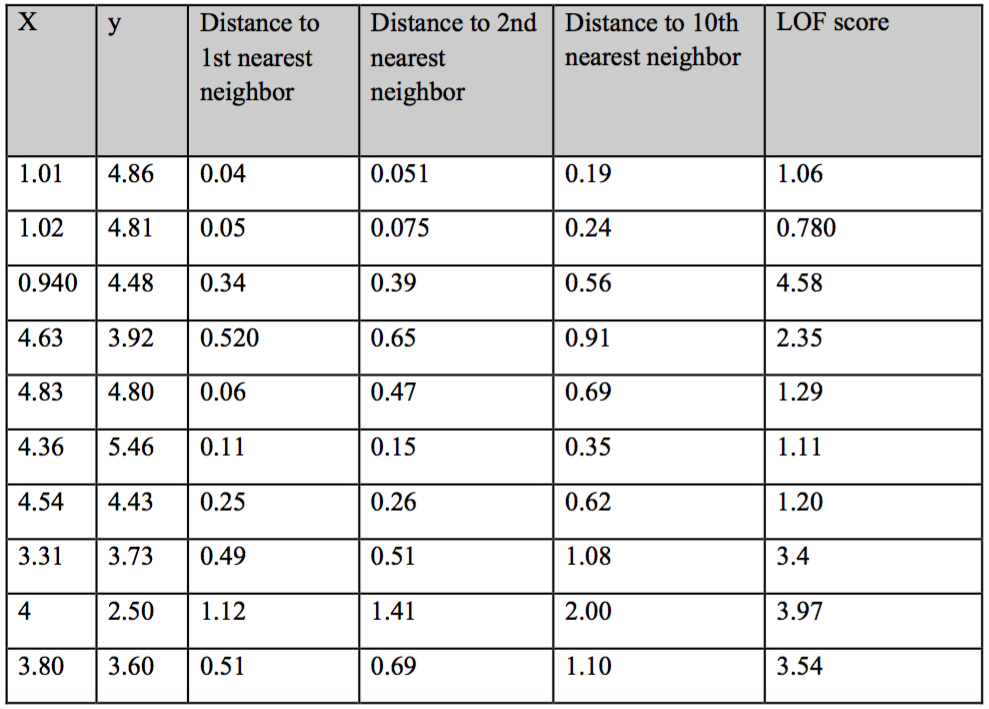
**Outliers (pick two points from the set {1,2,3,4,5}):** 4, 5

(b) Based on your answer in (a), comment on one apparent weakness of the kth-nearest neighbor approach for anomaly detection in this scenario. Suggest an alternative anomaly detection approach that does not have this limitation.

**Answer:** The KNN based approach is sensitive to the choice of k. Any non-proximity based anomaly detection that is not sensitive to the choice of parameter (e.g., Clustering-based (non-K-means), Reconstruction based).

1. Consider the 2-dimensional data set shown in the figure below. There are 90 points that belong to two well-defined clusters.

   
A sample of the points is taken, and a few statistics are computed for them in the table below.



1. Which three points would you consider outliers by visual inspection?

(0.940, 4.48), (4, 2.50) and (3.80, 3.60)

1. Using the table, state which three points would be considered outliers using the distance to kth-nearest neighbor algorithm (assume k=1,2,10) [1\*3]

k = 1: (4,2.5), (4.63,3.92), (3.8,3.6),

k = 2: (4,2.5), (3.8, 3.6), (4.63, 3.92)

k = 10: (4,2.5), (3.8,3.6), (3.31,3.73)

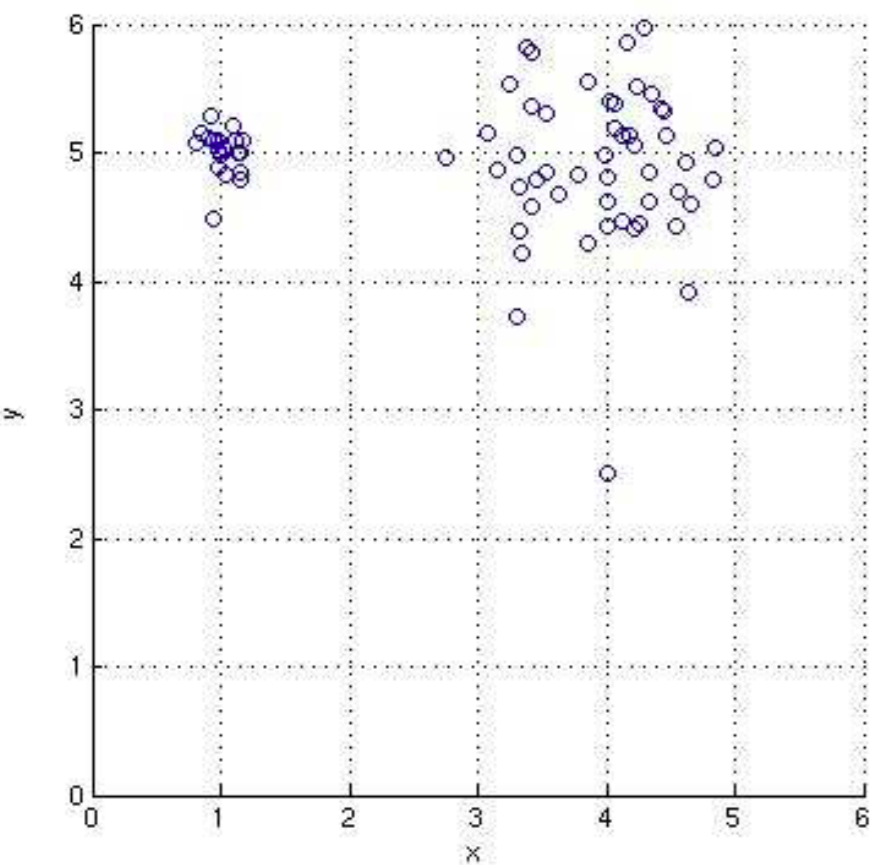
1. c. Using the same table, state which three points would be considered outliers using the LOF algorithm.

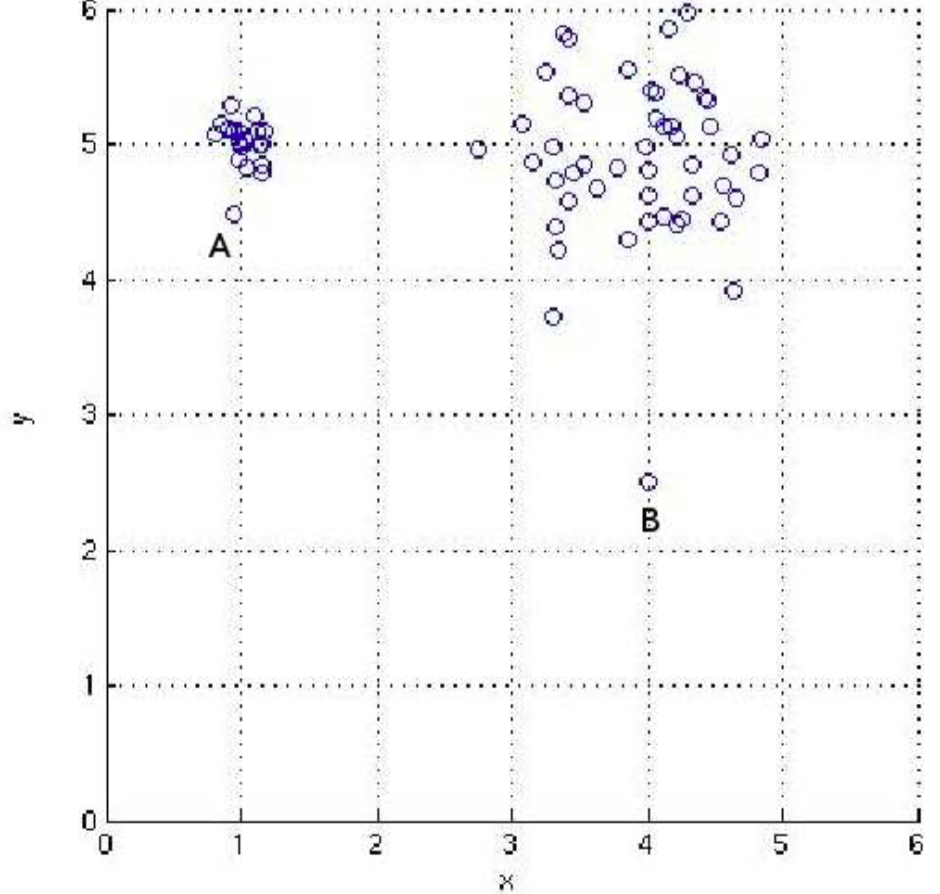
(0.940, 4.48), (4, 2.50) and (3.80, 3.60)

1. Based on your answers to parts (a)-(c), what can you say about the relative performance of the distance to the kth-nearest neighbor and LOF algorithms?

The kth nearest neighbor approach is sensitive to the choice of k. The LOF approach performs better as it is a relative density-based approach, and it can handle the varying density in this data set.

1. Consider the 2-dimensional data set shown in the figure below. There are 70 points that belong to two well-defined clusters.

Which points would you consider outliers by visual inspection? Mark these points on a printout of the figure.   


Answer: points A and B in the figure below. 

1. Comparison of two anomaly detection approaches: distance to the kth nearest neighbor and LOF.
2. Using the following table, state which points would be considered outliers using the distance to the kth-nearest neighbor algorithm (assume k=10). Use a threshold of 0.5 on the distance to the 10th nearest neighbor to determine outliers.

|  |  |  |
| --- | --- | --- |
| X | Y | Distance to 10th nearest neighbor |
| 1.02 | 5.04 | 0.15 |
| 4.00 | 2.50 | 2.01 |
| 4.02 | 5.39 | 0.42 |
| 3.31 | 3.73 | 1.13 |
| 1.11 | 5.09 | 0.16 |
| 0.94 | 4.48 | 0.56 |
| 4.07 | 5.38 | 0.38 |
| 4.17 | 5.85 | 0.72 |
| 4.63 | 3.92 | 0.90 |

The points (4.00,2.50), (3.31,3.73), (0.94,4.48), (4.17,5.85) and (4.63,3.92) are the outliers determined by the kth-nearest neighbor algorithm since their distance to the 10th neighbor is higher than the threshold.

1. Using the table shown below, state which points would be considered outliers using the LOF algorithm. Use a threshold of 2.5 for the LOF score to determine outliers.

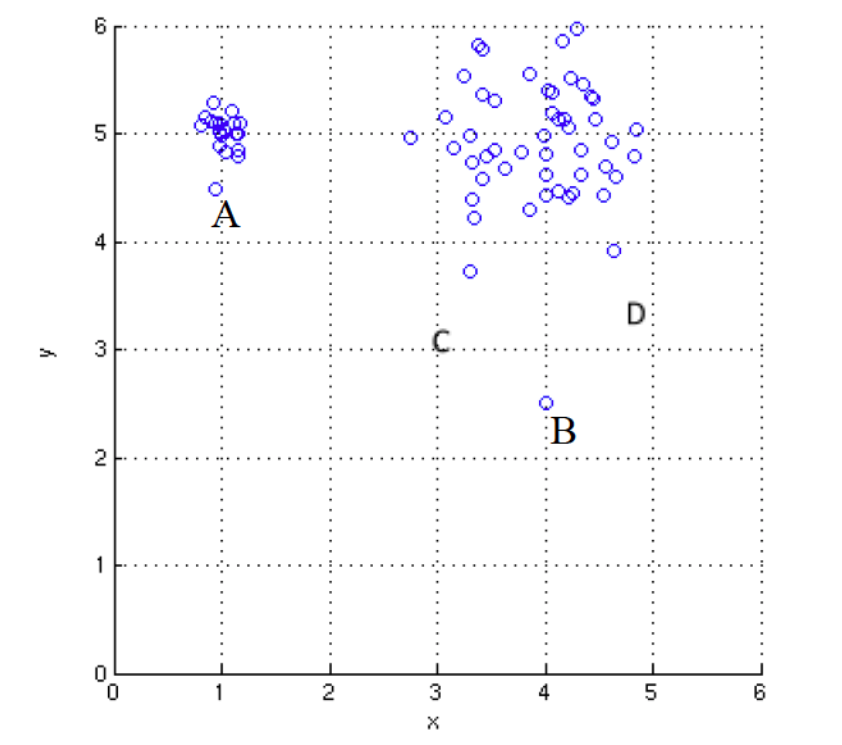
|  |  |  |
| --- | --- | --- |
| X | Y | LOF Score |
| 1.02 | 5.04 | 0.83 |
| 4.00 | 2.50 | 4.88 |
| 4.02 | 5.39 | 1.05 |
| 3.31 | 3.73 | 2.69 |
| 1.11 | 5.09 | 1.02 |
| 0.94 | 4.48 | 3.88 |
| 4.07 | 5.38 | 0.95 |
| 4.17 | 5.85 | 1.57 |
| 4.63 | 3.92 | 2.35 |

The points (4.00,2.50), (3.31,3.73) and (0.94,4.48) are the outliers determined by the LOF algorithm since their LOF score is higher than the threshold.

1. Based on your answers to parts (a) and (b), what can you say about the relative performance of the distance to the kth-nearest neighbor and LOF algorithms?

The outlier detection algorithm based on the distance to the kth-nearest neighbor marked more non-outlier points as outliers because the distance between A and its kth nearest neighbors is less than the distance between some pairs of points in the larger cluster. LOF, which is based on local density, made only one false discovery as against three made by the distance-based algorithm.

1. Consider the 2-dimensional dataset shown in the figure below. The dataset contains 70 points. Given that there are TWO outliers in the dataset, your goal is to discover the two outliers.



1. By visual inspection, which two points would you consider as the outliers: A, B, C or D?

A and B.

1. Assume that you are asked to use the KNN method to detect outliers with k = 1. Which two points would you consider as the outliers: A, B, C or D? You can find the distances to their 1st nearest neighbor in the following table.

|  |  |
| --- | --- |
| Point | Distance to 1st nearest neighbor |
| A | 0.3 |
| B | 1.4 |
| C | 0.35 |
| D | 0.4 |

B and D

1. Assume that you are asked to use the LOF method to detect outliers. Which two points would you consider as the outliers: A, B, C or D? You can find their LOF scores in the following table.

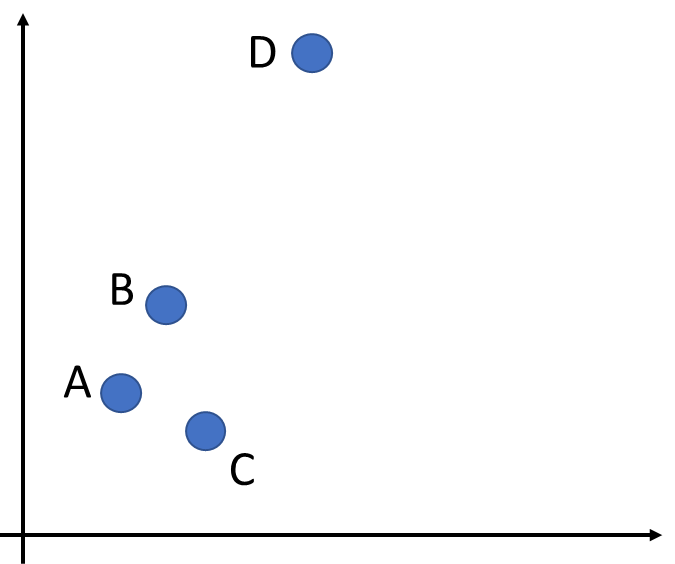
|  |  |
| --- | --- |
| Point | LOF score |
| A | 3.88 |
| B | 4.88 |
| C | 2.69 |
| D | 2.35 |

A and B

1. Based on your answers to question a-c, what can you say about the relative performance of the distance to the kth-nearest neighbor and LOF algorithms?

The outlier detection algorithm based on the distance to the kth-nearest neighbor marked non-outlier points as outliers because the distance between A and its kth nearest neighbors is less than the distance between some pairs of points in the larger cluster. LOF, which is based on local density, made no false discoveries.

1. For each dataset given below, explain how you will design a distance-based outlier detection method that will correctly identify the outliers.
2. Consider the two-dimensional dataset given below. Assume the data point D is an outlier.

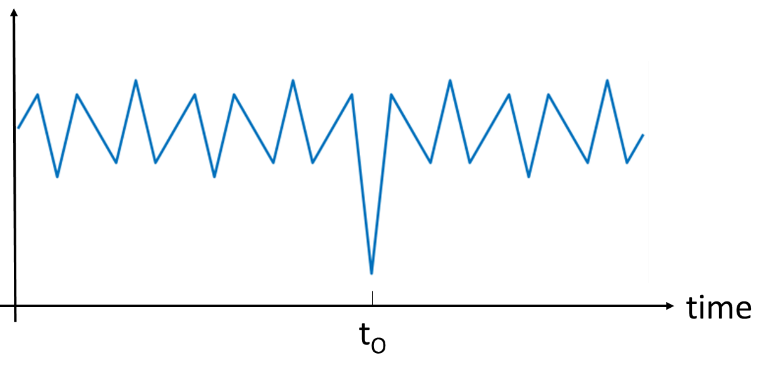


Apply distance-based outlier detection with k=1 using Euclidean distance for finding 1-nearest neighbor.

1. Consider the two-dimensional dataset given in part (a). Assume the data point C is an outlier.

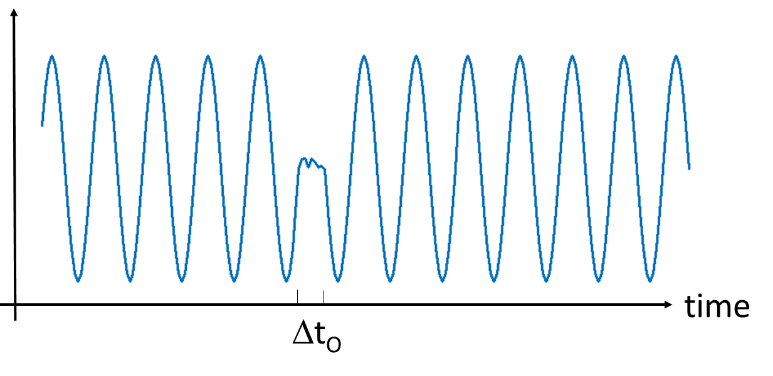
Apply distance-based outlier detection with k=1 using cosine similarity for finding 1-nearest neighbor.

1. Consider the univariate time series data shown below. Assume the value at time tO is an outlier.



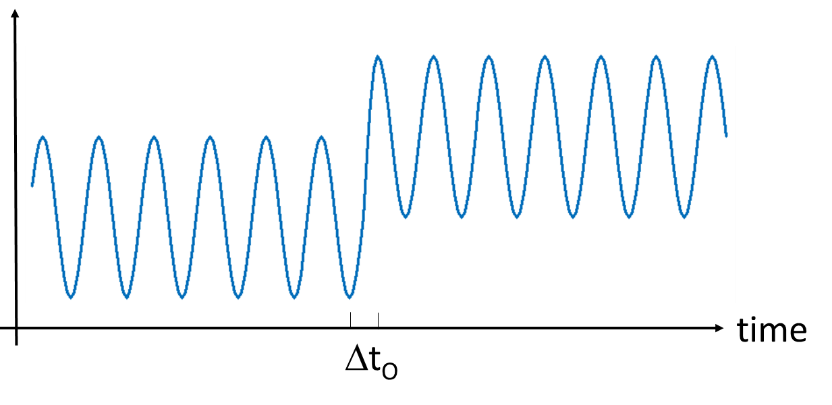
Apply distance-based outlier detection with k=1 to each time step using the difference from the previous time step as a distance for finding 1-nearest neighbor outliers.

1. Consider the univariate time series data shown below. Assume the values within the segment labeled as ΔtO are outliers.



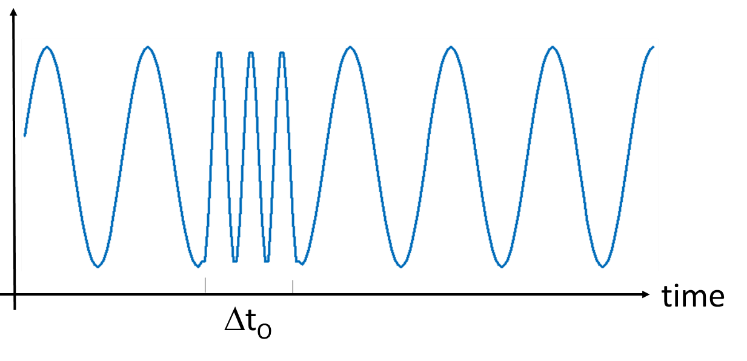
First, partition the time series into fixed-size (potentially overlapping) windows of width at least ΔtO and create a feature vector to represent the sequence of values in each window. Apply distance-based outlier detection on the windows using Euclidean distance on the sequence vectors. The number of nearest neighbors (k) depends on the window size and whether the time windows are overlapping. If the windows are disjoint and the window size is equal to the period of the time series, then k=1 is sufficient to detect the outlying segment. If windows are overlapping, then it may be necessary to increase the value of k, as several overlapping windows may contain anomalous data.

1. Consider the univariate time series data shown below. Assume the values within the segment labeled as ΔtO are outliers.



First, partition the time series into fixed-size (potentially overlapping) windows of width at least ΔtO and create a feature vector to represent the sequence of values in each window. Apply distance-based outlier detection on the windows using Euclidean distance on the sequence vectors. The number of nearest neighbors (k) depends on the window size and whether the time windows are overlapping. If the windows are disjoint and the window size is equal to half of the periodicity of the time series, then k=1 is sufficient to detect the outlying segment. If windows are overlapping, then it may be necessary to increase the value of k, as several overlapping windows may contain anomalous data.

1. Consider the univariate time series data shown below. Assume the values within the segment labeled as ΔtO are outliers.



First, partition the time series into fixed-size (potentially overlapping) windows of width at least ΔtO and create a feature vector to represent the sequence of values in each window. Apply distance-based outlier detection on the windows using Euclidean distance on the sequence vectors. The number of nearest neighbors (k) depends on the window size and whether the time windows are overlapping. If the windows are disjoint and the window size is equal to ΔtO, which is the same as the periodicity of the normal part of the time series, then k=1 is sufficient to detect the outlying segment. If windows are overlapping, then it may be necessary to increase the value of k, as several overlapping windows may contain anomalous data.

1. Consider the problem of applying distance-based outlier detection to high-dimensional data.
2. Explain the challenges in applying distance-based outlier detection to such a dataset.

The notion of distance may not be meaningful for high-dimensional data due to the difficulty of distinguishing outliers from the normal points. Both the storage and computational complexities also increase with higher dimensionality of the data.

1. Suppose we apply principal component analysis (PCA) to reduce the dimensionality of the data first, followed by applying distance-based outlier detection. For each two-dimensional dataset shown below, state whether the PCA-reduced (from 2-dimension to 1-dimension) distance-based outlier detection method can detect the given outlier (shown in red).



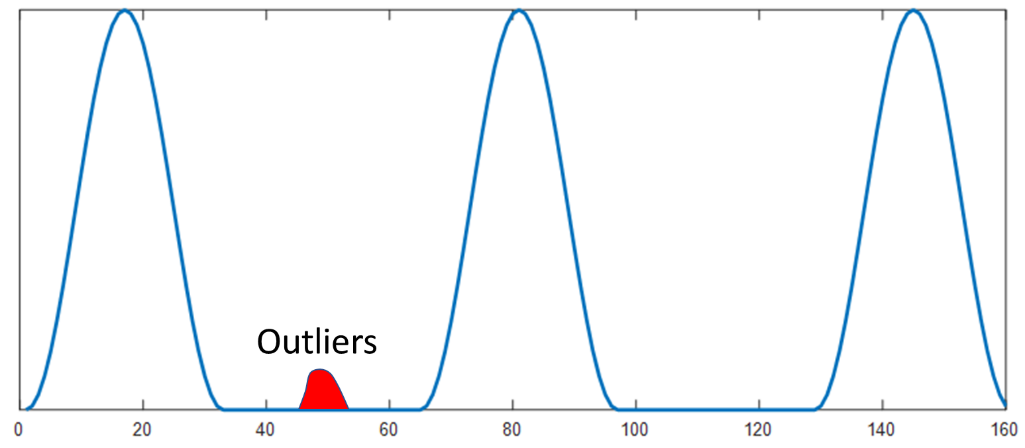
Yes. The principal component will lie along the main diagonal, which is the direction that corresponds to the maximum variance of the data. After projecting each data point along the first principal component, the outlier will have a further distance to its nearest neighbor than other points.



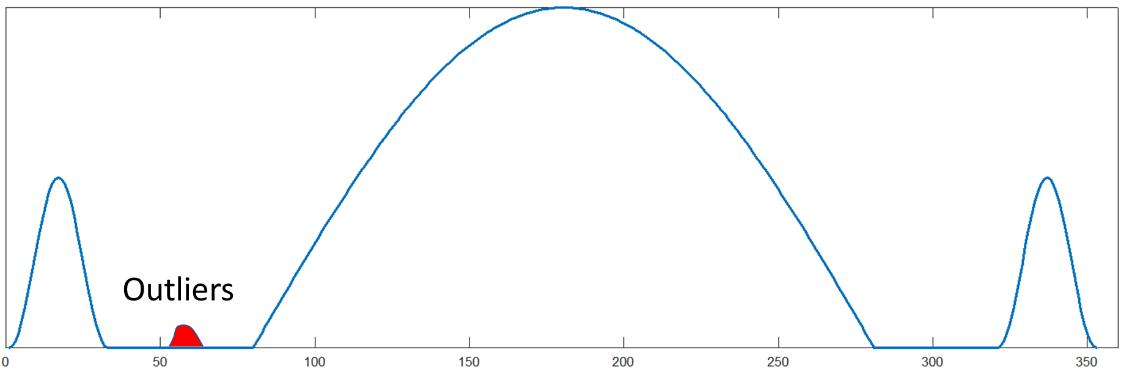
No. The principal component will lie along the main diagonal, which is the direction that corresponds to the maximum variance of the data. After projecting each data point along the first principal component, the outlier will be projected to a point that is close to other normal points in the data. Thus, it will not be able to detect the outlier.

# **Clustering-based Approaches**

1. Suppose we use a clustering-based approach to find anomalies in which the data is clustered into k clusters using K-means, and all points in the smallest cluster are labeled to be anomalous. Assume that this approach is applied to the following 1-D data, and the number of clusters is set to 4. State whether the anomalies can be detected using this approach. For each dataset, we have plotted its probability density function (i.e., histogram) to depict the distribution of points. The highlighted regions indicate the outliers.

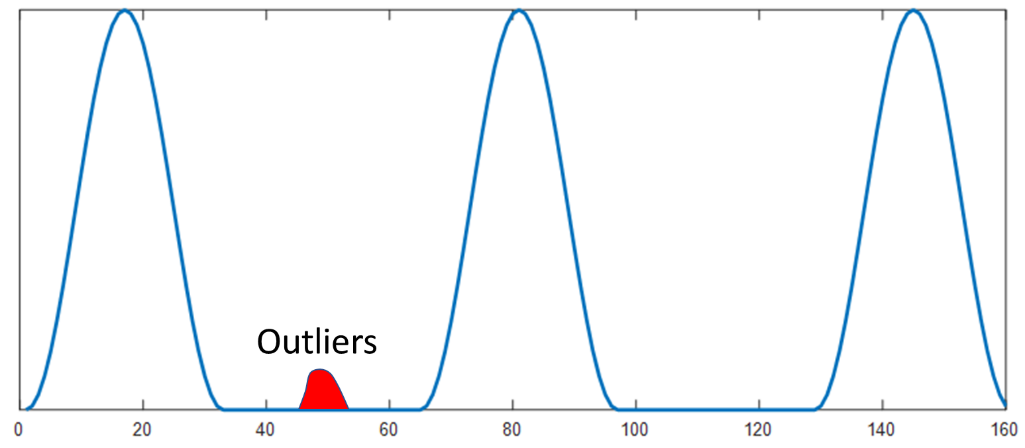


Yes. The outlying cluster is well-separated from the other 3 clusters of normal points.

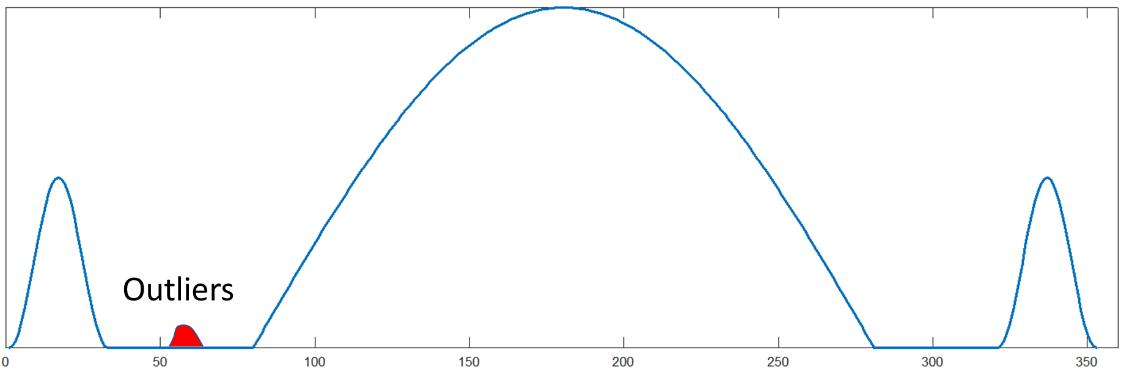


No. K-means will split the large cluster in the middle to smaller clusters. The outliers could be absorbed into one of these clusters.

1. Suppose we apply hierarchical clustering (MIN and MAX algorithm) to detect anomalies in each of the following 1-dimensional datasets. Assume the number of clusters is set to 4. State whether the anomalies can be detected using these approaches. For each dataset, we have plotted its probability density function (i.e., histogram) to depict the distribution of points. The highlighted regions indicate the outliers.



Both MIN and MAX can detect the outlying clusters, which are well-separated from the other 3 clusters of normal points.

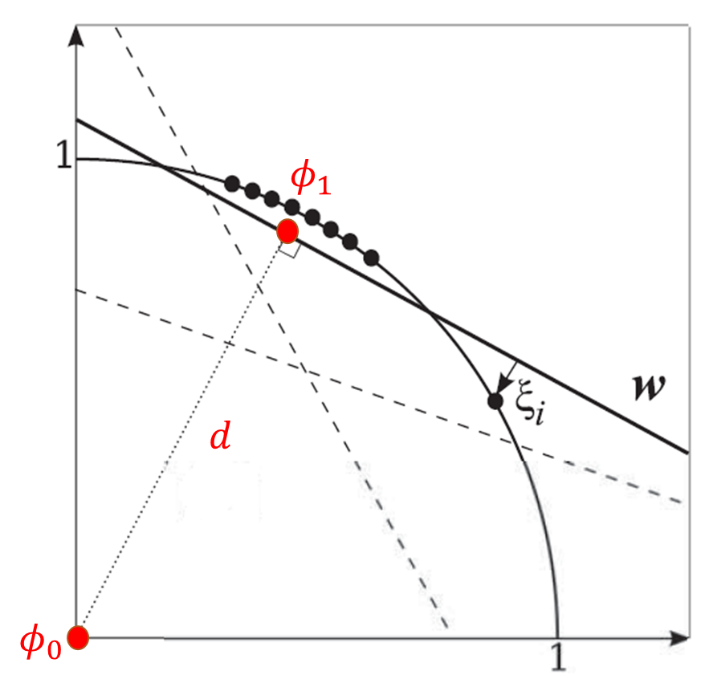


MIN can detect the outlying cluster. However, for the MAX algorithm, the outliers could be absorbed into one of these clusters.

# **One-class Classification**

* + 1. Show that the perpendicular distance between the hyperplane shown in Figure 9.14 to the origin is given by .

We can use a geometric approach to prove this. Consider the scenario shown in the figure below.



Let φ0 and φ1 be a pair of points in the transformed high-dimensional feature space, where φ0 corresponds to the origin and φ1 is the orthogonal projection of φ0 onto the hyperplane. Thus, we have <w, φ0> = 0 and <w, φ1> = ρ. Subtracting them, we have <w, φ1 -φ0> = ρ. Since w is perpendicular to the hyperplane, and thus, parallel to φ1 -φ0, their inner product is given by <w, φ1 -φ0> = ||w|| ||φ1 -φ0|| = ρ. By setting ||φ1 -φ0|| = d, we have ||w|| d = ρ, or equivalently, d = ρ/||w||.

* + 1. Consider the following 4 data points: (5,5), (5,6), (6,5), (1,9).

1. Apply the polynomial kernel (see Equation 4.98 in the book) with degree p = 2 to the data and compute the similarity between every pair of points. Explain the limitation of using the polynomial kernel on the data for one-class SVM.

The kernel matrix is given as follows:

The limitation of applying the polynomial kernel to the data is that the similarity between each point to itself K(x,x) > 1 and less than its similarity to other data points in the data. For example, the similarity between (5,5) to itself is less than its similarity to (5,6).

1. Normalize each data point to have unit length. Compute the following polynomial kernel with degree p=2 using the normalized points:

After normalization, the data points become (0.7071,0.7071), (0.6402,0.7682), (0.7682,0.6402), and (0.1104, 0.9939). The kernel matrix is given by

1. Express the polynomial kernel function given in part (b) with p = 2 as an inner product of two feature functions, φ(x) and φ(y), where x and y are 2-dimensional vectors. Show that the feature vectors φ(x) and φ(y) reside on the arc of a hypersphere with unit radius.

Let x = (x1, x2) and y = (y1, y2). Assuming p=2, we have

Thus, if φ(x) = and φ(y) = , then their inner product will be equal to K(x,y). Also, note that

|| φ(x) || =

Thus, each feature vector resides on the arc of a 5-dimensional hypersphere with unit radius.

# Reconstruction-Based Approaches

1. Consider the 2-dimensional dataset shown in the table and figure below:

|  |  |
| --- | --- |
| x1 | x2 |
| 1 | 2 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 2 | 5 |



For this question, you will apply a PCA-based reconstruction method for anomaly detection.

1. Compute the 2x2 covariance matrix of the data.

1. Derive the first principal component of the data, which corresponds to the eigenvector of the covariance matrix with the largest eigenvalue.

The principal component can be found by solving the eigenvalue equation Cv = λv or equivalently (C - λI) v = 0 for v≠0 (see Equation A.13 in Appendix A). Since v≠0, the non-trivial solution is found by solving the following characteristic polynomial equation: det[C - λI] = 0:

The solution is . Since we are interested in the first principal component, we choose the solution for . The principal component is obtained by solving the following equation:

The solution after solving the system of linear equations above is .

1. Each data point x can be projected to its first principal component v as follows: z = vT(x-μ), where μ is the (column-wise) average of the data. The reconstructed data point is obtained as follows = vTz + μ. Compute the reconstructed value for each data point in the table.

The mean values of the two features are .

|  |  |
| --- | --- |
| x1 | x2 |
| 1 | 2 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 2 | 5 |

|  |
| --- |
| z |
| -1.7970 |
| -1.2029 |
| 0.1956 |
| 1.5941 |
| 1.2103 |

|  |  |
| --- | --- |
| 1 | 2 |
| 1.3324 | 1.0448 |
| 1.6873 | 2.2324 |
| 2.5162 | 3.3573 |
| 3.3470 | 4.4823 |
| 3.1190 | 4.1735 |

1. Based on your answer in part (c), calculate the reconstruction error of each data point using the following formula:

Reconstruction error between (x, ) =

Which data point is an anomaly?

|  |
| --- |
| Reconstruction error |
| 0.1708 |
| 0.1530 |
| 0.3617 |
| 0.6589 |
| 1.9353 |

#### Since the last data point has the highest reconstruction error, it is likely an anomaly.

# **Information Theoretic Approaches**

* + 1. The table below includes credit card activities during March 2018 from 100 different clients. Answer the following questions based on your observation.

|  |  |  |
| --- | --- | --- |
| Yearly Income ($) | Credit card activity on March (amount spent in $) | Frequency |
| < 30K | Medium | 2 |
| < 30K | Low | 17 |
| [30K, 80K] | Low | 20 |
| [81K, 120K] | Medium | 20 |
| [121K, 200K] | High | 15 |
| > 200K | Low | 4 |
| > 200K | High | 22 |

1. What is the information content of the above data? Use the entropy measure to compute that.
2. We want to identify those clients that have unusual/suspicious account activity during March. Do you observe any patterns in the data? Can you identify any irregularities? Explain briefly.
3. Argue about whether there are anomalies in the data or not by using an information-theoretic approach.

Answer:

i. Entropy = 2.55

ii. We observe the following pattern: accounts with high income have high activity, and accounts with low income have lower activity.

Also, we identify two cases of irregularities: the accounts with high income (>200K) combined with low activity and the accounts with low income (<30K) combined with medium activity, which are both quite unusual.

iii. The information content of the data, as measured by the entropy, is 2.55. If we eliminate the 4 instances that correspond to the high income (>200K) unusual case and the 2 instances that correspond to the low income (<30K) unusual case, the entropy of the resulting data becomes 2.307, which results in a gain of 2.55-2.307=0.243.

# **Comparison of Anomaly Detection Approaches**

#### Comparison of anomaly techniques

(i) Give one advantage and one disadvantage of statistical approaches over cluster-based

approaches for anomaly detection.  
  
Advantage: If the distribution is known, the anomaly score, typically the p-value, is easy to compute and has an interpretation that is well-founded in statistics and easy to understand. It is also generally more computationally efficient than cluster-based approaches.

Disadvantage: Data distribution is rarely known exactly, and thus any statistical distribution usually only approximates the actual distribution to some degree.

(ii) Give one advantage and one disadvantage of distance-based approaches (e.g., distance

to kth-nearest neighbor) over statistical approaches for anomaly detection.

Advantage: Easy to compute. No knowledge of the statistical distribution of the data is required.

Disadvantage: Sensitive to the choice of parameters. Cannot handle widely differing density.

#### Comparison of anomaly techniques

a. State one difference and one similarity between clustering-based anomaly detection and proximity-based anomaly detection techniques.

**Similarity:** Both of them work with the notion of similarity, and the similarity could be density, contiguity, etc

**Difference:** Proximity-based anomaly detection treats all points equally when an anomaly score is computed, clustering-based techniques use the notion of cluster centroids or a cluster representative to assign scores

b. Suppose you are using single-link algorithm for anomaly detection. You build a dendogram on the entire data set and then break it into two clusters. Finally, you detect the smaller cluster as an anomaly. Describe one advantage and one disadvantage of this method over using k-means for anomaly detection.

**Advantage:** It will give the same clustering result, and hence will mark the same set of points in different runs

**Disadvantage:** Higher computational and storage requirement

#### Comparison of anomaly techniques

1. Give one similarity and one difference between statistical approaches and clustering-based approaches for anomaly detection.

Similarities:

* 1. In both cases, a model is being built for normal data.
  2. Anomalies can distort clusters and the parameters of a distribution.
  3. It can be difficult to decide on a clustering technique, just as it can be difficult to decide what distribution to use for a statistical approach.  
     Differences:
  4. Statistical approaches require a priori knowledge of the distribution, specifically of the parametric form of the distribution.
  5. Parameters are estimated from the data. Most clustering approaches are non-parametric.
  6. Statistical approaches typically perform poorly for higher-dimensional data as it is difficult to estimate the parameters of the distribution. However, there are clustering-based approaches that work well for higher-dimensional data.

1. Give one advantage and one disadvantage of a distance-based approach over statistical approaches for anomaly detection.

Advantages:

* 1. Distance-based detection is non-parametric, i.e., doesn’t require knowledge of the distribution and is easy to implement.
  2. Anomalies can distort the parameters of the distribution but don’t affect distance-based approaches.

Disadvantages:

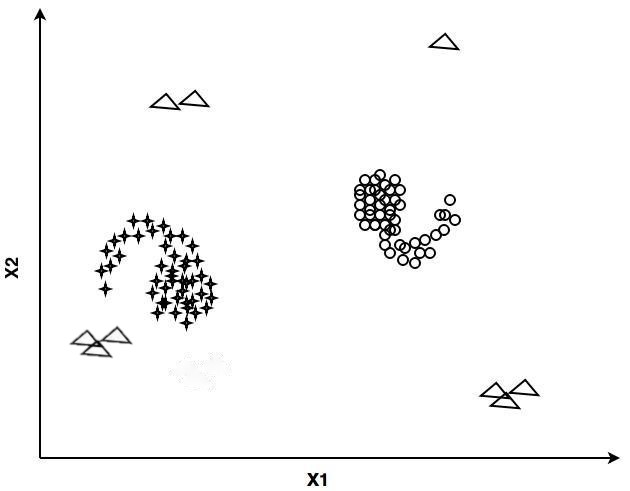
* 1. Scores generated by statistical approaches can be easily interpreted in a probabilistic framework. Statistical approaches have a firm mathematical foundation.
  2. The runtime complexity of distance-based approaches is quadratic in the number of data points, which can be problematic for large datasets. Statistical approaches can be more efficient.
  3. Distance-based approaches are sensitive to the choice of K.
  4. A distance-based approach cannot handle regions of differing densities since it uses a global threshold that cannot account for density differences. A statistical approach, such as the mixture of Gaussians (EM), can take care of this.

#### You are given a data set containing the height, weight, age, and blood pressure of a representative sample of people from a major metropolitan area. Comment on the suitability of using a statistically based versus a cluster-based outlier detection scheme to identify people with anomalous characteristics for this data set.

For statistically-based anomaly detection, you would be faced with the choice of whether to model the data as a single multivariate distribution or as a mixture of distributions (groups). You would also need to decide which type of distributions to use and what significance level to use for classifying an object as an outlier. There are not a lot of multivariate outlier tests, but the Mahalanobis distance test mentioned in the book would be applicable if the data had a multivariate normal distribution, which it might. Alternatively, univariate outlier tests could be applied for each attribute.

For cluster-based outlier detection, you would face the issue of what type of clustering technique should be used and the number of clusters that should be found. You would also have to decide the threshold for what constituted an outlier. To compare, a clustering approach does not require an assumption about the type of distribution, although it does require a choice of clustering techniques. Both approaches require estimating the number of clusters or distributions, and both require deciding on a threshold for calling an object an outlier. The statistics-based approach might have more credibility, but this may not be deserved unless the type of distribution chosen can be validated. For clustering-based approaches, the data would need to be normalized, while for the statistics approach using the Mahalanobis distance, the normalization would occur automatically. For this data, the distribution of data is probably Gaussian, and thus, the statistical approach might be preferred.

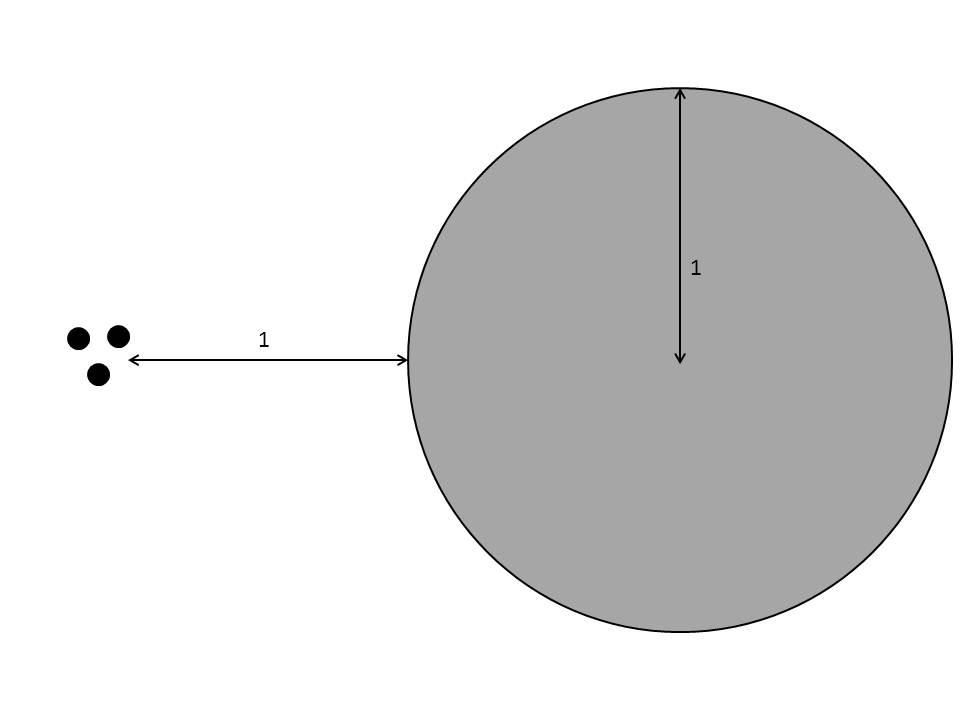
#### Consider the data set shown in the figure below. The figure shows a scatter plot of data points using two attributes X1 and X2. The data consists of two clusters of normal points shown as circles and stars. There are a few anomalous points shown by triangles. Our goal is to identify these triangles as anomalies.



Given the methods in the table below, pick the best method to isolate the outliers and briefly describe the problem with the rest. In case you think none of the methods is suitable, explain the shortcomings of all methods for this data set.

|  |  |
| --- | --- |
| Method | Outlier score |
| KNN (k=4) | Distance to the kth nearest neighbor |
| K-means (k=2) | Distance to the nearest cluster centroid |

KNN is better suited to identify anomalies in this data set. K-means will create centroids at the center of the circular part of the two clusters, and with these cluster centers, the tail of the star cluster is at the same distance as one of the anomaly groups, so it will be hard to capture that without misclassifying the tail as anomalous.



#### Consider the above distribution of points where the grey circle denotes a dense circular cluster of points with 10,000 points, and the 3 dots represent three points that are outliers. The relevant distances have been marked in the figure.

1. We run k-means clustering with k = 2, where the initial choice of centroids is the two points in the data that are farthest away from each other. The smallest cluster is then declared as the set of outliers.

Will this approach be able to detect the 3 outliers? Explain

What if the distance between the three points and the circular set of points were 0.5 instead of 1? Explain.

What if the distance between the three points and the circular set of points were 2 instead of 1? Explain.

No. Since the centroids are initialized as the farthest two points, some points of the grey circle at the left side will be in the same cluster with the 3 outliers in the first iteration. Then, both centroids will end up in the grey circle. Thus, the 3 outliers cannot be detected as anomalies by themselves. Instead, k-means will detect the 3 points and many points in the grey circle as anomalies, and hence this approach will fail.

When the distance between the three points and the circular set of points is 0.5, the result is the same.

When the distance between the three points and the circular set of points is 2 the points in the circle, and the 3 points are found as two separate clusters since the distance between the two sets of points is sufficient to keep the centroid in the set of points from being pulled into the circular set of points.

1. Suppose we run the MIN method and stop merging clusters when the number of clusters is two. Will this approach be able to detect the 3 outliers? Explain.

Yes. The MIN method will detect the 3 outliers. Since the distance between outliers and the grey circle is much larger than their within cluster distance, the two clusters are merged at the very end. Hence, if we stop merging clusters when the number of clusters is two, we will have the 3 outliers as one cluster.

#### For each of the following anomaly detection techniques, fill in one strength and one weakness.

|  |  |  |
| --- | --- | --- |
| Anomaly detection techniques | Strengths | Weaknesses |
| Statistical Approaches | Statistical approaches to outlier detection have a firm theoretical foundation and build on standard statistical techniques. When there is sufficient knowledge of the data and the type of test that should be applied, these approaches are statistically justifiable and can be very effective. They can also provide confidence intervals associated with the anomaly scores, which can be very helpful in making decisions about test instances, e.g., determining thresholds on the anomaly score. | If the wrong model is chosen, then a normal instance can be erroneously identified as an outlier. For example, the data may be modeled as  coming from a Gaussian distribution but may actually come from a distribution that has a higher probability (than the Gaussian distribution) of having values far from the mean. Statistical distributions with this type of behavior  are common in practice and are known as **heavy-tailed distributions**. Also, we note that while there are a wide variety of statistical outlier tests for single attributes, far fewer options are available for multivariate data, and these tests  can perform poorly for high-dimensional data. |
| Distance-Based Approaches | Proximity-based approaches are non-parametric in nature and hence are not restricted to any particular form of distribution of the normal and anomalous  classes. They have broad applicability over a wide range of anomaly detection problems, where a reasonable proximity measure can be defined between  instances. They are quite intuitive and visually appealing since proximity-based  anomalies can be interpreted visually when the data can be displayed in two- or three-dimensional scatter plots. | The effectiveness of proximity-based methods depends greatly on the choice of the distance measure. Defining distances in high-dimensional  spaces can be challenging. Another challenge common to all proximity-based methods is their high computational complexity. Given *n* points, computing the anomaly score for every point requires considering all pair-wise distances, resulting in an *O*(*n*2) running time. For large data sets, this can be too expensive. Choosing the right value of parameters (*k* or *d*) in distance-based methods is also difficult and often requires domain expertise. Finally, distance-based measures don’t handle differing densities well. |
| Density-Based Approaches | Since density-based approaches are typically based on distance, the same advantages apply. In addition, density-based approaches can naturally detect outliers as low-density areas. If a relative density approach is used, as in LOF, areas of differing density can be handled. | Since density-based approaches are typically based on distance, the same disadvantages apply, except that density-based approaches, such as LOF, can handle differing density. |
| Clustering-Based Approaches | Clustering-based techniques can operate in an unsupervised setting as they do not require training data consisting of only normal instances. Along with identifying anomalies, the learned clusters of the normal class help in understanding the nature of the normal data. Some clustering techniques, such as K-means, have linear or near-linear time and space complexity, and thus, an anomaly detection technique based on such algorithms can be highly efficient. | The performance of clustering-based anomaly detection methods is heavily dependent upon the number of clusters used as well as the presence of  outliers in the data. Each clustering algorithm is suitable only for a certain type of data; hence the clustering algorithm needs to be chosen carefully to effectively capture the cluster structure in the  data. |