Cognizant

Welcome
to the course

Anomaly Detection

Anomaly Detection

Anu's Concerns

Meet Anu from the Network team. She is worried about an abnormal behavior detected among the network servers for the past two days.

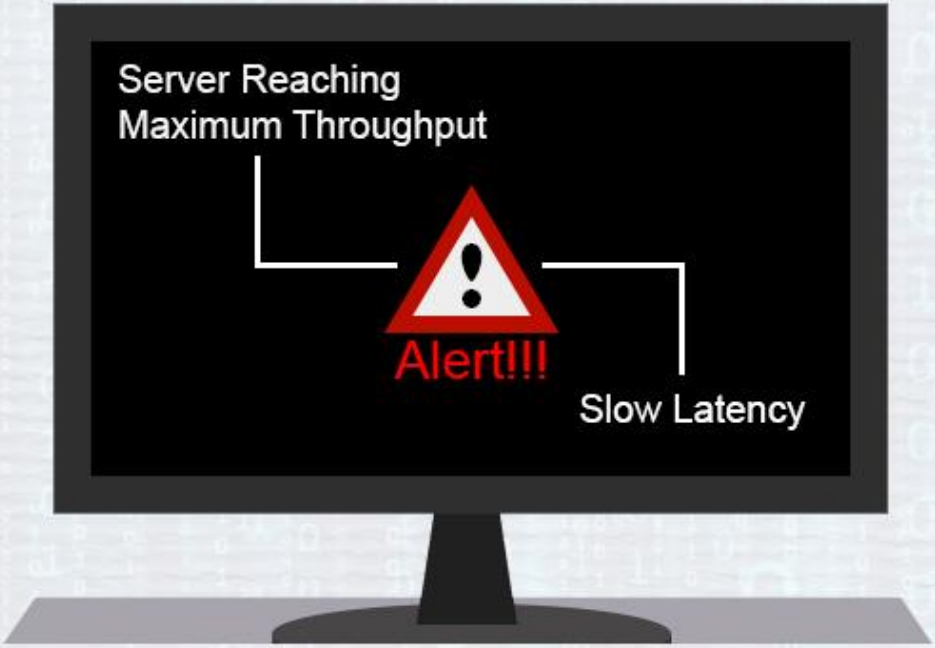
Anomaly Detection

The Problem with the Server

She occasionally receives an alert:

- When the server reaches the maximum load.
- When latency (response) is too slow from the server.

These are abnormal behaviors that happen on rare occasions.



The Mystery around the Problem

Since such abnormal behavior is rarely observed, very little information is available about such scenarios from the past. Most of the data available are related to normal server behavior.

CONTINUE

Pressing Need for a Solution!

However, Anu cannot afford to wait any longer! She has to detect the abnormal behavior immediately. Only then will she be able to resolve the network issues on time, and prevent an adverse impact on the productivity of the affected associates.

CONTINUE

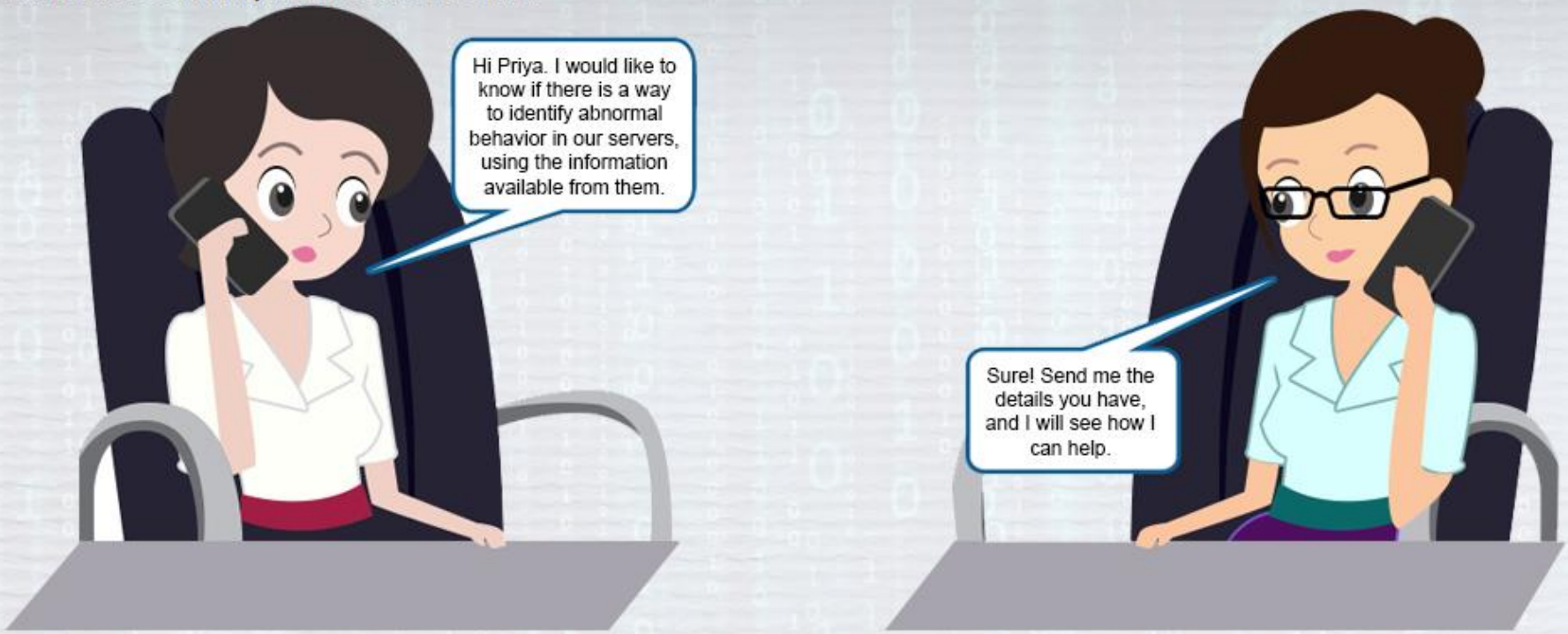
The Ray of Hope

Thankfully, her sources inform her that Priya, a data scientist, has solved similar problems in the past.

CONTINUE

Hence, the Meeting Happened

Anu decides to call Priya and discuss the issue.



Anomaly Detection

Topics for the Day

By the end of this module, you will be able to:

- Describe Anomaly Detection
- Identify the situations where Anomaly Detection is applicable
- Identify the suitable data for Anomaly Detection
- Explain the process of Anomaly Detection

Anomaly Detection

What is Anomaly Detection?

Any rare occurrence that lies outside the boundary of normal occurrences is called an anomaly.

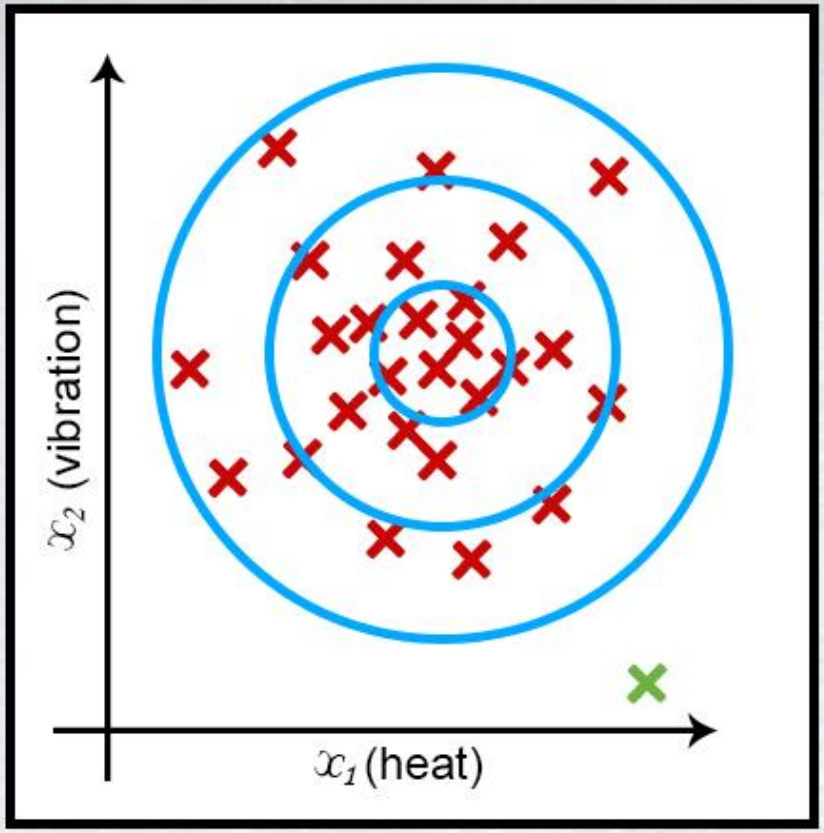
Examples of anomaly detections are:

- Cyber intrusion detections
- Financial fraud detection
- Rare disease identification, etc.

Since, we rarely get data related to such occurrences, they need to be handled in a special way.

The given image is a representation of aircraft engine heat generated vis-a-vis the amount of vibration observed.

Note that the green data point is significantly different from other data points, with abnormally high heat and low vibration.



Why Do We Need Anomaly Detection?

Let's look at an example of how anomaly detection can be useful.

A credit card company stores data related to each transaction. Less than 0.01% of these transactions turn out to be fraudulent. Based on the data collected, the credit card company can identify potentially fraudulent transactions in the future, so that appropriate action can be taken to prevent such occurrences.

To identify such fraudulent activities, the company has to take the help of anomaly detection, because:

- With such few fraudulent transactions, it will be impossible for a learning algorithm to understand all types of fraudulent transactions.
- If fraudsters get innovative and try new techniques, they will not get detected as fraudulent by the machine learning solution.

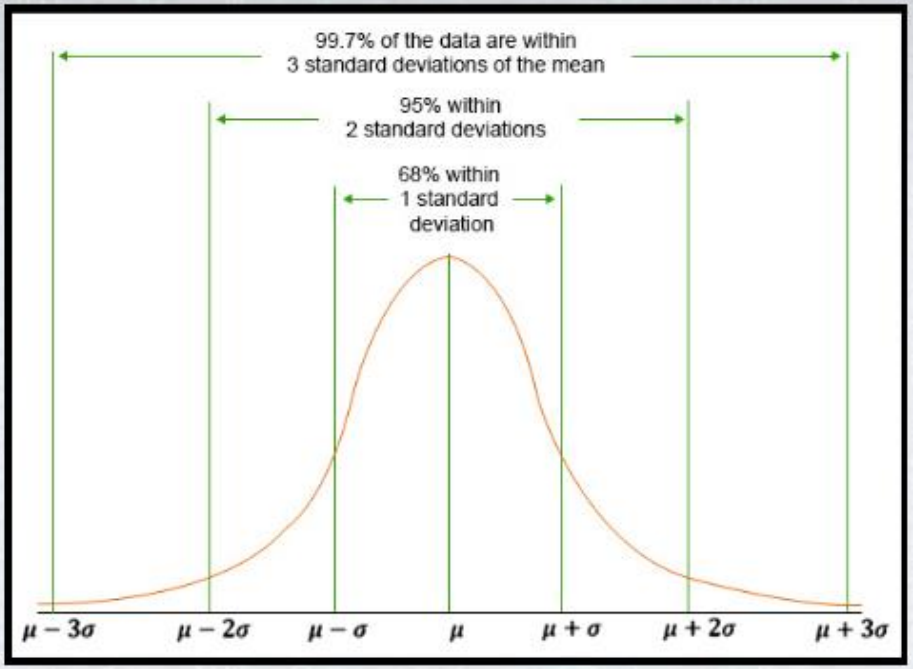
The Secret Lies in the Distribution of Data

Is the problem faced by Anu similar to the challenges that the credit card company faced? Priya must figure this out.

She starts by checking if the data is normally distributed. In a normal distribution, the data follows the bell shape curve.

Refer to the graph here to view the normal distribution of data.

Now that she knows about the distribution of data, Priya will have to follow a few steps to do the data analysis.



μ—It is the average value of all the data points
σ—It is the distribution of data

Note: As we move farther and farther from mean, the number of occurrences become lower and lower. Typically, beyond 3 sigma distance from mean, an occurrence of a value becomes rare.

Anomaly Detection

Step 1: Understanding the Data

To detect anomalies in a dataset, one has to first understand the data. Priya, therefore, starts from the dataset she received.

While taking a look at the data, Priya notices that the dataset consists of 10000 rows and 3 columns. She notes down the details provided in each column.

Column 1: Throughput by the server.
Column 2: Latency of response by the Server.
Column 3: Anomalous/Not Anomalous label (1/0).

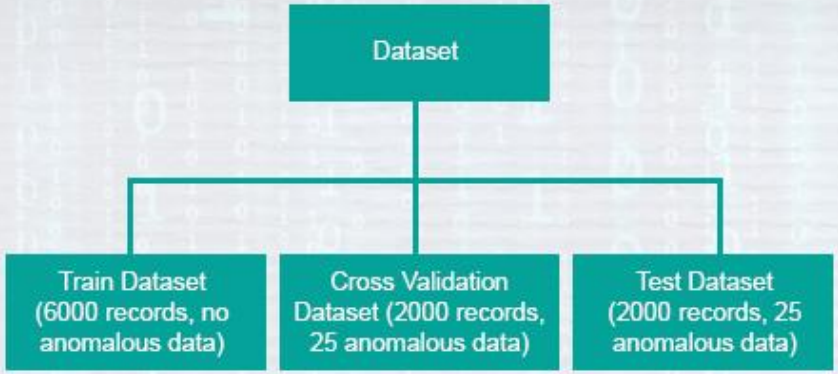
Note: Anomalous data points are very few compared to the normal data points (~0.5% of data set).

Throughput	Latency	Anomalous
Numeric	Numeric	Integer
14.95964562	16.6961271	0
14.52675873	14.61560122	0
11.77803878	14.73074424	0
12.85649088	14.00372929	0

Step 2: Splitting the Data

The second step to apply Anomaly Detection is to split the dataset into three parts - **Train Dataset, Cross Validation Dataset,** and **Test Dataset.**

Priya should split the dataset in the following manner:



Note: 60% Train, 20% Cross Validation, and 20% Test dataset.

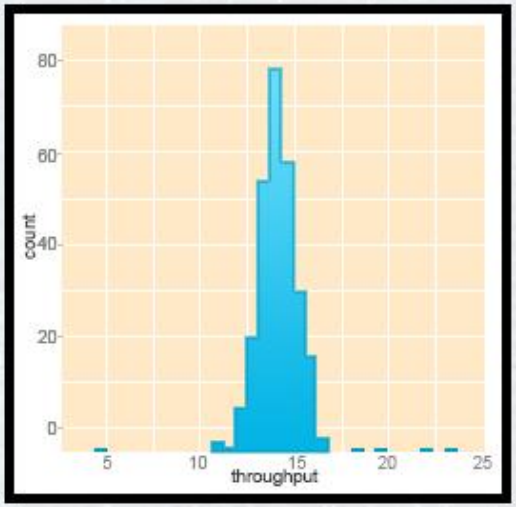
Step 3: Visualizing the Data

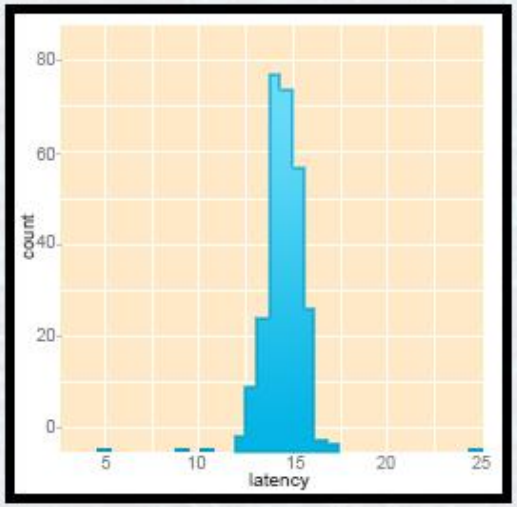
As the next step, Priya plots the data points as histograms to see if the data points form a bell shape curve.

Next, she plots the two features shared by Anu - **Throughput** and **Latency of Response.** To her relief, both the features are normally distributed.

Refer to the graphs to observe the distribution of Throughput and Latency of Response.

A typical throughput value occurs between 11 and 16, while latency value is between 12 and 18.





Step 4: Selecting a Feature(Columns)

Now, Priya should select the important features in the dataset for anomaly detection. While doing that, she recalls a few best practices for choosing features. Priya knows that she has to choose features that might take on unusually large or small values in the event of an anomaly.

Click the Case Study button to view an example of how to choose a feature.

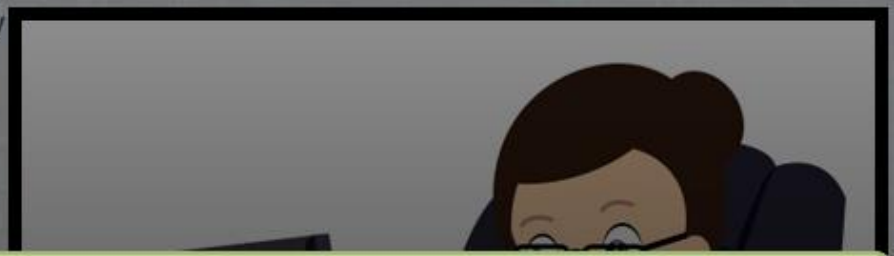
Case Study

Anomaly Detection

Step 4: Selecting a Feature(Columns)

Now, Priya should select the important features in the dataset for anomaly detection. While doing that, she recalls a few best practices for choosing features. Priya knows that she has to choose features that might take on unusually large or small values in the event of an anomaly.

Click the Case Study button to view an example of how to choose a feature.



Case Study

The Context

Let's imagine a data center. The features that are observed for monitoring computers in the center are:

- Memory
- Disk access
- CPU load
- Network traffic

Instances of Anomaly

Let's say that anomalies are more likely to occur when the computer gets stuck in a long loop. In that case, the CPU load is likely to be quite high, and the network traffic quite low.

Spotting Anomalies

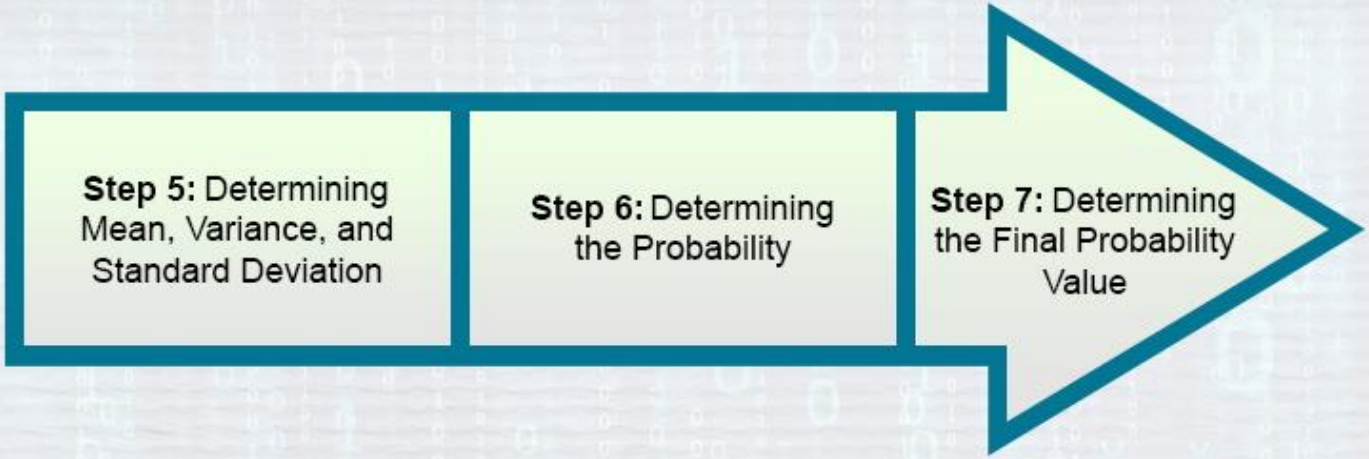
Since, network traffic and CPU load are related, we can engineer a new feature which is a ratio of CPU load to network traffic. This single feature helps in identifying anomalies.

Close

Anomaly Detection Process

Anomaly detection is done in three phases. First, the **Mean, Variance,** and **Standard Deviation** are calculated. Based on that calculation, the **Probability** and **Final Probability Value** are determined.

Take a look at the diagram to view the steps.



Step 5: Determining Mean, Variance, and Standard Deviation

After analyzing all the data, Priya is now ready to initiate the process of Anomaly Detection. She needs to determine the Mean, Variance, and Standard Deviation.

Click each button to learn how to determine mean, variance, and standard deviation.

Mean

Variance

Standard Deviation

Step 5: Determining Mean, Variance, and Standard Deviation

After analyzing all the data, Priya is now ready to initiate the process of Anomaly Detection. She needs to determine the Mean, Variance, and Standard Deviation.

Click each button to learn how to determine mean, variance, and standard deviation.

Mean

Variance

Standard Deviation

Mean (μ)

It is the sum of all data points, divided by the total number of data points. The formula to calculate mean is given below.

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

i,m: Number of observations. j: Number of features

Close

Step 5: Determining Mean, Variance, and Standard Deviation

After analyzing all the data, Priya is now ready to initiate the process of Anomaly Detection. She needs to determine the Mean, Variance, and Standard Deviation.

Click each button to learn how to determine mean, variance, and standard deviation.

Mean

Variance

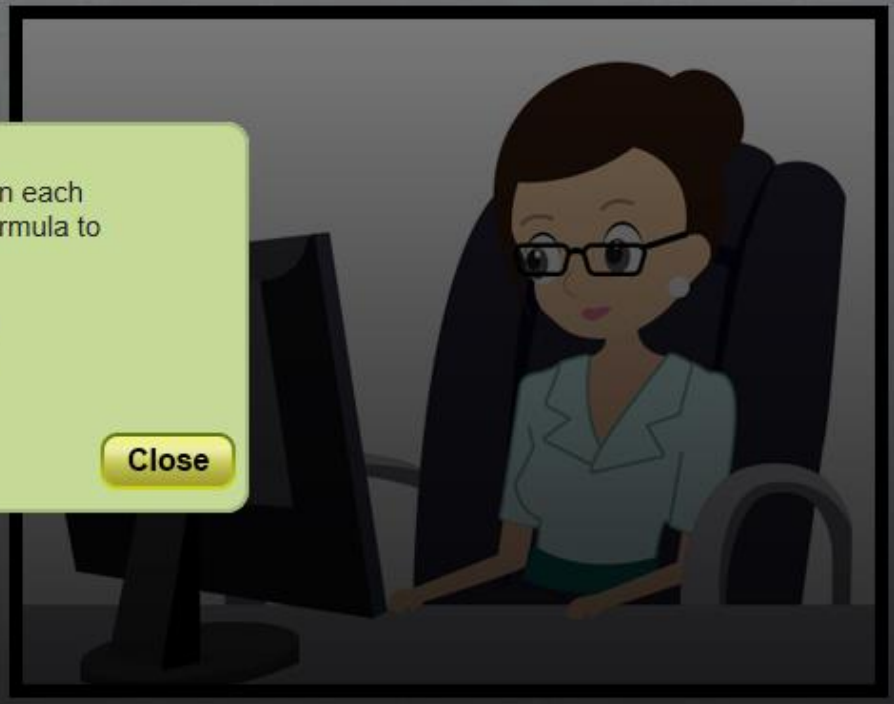
Standard Deviation

Variance (σ^2)

It is the sum of squared difference between each observed data point and the mean. The formula to calculate variance is given below.

$$\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2$$

Close



Anomaly Detection

Step 5: Determining Mean, Variance, and Standard Deviation

After analyzing all the data, Priya is now ready to initiate the process of Anomaly Detection. She needs to determine the Mean, Variance, and Standard Deviation.

Click each button to learn how to determine mean, variance, and standard deviation.

Mean

Variance

Standard Deviation

Standard Deviation (σ)
It is the square root of variance. Standard deviation is indicative of the width of the bell curve.

Close

Step 6: Determining the Probability

Once the mean, variance, and standard deviations are calculated, Priya determines the data point's probability [p(x)] of being within the normal distribution. This probability is determined for each data point in the dataset, based on the mean and standard deviation.

She should use the Gaussian probability density estimate formula to determine the probability. The formula is given below.

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ = Mean
σ = Standard Deviation
$\pi \approx 3.14159$
$e \approx 2.71828$

Step 7: Determining the Final Probability Value

Priya then multiplies the probability of each data point to determine the final probability value. The formula that she uses to determine the final probability is given below.

$$p(x) = p(x_1; \mu_1, \sigma_1^2) * p(x_2; \mu_2, \sigma_2^2) * \ldots * p(x_n; \mu_n, \sigma_n^2)$$

She picks up an Epsilon value. Any probability value less than that value is then flagged as an anomaly. Therefore, an anomaly can be expressed as follows:

$$p(x) < \varepsilon$$

Step 8: Picking the Epsilon Value

After deriving mean, variance, and standard deviation, Priya finds the p(x) value for some of the data points. Her findings are given below:

0.042497326
0.067815344
0.083429987
0.062672099
0.008289868

Based on her findings, she calculated the Epsilon value (ε) as **0.008289868.** She picked up this value as this is the minimum probability value of the observations. This is one strategy to pick up the Epsilon value.

Step 9: Adjusting the Epsilon Value

The next step is to adjust the Epsilon value. To adjust the value, Anu should follow the confusion matrix and select the best output.

Click each button to learn more about the elements of the confusion matrix.

True Positive | Recall | Precision | F1 Score

Depending on the outcome, one can choose to alter the epsilon value by analyzing the Cross-Validation data set, thus improving the overall accuracy.

Anomaly Detection

Step 9: Adjusting the Epsilon Value

The next step is to adjust the Epsilon value. To adjust the value, Anu should follow the confusion matrix and select the best output.

Click each button to learn more about the elements of the confusion matrix.

True Positive | Recall | Precision | F1 Score

True Positive

They are the actual anomalous data that are correctly classified as anomalous data.

Close

Depending on the outcome, one can choose to alter the epsilon value by analyzing the Cross-Validation data set, thus improving the overall accuracy.

Step 9: Adjusting the Epsilon Value

The next step is to adjust the Epsilon value. To adjust the value, Anu should follow the confusion matrix and select the best output.

Click each button to learn more about the elements of the confusion matrix.

True Positive | Recall | Precision | F1 Score

Recall

It is the fraction that is rightly predicted as anomalous among all data points that are actually anomalous.

Formula: Recall = TP/(TP+FN)

Close

Depending on the outcome, one can choose to alter the epsilon value by analyzing the Cross-Validation data set, thus improving the overall accuracy.

Step 9: Adjusting the Epsilon Value

The next step is to adjust the Epsilon value. To adjust the value, Anu should follow the confusion matrix and select the best output.

Click each button to learn more about the elements of the confusion matrix.

True Positive | Recall | Precision | F1 Score

Precision

It is the fraction that is actually anomalous among all the data points predicted as anomalous.

Formula: Precision = TP/(TP+FP)

Close

Depending on the outcome, one can choose to alter the epsilon value by analyzing the Cross-Validation data set, thus improving the overall accuracy.

Anomaly Detection

Step 9: Adjusting the Epsilon Value

The next step is to adjust the Epsilon value. To adjust the value, Anu should follow the confusion matrix and select the best output.

Click each button to learn more about the elements of the confusion matrix.

True Positive | Recall | Precision | F1 Score

F1 Score
The F1 score can be interpreted as a weighted average of the Precision and Recall.

Formula: F1 Score = (2*Precision*Recall)/(Precision + Recall)

Close

Depending on the outcome, one can choose to alter the epsilon value by analyzing the Cross-Validation data set, thus improving the overall accuracy.

Anomaly Detection vs. Supervised Learning

At the outset, Anomaly Detection appears similar to Supervised Learning with the classification model. However, they are not the same.

The differences between Anomaly Detection and Supervised Learning are provided in the table below.

Anomaly Detection	Supervised Learning
It requires a small number of positive examples and a very large number of negative examples.	It requires a large number of both positive and negative examples.
There are many different "types" of anomalies. It is hard for any algorithm to learn from positive examples what the anomalies look like. Future anomalies may look nothing like any of the present anomalous.	The algorithm requires enough positive examples to get a sense of what positive examples are like. Future positive examples are likely to be similar to the ones in the training set.
Anomaly Detection can be used for cases such as fraud detection, manufacturing defects, or monitoring machines in a data center.	Supervised Learning can be used for cases such as email spam classification, weather prediction (e.g. sunny/rainy etc.), or cancer classification.

Summary

Some of the key learning points of the module are mentioned below:

- Anomaly Detection is a method to detect anomaly in a dataset. Anomaly refers to any data point that falls outside the normal distribution.
- Anomaly Detection works best when there is a large number of normal data points, and very few anomalous data points.
- The dataset is first categorized into three sets. These sets are used to calculate the probability density. Finally, the Epsilon value is adjusted using the confusion matrix.
- In Anomaly Detection, first the mean, variance, and standard deviations are determined. Based on the outcome, the probability value is calculated. At last, the final probability value is determined.

Click the Download PDF button to download a PDF version of the course.

Download PDF

Cognizant

Thank You