1. Describe global, contextual, and collective outliers.

A global outlier is a data point that significantly deviates from the rest of the data points. Contextual outliers are data points that deviate based in a selected context. Collective outliers are a group of data points that collectively deviate from the rest of the data.

1. Give a brief description of the differences between supervised, unsupervised, and semi-supervised methods of outlier detection.

Supervised methods use user-labeled examples to determine outliers. We can treat it as a classification problem. Unsupervised methods do not use user-labeled examples; it assumes normal objects are somewhat clustered. Semi-supervised methods use a mixture of both supervised and unsupervised. We can use some labeled objects to help improve unsupervised methods.

1. How does proximity method find outliers?

Proximity method finds outliers by finding the data points that are farther away than the others. It does this using distanced-based and density-based outlier detection.

1. What makes outlier detection more challenging for high-dimensional data?

Outlier detection is more difficult in high-dimensional data because objects become clouded by noise, they’re sparse, and it’s difficult to determine why an outlier is an outlier if detected.

1. Please write any questions you may have.

What is a practical use-case for finding outliers where the data is high-dimensional?