Anomaly Detection with Isolation Forests

Recap from Last session

- Introduction to unsupervised learning
- Introduction to clustering
- Clustering with K Means
- Market segmentation with K Means

Agenda for today

- What is anomaly detection?
- Why it is important?
- Common techniques for anomaly detection
- Isolation forest for anomaly detection
- R implementation of the algorithm on a data set

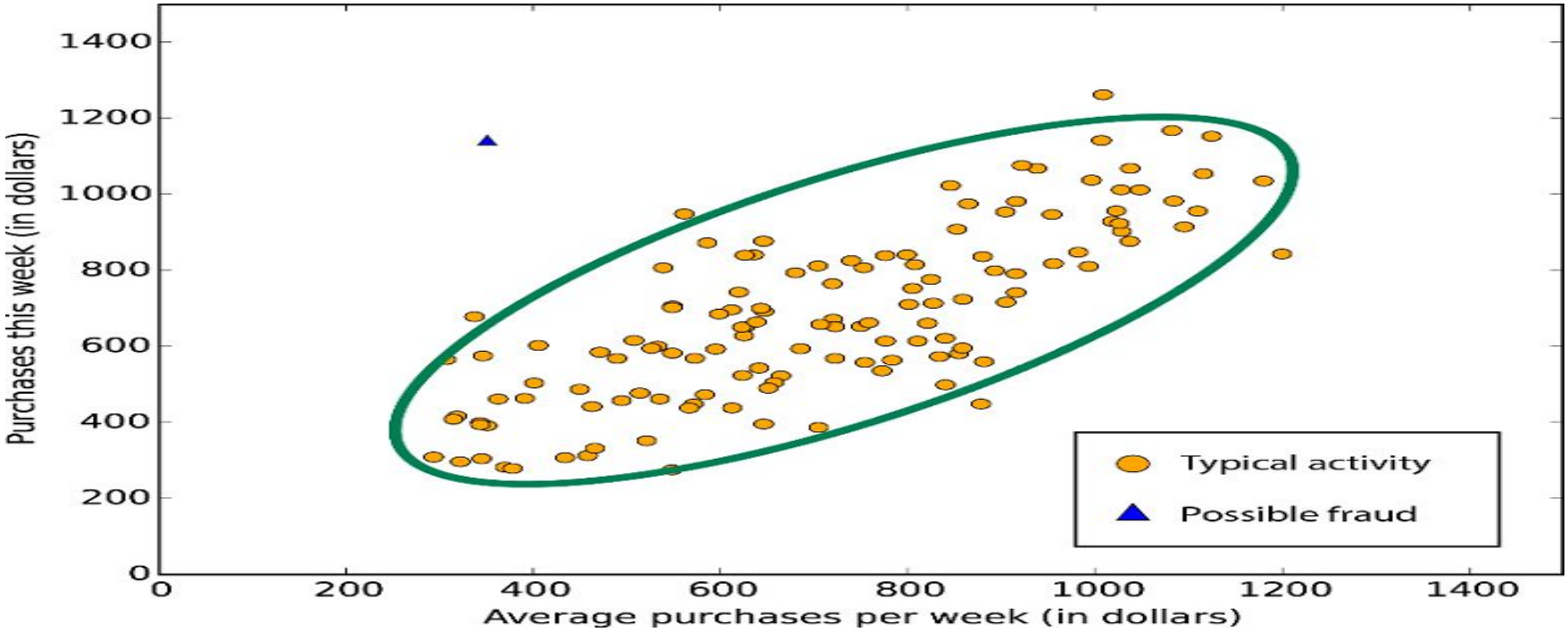
What is Anomaly detection?

- Identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data

- Aberration from normal behavior

- Outliers in higher dimensions are really hard to capture, so we need some mechanism to detect those

What does an anomaly looks like?



Why is it important?

- Anomalous data can be connected to some kind of problem or rare event such as e.g. bank fraud, medical problems, structural defects, malfunctioning equipment etc

- This connection makes it very interesting to be able to pick out which data points can be considered anomalies, as identifying these events are typically very interesting from a business perspective.

Common Techniques

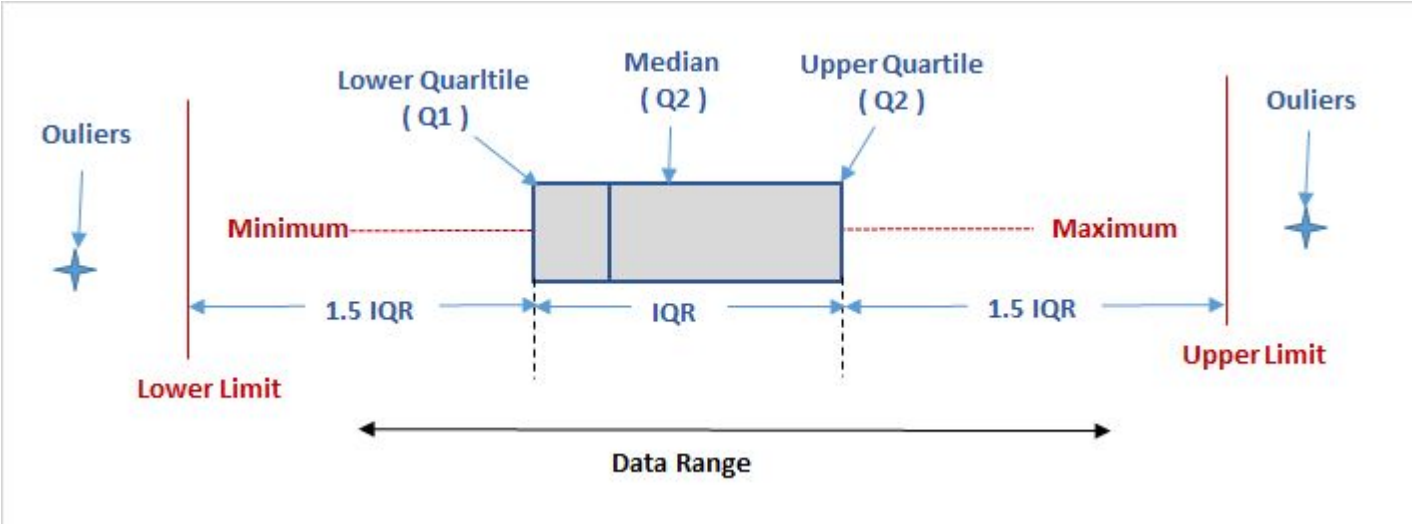
1) Data visualization

- If we have a data set having one or two dimensions, we can easily visualize those.

- Visualizations, such as **box plot** or **scatter plot** can help us identifying the outliers

- **Limitation:** It is only applicable to one and two dimensions.

Box plot in action



2) Dimensionality reduction

- In case of more than two dimensions, we have algorithms such as **PCA** which helps us reduce the dimensions

- In principle, we can reduce the dimensions to 2 variables using PCA and can then visualize them using a box plot to detect anomaly

3) Multivariate data

- Here, we assume that the data is drawn from any known multivariate distribution

- As an example, assuming that the data is drawn from a multivariate normal, we first calculate the estimates of mean and variance from the data, and hence learning the entire distribution

- We then check for the probability for a certain sample. If the probability is less than a certain threshold, we mark them as anomaly.

4) Density based anomaly detection

- It uses approach similar to nearest neighbor to detect outliers. This algorithm is commonly refer to as **Local Outlier Factor**

- It is a calculation that looks at the neighbors of a certain point to find out its density and compare this to the density of other points later on

- If the density of a point is much smaller than the densities of its neighbors (LOF $\gg 1$), the point is far from dense areas and, hence, an outlier.

Isolation Forest

What is Isolation forest?

- It explicitly identifies anomalies instead of profiling normal data. In all previously mentioned approaches, the models profile normal data and then look for aberrations.

- It is almost equivalent to a tree based ensemble

- There are multiple trees which are build in this algorithm, just like how a random forest has multiple decision trees

How Isolation forest works?

- It grows multiple trees as mentioned in the last slide

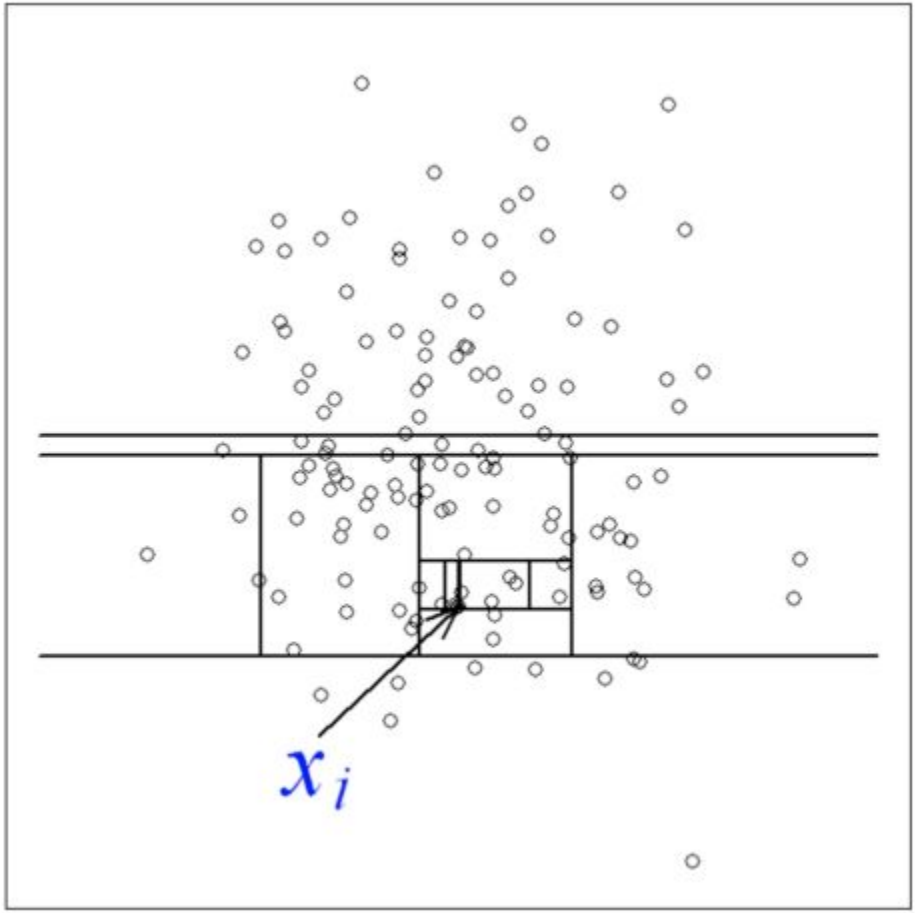
- The tree construction is different from a normal decision tree:
  - To grow a tree, select a **random** feature/variable/column
  - After selecting a random feature, select a **random split** between minimum and maximum of that feature

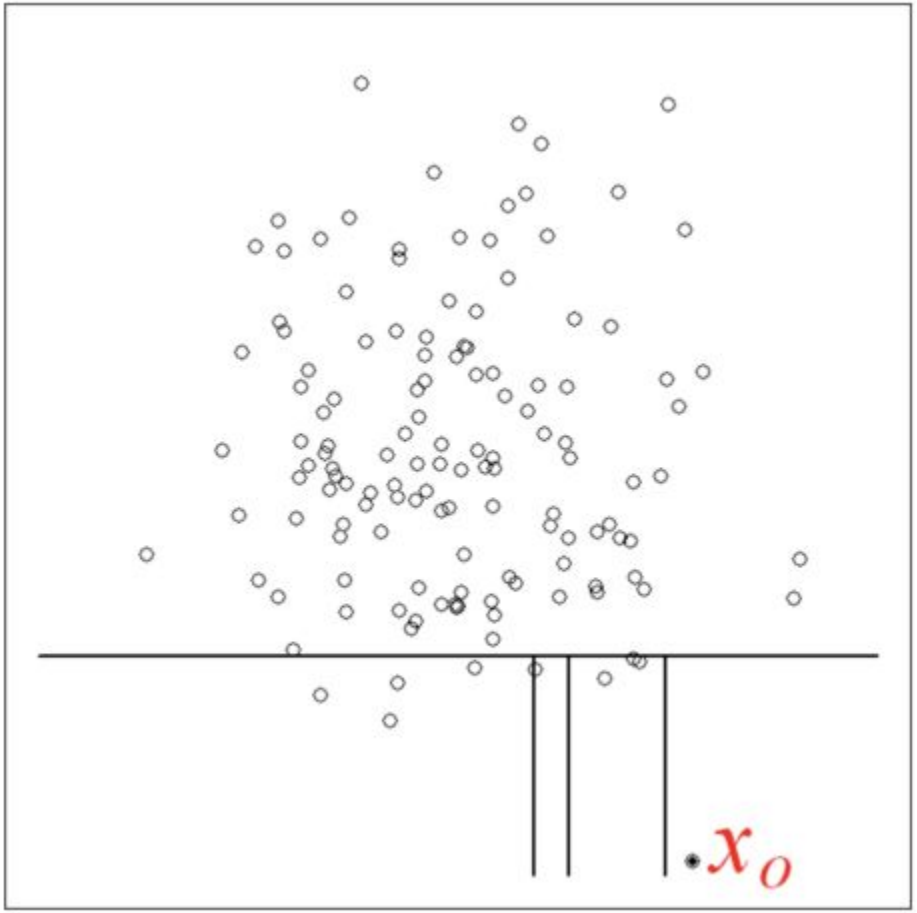
- The samples having **shortest path length** are going to be anomalous and outliers.

Intuition behind Isolation forest

- Anomalies are less frequent and different from normal samples

- Isolating anomaly observations is easier because only a few conditions are needed to separate those cases from the normal observations. On the other hand, isolating normal observations require more conditions.
  - A few condition implies a shorter path in the tree.


x_i


x_o