DATA SCIENCE
MEI/1

University of Beira Interior

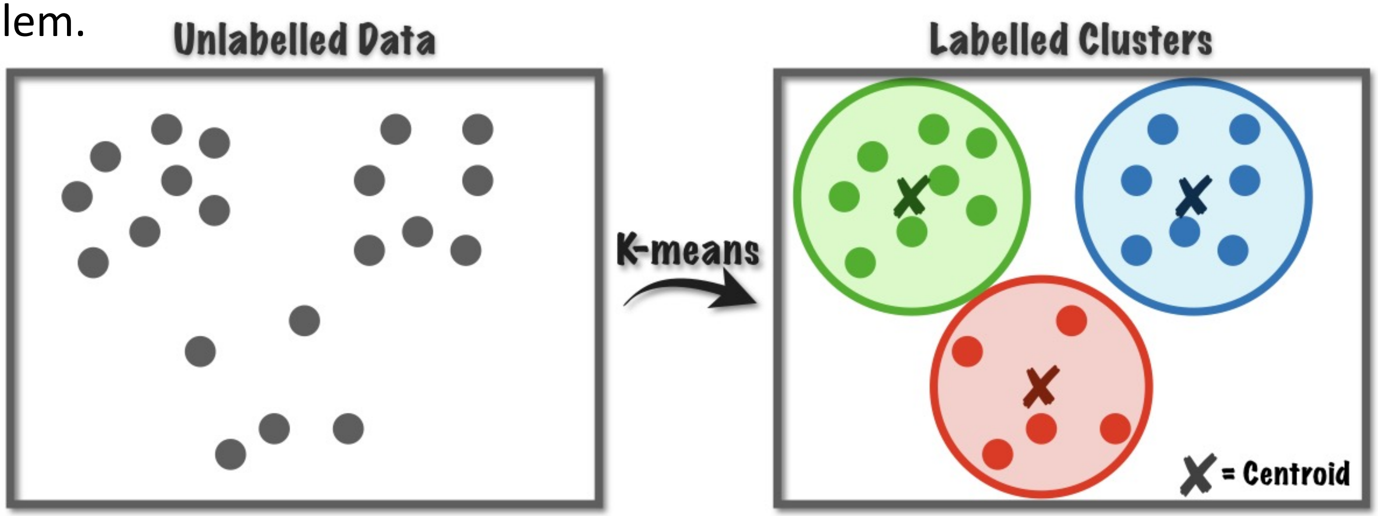
Department of Informatics

Hugo Pedro Proença,

hugomcp@di.ubi.pt, **2021/22**

Data Reduction and Outlier Detection

- When handling “Big Data” problems, it is expected that some of the instances/features are contaminated by noise and should not be considered in the subsequent processing phases (i.e., to build the model/hypothesis).
- Obtaining representative instances of the whole data set (e.g., using clustering techniques) can be quite effective to reduce the amount of data to consider.
  - Use the centroid of each cluster instead of using all instances
  - This can be an extreme solution that looses much important information of the problem.



Data Reduction and Outlier Detection

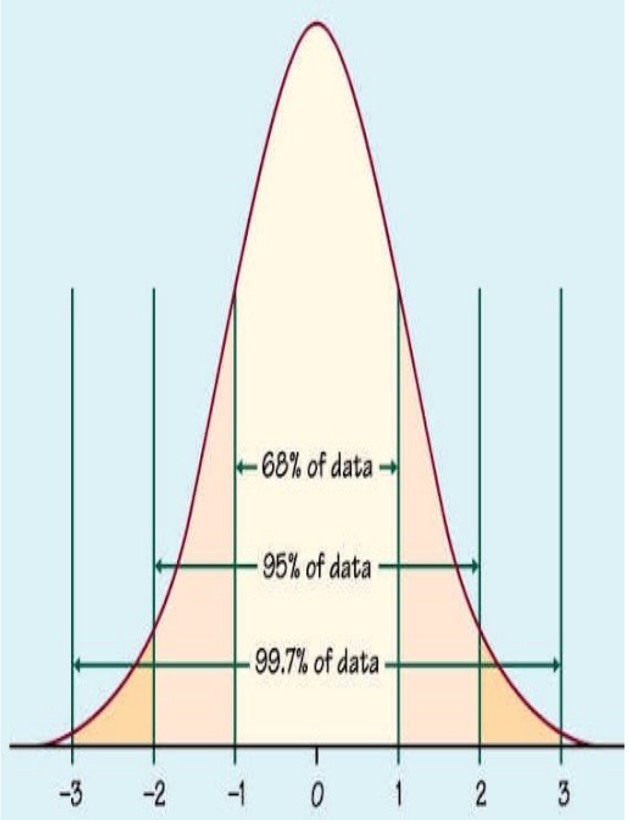
- Instead, a more reasonable solution might be to filter out some of the instances in the dataset, which can be considered not to faithfully represent the data distribution of the remaining elements.
  - These are the outliers, i.e., "*the observation which deviates so much from other observations as to arouse suspicion it was generated by a different mechanism*" (Hawkings, 1980).
- In practice, in outlier is an instance that diverges significantly from the general patterns in the dataset.
- At the coarse level, outliers can be divided into two groups:
  - **Univariate**, which can be found simply by analyzing the distribution of the values in an independent way (e.g., "Name='John Smith', Age=486);
  - **Multivariate**, which can only be found by analyzing the whole feature space, i.e., each feature has a reasonable value, but the whole instance does not.

Data Reduction and Outlier Detection

- Some of the most important methods to detect outliers are:
  - Z-score (parametric method);
  - Probabilistic and Statistical modeling (parametric method);
  - Linear regression models (e.g., PCA)
    - Use the reconstructed (smoothed) versions of the data, considering exclusively the “k” principal coefficients.
  - Proximity-based models (non-parametric);
  - Information theory models;
  - Sparsity analysis methods.

Data Reduction and Outlier Detection

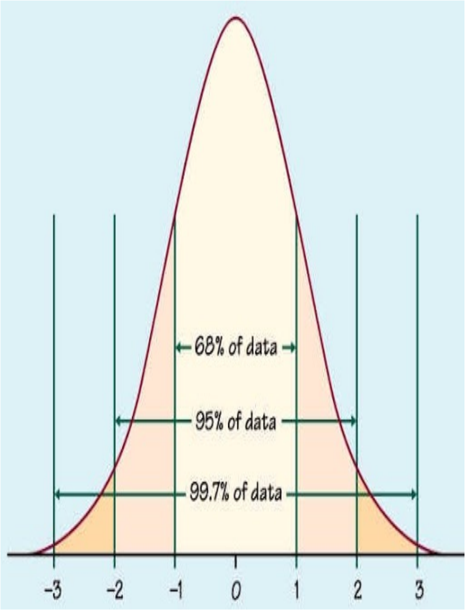
- For univariate analysis, the z-score (a.k.a. standard score) of an instance/feature is a measure of the number of standard deviations of each data point from the overall mean.
  - Most frequently, it assumes a Gaussian distribution for the data
  - In cases where features are not described by Gaussian distributions, data can be transformed (scaled)
- Once the appropriate transformations of the feature space are done, the z-score of an instance is given by:



$$z = \frac{x - \mu}{\sigma}$$

Z-Scores and Boxplot Exercise

- Consider the following values, measuring the height of students in a 5th grade class, given in centimeters.



130	132	138	136	131	153	131	133	129	133	110	132	129	134	135

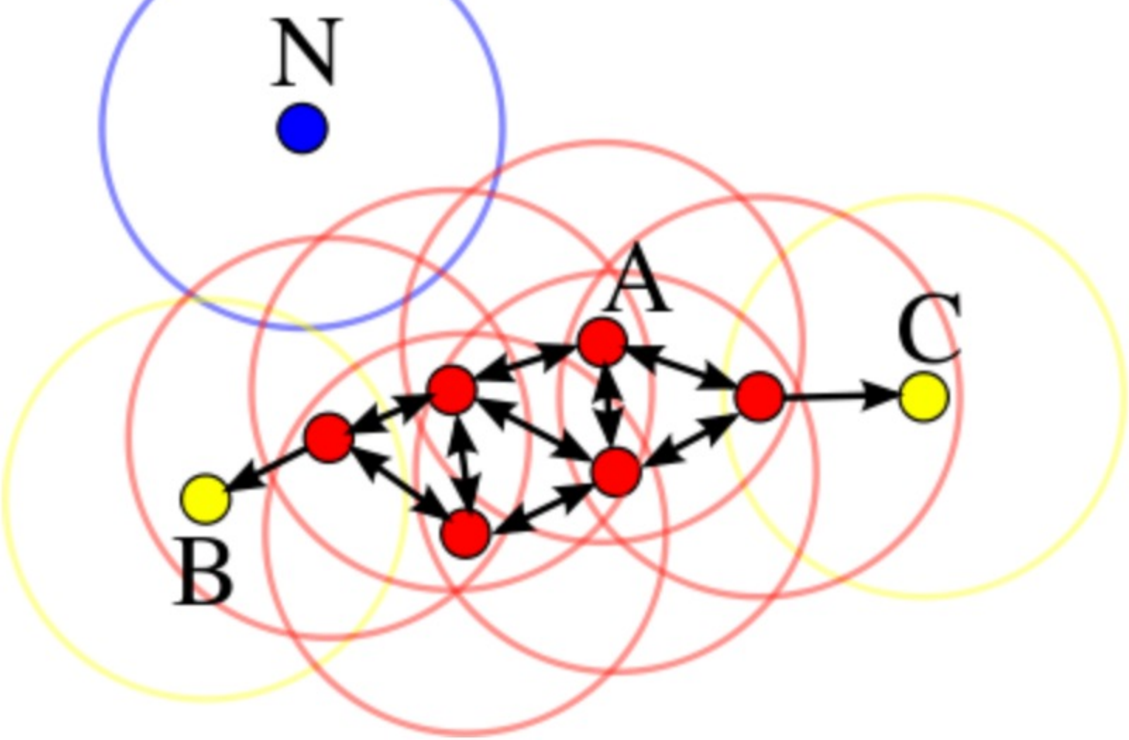
132	135	134	133	132	130	131	134	135	135	134	136	133	133	130

- Assuming that any values with |z-score|>= 3 are outliers, detect the outliers in this set
- Draw the corresponding boxplot.

$$z = \frac{x - \mu}{\sigma}$$

Data Reduction and Outlier Detection

- Density-based spatial clustering of Applications with Noise (DBSCAN)
- This algorithm finds neighbors of each instance on a n-dimensional space, that are closer than a specific radius value (ε)
- The elements in the dataset are then classified into three types:
  - Core Point (e.g., A). When the neighborhood contains at least “k” points
  - Border Point (e.g., C). Points that lie in a cluster, but their neighborhood do not contain “k” points.
  - Outlier (e.g., N). Points that lie in no cluster, i.e., that are not connected to any other.

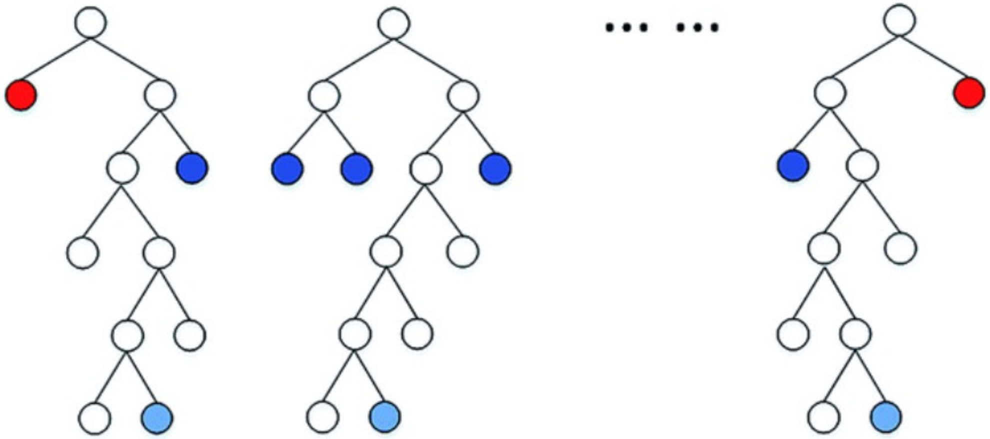


Data Reduction and Outlier Detection

- Any core point will form a cluster with all the points that are reachable from it.
- Here, note that the concept of “reachability” is non-symmetric as, by definition, no point may be reachable from a non-core point.
- Cluster satisfy two properties:
  - All points within the cluster are mutually reachable;
  - If a point is reachable from any point of the cluster, it is part of the cluster also (for boundary points).
- In this algorithm, the number of points “k” and the distance value used (ε) are of most importance to adjust the number of outliers that should be detected.
- Note that, in opposition to other clustering methods, DBSCAN estimates the number of cluster in a fully autonomous way.

Data Reduction and Outlier Detection

- **Isolation Forests** are a relatively novel non-parametric method for large datasets in multi dimensional feature spaces.
- The algorithm is based in groups of binary decision trees.
- The algorithm isolates each instance in the dataset and classifies them into outliers or inliers. This depends on how many steps are required to separate the instance.
  - A non-outlier will be difficult to isolate
  - In opposition, outliers will be much easier to isolate.



Data Reduction and Outlier Detection

- The key concept in Isolation Forests is the **isolation number.**
- The isolation number is the number of splits needed to isolate a data point. This number of splits is obtained by following these steps:
  1. Select the instance to isolate.
  2. For each feature, set the range to isolate between the minimum and the maximum.
  3. Choose a feature randomly.
  4. Pick a value that's in the range, again randomly:
     1. If the chosen value keeps the point *above*, switch the *minimum* of the range of the feature to the value.
     2. If the chosen value keeps the point *below*, switch the *maximum* of the range of the feature to the value.
  5. Repeat steps 3 & 4 until the feature is isolated. That is, until the point is the only one which is inside the range for all features.
  6. Count how many times you've had to repeat steps 3 & 4. This value is the **isolation number.**
- Due to its stochastic nature, the process should be repeated for each instance several times, to assure the stability of results.

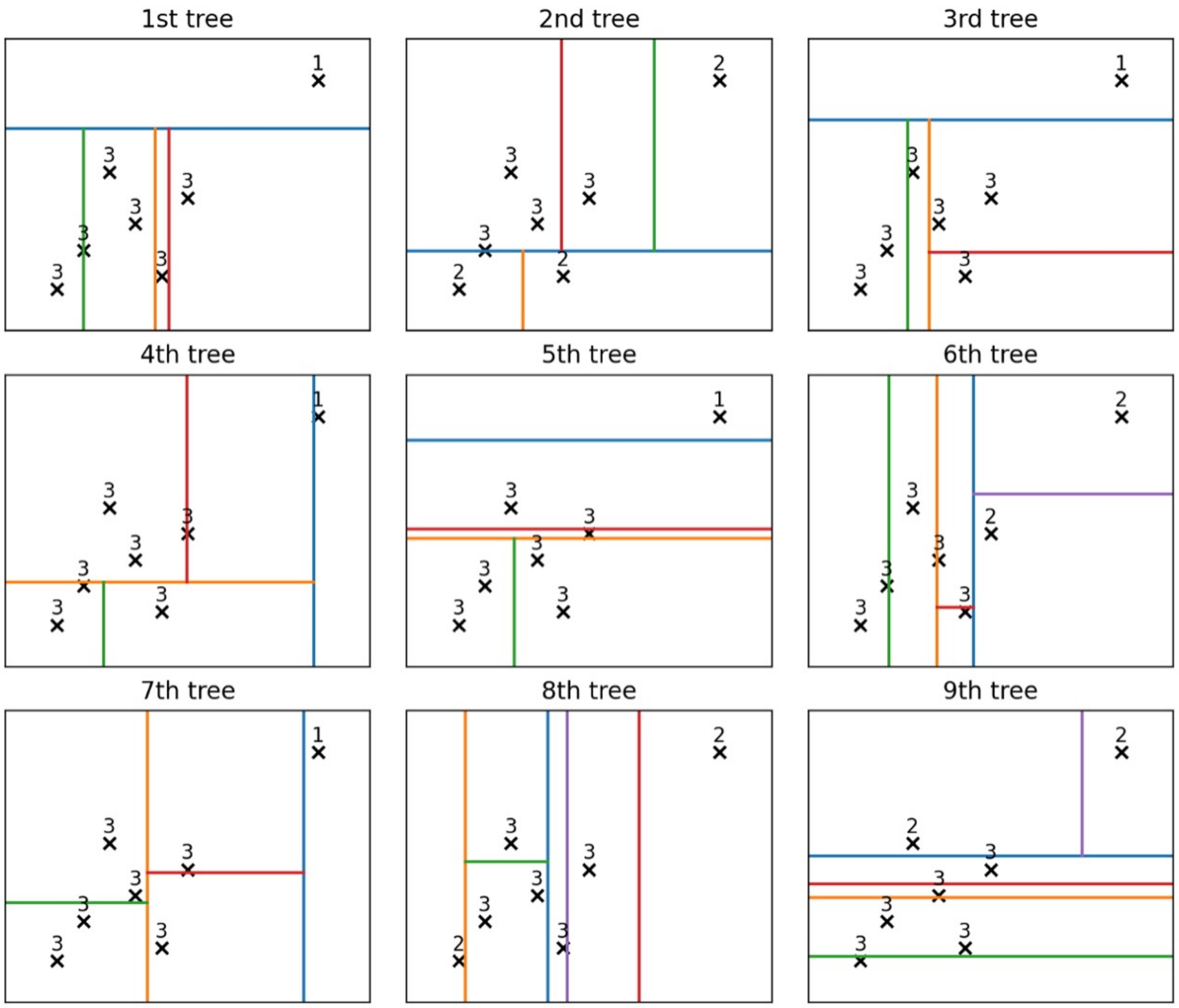
Isolation Forest Exercise

- Consider the following dataset. Use the Isolation Forest algorithm to identify the outliers in the set.

X	Y
1.0	2.1
1.1	2.4
1.2	3.0
1.3	2.6
1.4	2.2
1.5	2.8
2.0	3.7

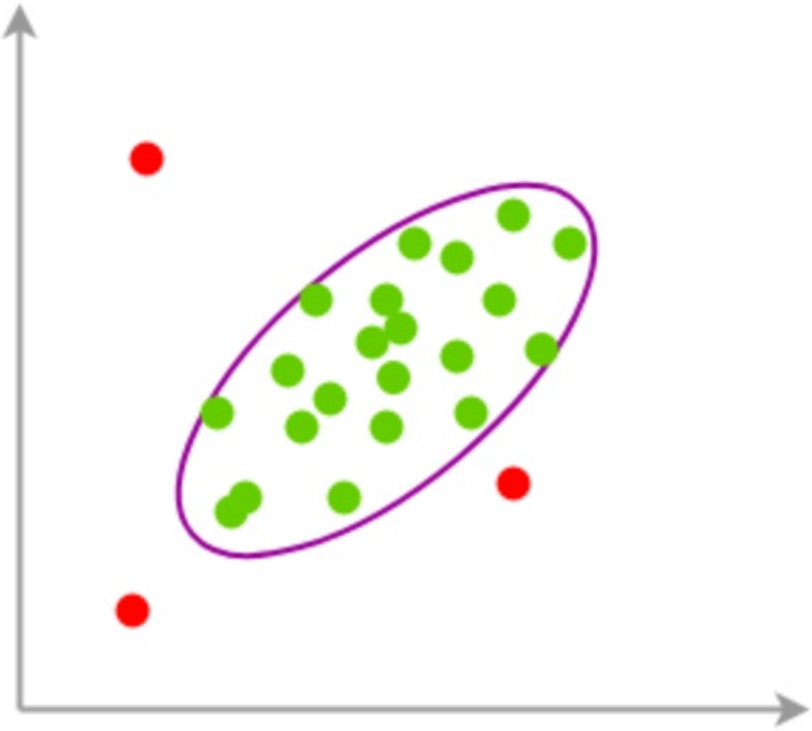
Isolation Forest Exercise (Solution)

- Being intrinsically stochastic, one possible solution for 9 trees would be:



Data Reduction and Outlier Detection

- **Elliptic Envelope**. This is an algorithm with a very simple intuition. We draw an ellipse around the data points.
- However, one constraint is that it also assumes the Gaussian distribution of data.
- For example, if the dataset has “n” features and all are Gaussian, then the feature space forms a multi-dimensional Gaussian
- We define a hypersphere (ellipsoid) that covers the normal data, and data that falls outside this shape is considered an outlier. An efficient implementation of this technique for multivariate data is known as the Minimum Covariance Determinant, or MCD for short.



Data Reduction and Outlier Detection

- MCD is a method for estimating the mean and covariance matrix in a way that tries to minimize the influence of anomalies. The idea is to estimate these parameters from a subset of the data that has been chosen to (hopefully) not contain anomalies.
- More specifically, we start by taking all possible subsets of the data, of a specified size. Estimate the mean and covariance matrix for each subset. Then, keep the estimates for the subset whose covariance matrix has the smallest determinant. The chosen covariance matrix is finally multiplied by a 'consistency factor'.
- The idea behind minimizing the determinant is that the determinant of a covariance matrix measures how broad the distribution is.
- MCD therefore selects the subset of the data that is most tightly distributed. This is to exclude anomalies, which are likely to lie further away from the rest of the data.
- In practice, one can't perform a brute force search over all possible subsets of the data, because there are too many. So, practical MCD algorithms are concerned with how to perform this procedure in a computationally efficient way.