Unsupervised learning
Part 4: Outlier detection

By: Noureddin Sadawi, PhD

University of London

What is an outlier?

- ‘A data value that is very different from most of the data’

  (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

- ‘A person, thing, or fact that is very different from other people, things, or facts, so that it cannot be used to draw general conclusions’

  https://dictionary.cambridge.org/dictionary/english/outlier

Outlier detection vs novelty detection

- **Outlier detection:** The training data contains outliers which are defined as observations that are far from the others. Outlier detection estimators thus try to fit the regions where the training data is the most concentrated, ignoring the deviant observations
- **Novelty detection:** The training data is not polluted by outliers and we are interested in detecting whether a **new** observation is an outlier. In this context an outlier is also called a novelty

  https://scikit-learn.org/stable/modules/outlier_detection.html

Anomaly detection

- ‘Outlier detection and novelty detection are both used for anomaly detection, where one is interested in detecting abnormal or unusual observations.’
- ‘Outlier detection is then also known as unsupervised anomaly detection and novelty detection as semi-supervised anomaly detection.’

  https://scikit-learn.org/stable/modules/outlier_detection.html

Anomaly detection

- ‘In the context of outlier detection, the outliers/anomalies cannot form a dense cluster as available estimators assume that the outliers/anomalies are located in low density regions.’
- ‘On the contrary, in the context of novelty detection, novelties/anomalies can form a dense cluster as long as they are in a low density region of the training data, considered as normal in this context.’

  https://scikit-learn.org/stable/modules/outlier_detection.html

Outlier detection using STD

- If data follows a normal distribution (i.e. Gaussian distribution) then its standard deviation can be used as a good reference for identifying outliers.
- Usually:
  - 68% of the data is within 1 STDs from the mean
  - 95% of the data is within 2 STDs from the mean
  - 99.7% of the data is within 3 STDs from the mean.
- A common practical method to identify outliers is to use 3 STDs from the mean as a cut-off point (for normally distributed data).

Isolation forest

- ‘…”isolates” observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.’
- 'Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node.’

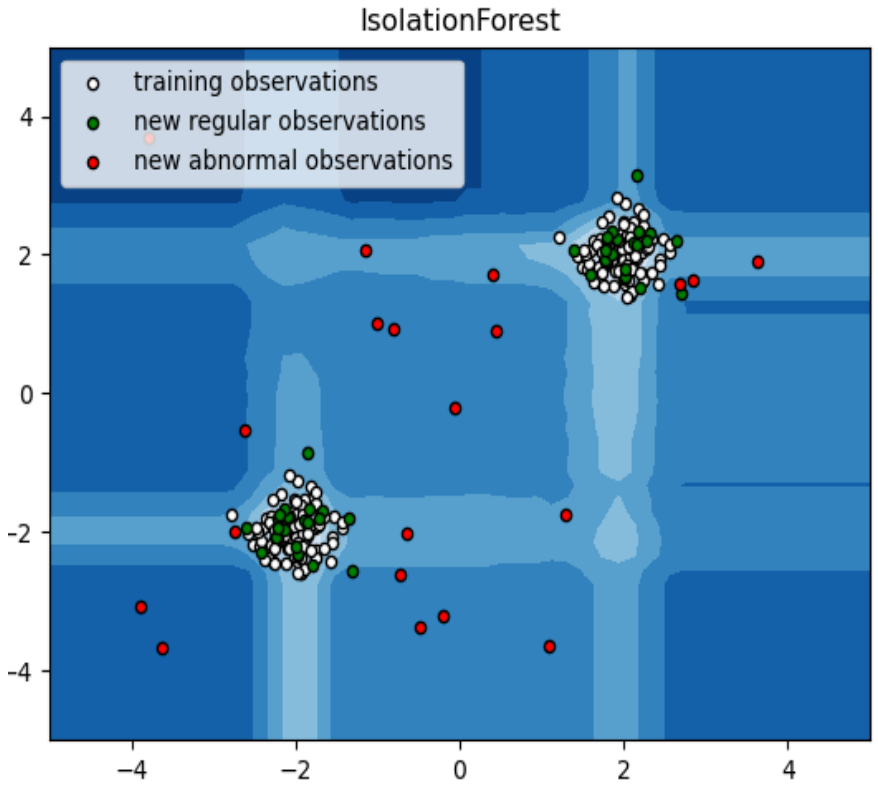
  https://scikit-learn.org/stable/modules/outlier_detection.html

Isolation forest

- **‘This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.’**
- ‘Random partitioning produces noticeably shorter paths for anomalies.’
- ‘Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.’

  https://scikit-learn.org/stable/modules/outlier_detection.html

Isolation forest



https://scikit-learn.org/stable/modules/outlier_detection.html

Good explanations

- A walk through the isolation forest, by Jan van der Vegt @ PyData 2019. https://youtu.be/RyFQXQf4w4w
- ‘Unsupervised anomaly Detection with isolation forest’ by Elena Sharova @ PyData 2018. https://youtu.be/5p8B2Ikcw-k
- ‘Anomaly detectcion: algorithms, explanations, applications’ by Thomas Dietterich, Anomaly Detection @ Microsoft Research 2018. https://youtu.be/12Xq9OLdQwQ