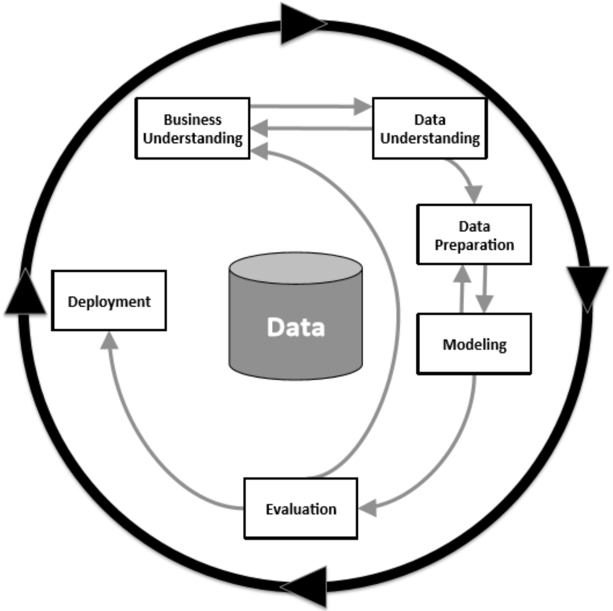
Advanced Data & Network Mining

Modelling - Unsupervised Learning
Anomaly Detection

2023-24

terri.hoare@dbs.ie

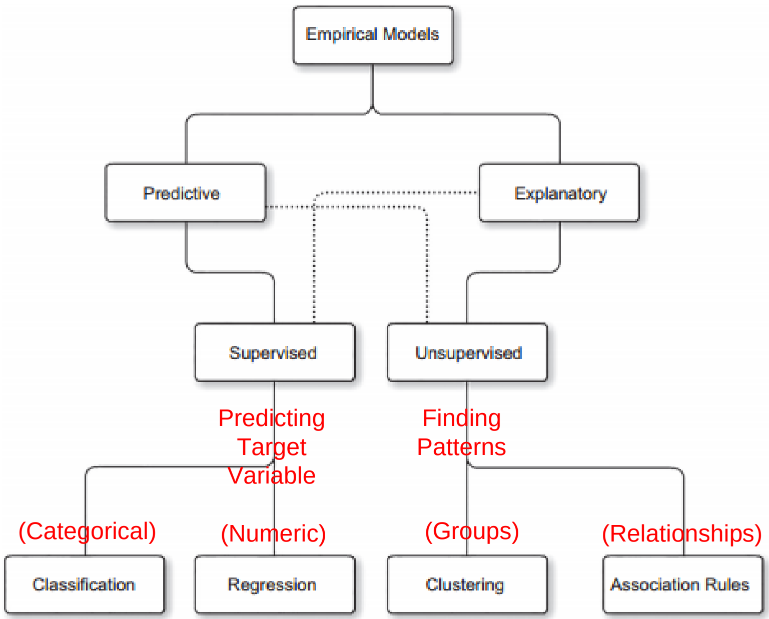
Recap on Challenges and Methodology
CRISP-DM (Cross Industry Standard Process for Data Mining)



The CRISP-DM process, including the six key phases and the important relationships between them (adapted from Wirth and Hipp, 2000, repr. in Kelleher *et al.*, 2020, p.14)

Teaching materials 'Data to Insights to Decisions' (Kelleher, MacNamee and D'Arcy , 2015)

Data Mining
Taxonomy of Algorithms



Classification:

K-NN
Decision Trees
Rule Induction
Naïve Bayes
ANN + DeepLearning
SVM + FLM
Ensemble – RF + GBT

Regression:

Linear Reg.
GLM
Logistic Reg.

Clustering:

K-means
X-Means
K-Medoids
DBSCAN
SOMs
Anomaly-Distance
Anomaly-Density
Anomaly-Local Outlier factor

Association Rules:

Apriori
FP Growth
Collaborative Filtering
Content Based Filtering
Graph

Anomaly Detection
Concepts

An outlier is a data object that is markedly different from the other objects in the data set.

A high-income individual may be an outlier in a middle class neighbourhood data set, but not in the ownership of luxury vehicles data set. Outliers are rare and hence they stand out amongst the other data points. For example, the majority of computer network traffic is legitimate and the one malicious network attack would be the outlier.

Teaching materials ‘Data Science Concepts and Practice’ (Kotu, Deshpande, 2018)

Anomaly Detection
Causes of Outliers

- **Data errors**

Measurement errors, human errors, data collection errors
Outlier detection used as a routine pre-processing step for algorithms such as regression and neural networks

- **Normal variance in the data**

In a normal distribution, 99.7% of data points lie within three standard deviations from the mean. In other terms 0.26% or 1 in 370 data points lie outside three standard deviations from the mean. By definition, while they don’t occur frequently, they are legitimate data for example someone more than 7 feet tall in a human height data set

- **Data from other distribution classes**

The number of daily page views for a customer facing website from a user IP address usually range from one to a few dozens. Although an outlier, it is quite “normal” for bots registering thousands of page view calls to a website. All bot traffic falls under distribution of a different class “traffic from programs” rather than traffic from regular browsers

Anomaly Detection
Causes of Outliers cont.

- **Distributional assumptions**

Outlier data points can originate from incorrect assumptions on the data distribution, for example, data measuring the use of a school library will show an outlier as a result of a surge in use during exams. Similarly there will be a surge of sales on St Stephen's day in Ireland. Outliers in these cases are expected

Understanding why outliers occur is important in determining the action to be taken.

For credit card transaction fraud monitoring (high frequency, high amounts, large geographic separation between points on consecutive transactions) needs to be isolated and the credit card customer contacted immediately to verify the authenticity of the transaction

In other cases we need to filter out the outliers during data set pre processing because they skew the final outcome and we are looking to generalize conclusions

Anomaly Detection

Case Study : Detecting Click Fraud in Online Advertising

The rise in online advertising has underwritten many successful Internet business models and enterprises. Online advertisements make free Internet services like web searches, news content, social networks, mobile application, and many other services viable. One of the key challenges in online advertisements is mitigating click frauds. Click fraud is a process where an automated program or a person imitates the action of a normal user clicking on an online advertisement, with the malicious intent of defrauding the advertiser, publisher, or advertisement network. Click fraud could be performed by contracting parties or third parties, like competitors trying to deplete advertisement budgets or to tarnish the reputation of the sites. Click fraud distorts the economics of advertising and poses a major challenge for all parties involved in online advertising. Detecting, eliminating, or discounting click fraud makes the entire marketplace trustworthy and even provides competitive advantage for all the parties.

Detecting click frauds takes advantage of the fact that fraudulent traffic exhibits an atypical web browsing pattern when compared with typical clickstream data.

Teaching, 2018)

Anomaly Detection
Case Study : Detecting Click Fraud in Online Advertising

Fraudulent traffic often does not follow a logical sequence of actions and contains repetitive actions that would differentiate from other regular traffic. For example, most of the fraudulent traffic exhibits either one or many of following characteristics: they have very high click depth (number of web pages accessed deep in the website); the time between each click would be very low; a single session would have a high number of clicks on advertisements as compared with normal user; the originating IP address would be different from the target market of the advertisement; there would be very little time spent on advertiser's target website; etc. It is not one trait that differentiates fraudulent traffic from regular traffic, but the combination of the traits. Detecting click fraud is an ongoing and evolving process. Increasingly the click fraud perpetuators are getting sophisticated in imitating the characteristics of a normal web browsing user. Hence, click fraud cannot be fully eliminated; however it can be contained by constantly developing new algorithms to identify fraudulent traffic.

Anomaly Detection
Case Study : Detecting Click Fraud in Online Advertising

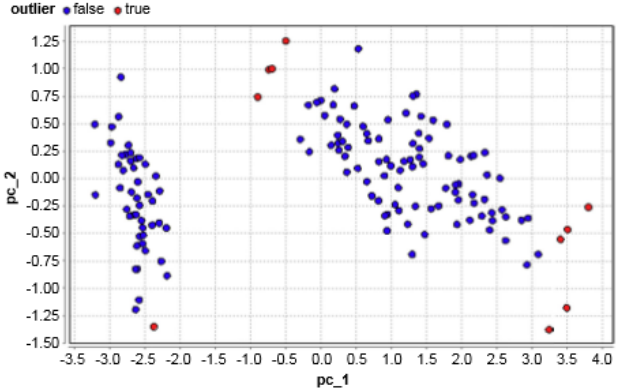
To detect click fraud outliers, first we need to prepare clickstream data in such a way that detection using data mining is easier. A relational column-row data set can be prepared with each visit occupying each row and the columns being traits like click depth, time between each clicks, advertisement clicks, total time spent in target website, etc. This multidimensional data set can be used for outlier detection using data mining. Clickstream traits or attributes need to be carefully considered, evaluated, transformed, and added in the data set. In multidimensional data space, the fraudulent traffic (data point) is distant from other visit records because of their attributes, such as number of ad clicks in a session. A regular visit usually has one or two ad clicks in a session, while a fraudulent visit would have dozens of ad clicks. Similarly, other attributes can help in identifying the outlier more precisely.

Anomaly Detection
Distance Based

Outlier identified based on distance if kth nearest neighbour

Learn Examples 11_Oulier_11.1_distance.rmp

- Assigns a distance score for each object that is the distance to the kth nearest data object. Three parameters specified for RapidMiner *Detect Outlier (Distances)* operator:- *K*, number outliers and *distance measure* (default Euclidean). A Boolean outlier attribute is added to Results which can be visualised as the third-dimension colour on the scatter plot chart type



Data set with outliers

Teaching materials 'Data Science Concepts and Practice' (Kotu, Deshpande, 2018)

Anomaly Detection
Distance Based Summary

Model
All data points are assigned a distance score based on nearest neighbour

Input
Accepts both numeric and categorical attributes. Normalization is required since distance is calculated

Output
Every data point has a distance score. The higher the distance, the more likely the data point is an outlier

Pros
Easy to implement. Works well with numeric attributes

Cons
Specification of k is arbitrary

Use Cases
Fraud detection, pre-processing technique

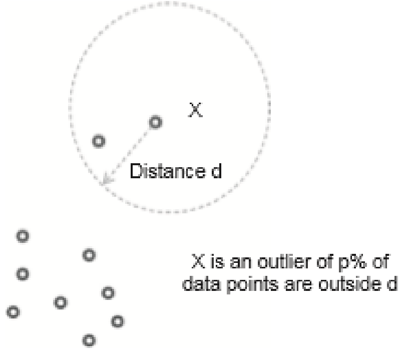
Teaching materials ‘Data Science Concepts and Practice’ (Kotu, Deshpande, 2018)

Anomaly Detection
Density Based

Outlier is identified based on data points in low density regions

Learn Examples 11_Oulier_11.2_oulier_density.rmp

A point X is considered an outlier if at least p fraction (must be set greater than 35%) of points lie more than a distance d from the point. Three parameters specified for RapidMiner **Detect Outlier (Density)** operator:- distance **d** (default 1), proportion **p** (default 95%) and **distance measure** (default Euclidean).



Outlier detection based on distance and propensity.

Anomaly Detection
Density Based Summary

Model
All data points are assigned a density score based on the neighbourhood

Input
Accepts both numeric and categorical attributes. Normalization is required since density is calculated

Output
Every data point has a density score. The lower the density, the more likely the data point is an outlier

Pros
Easy to implement. Works well with numeric attributes

Cons
Specification of distance parameter by the user. Inability to identify varying density regions

Use Cases
Fraud detection, pre-processing technique

Anomaly Detection
Local outlier factor

Outlier is identified based on calculation of relative density in the neighbourhood

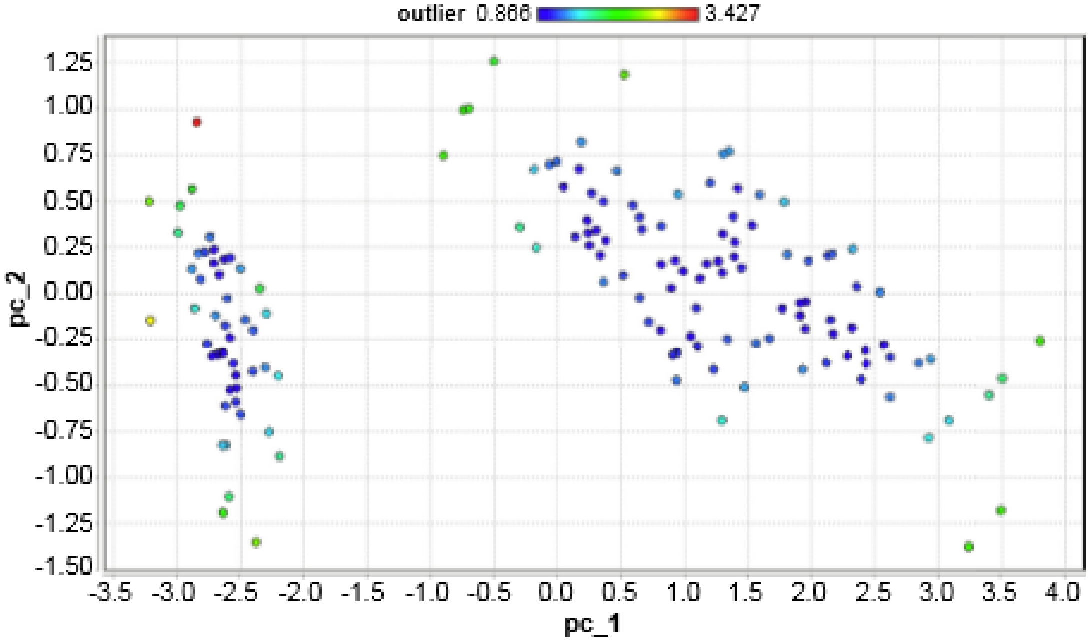
Learn Examples 11_Oulier_11.3__lof.rmp

A variation of density-based outlier detection addressing one of its key limitations namely detecting the outliers in varying density.

LOF takes into account both the density of the data point and the density of the neighbourhood of the data point. By comparing the density of the data point and density of all the data points in the neighbourhood, we can determine if the density of the data point is lower than the density of the neighbourhood. This scenario indicates an outlier.

Anomaly Detection
Local Outlier Factor Results cont.

Visually (colour set to outlier score) a point closer to red is considered an outlier



Output of LOF outlier detection

Teaching materials 'Data Science Concepts and Practice' (Kotu, Deshpande, 2018)

Anomaly Detection
Local Outlier Factor Summary

Model
All data points are assigned a relative density score based on the neighbourhood

Input
Accepts both numeric and categorical attributes. Normalization is required since density is calculated

Output
Every data point has a density score. The lower the density, the more likely the data point is an outlier

Pros
Can handle varying density scenario

Cons
Specification of distance parameter by the user.

Use Cases
Fraud detection, pre-processing technique

Data Mining
Anomaly Detection Discussion Points

- In theory any classification algorithm can produce a generalised model to detect an outlier if previously classified (labelled) data is available. However, as the probability of an outlier is very low , say less than 0.1%, the model can just predict the class as regular for all the data points and still be 99.9% accurate! This method clearly does not work for outlier detection since the Recall measure is 0%/ Balancing needed.

- In many practical applications like detecting network intrusion or fraud prevention in high volume transaction networks, the cost of not detecting an outlier is very high. The model can even have an acceptable number of false alarms, that is labelling a regular data point as an outlier.

Data Mining
Anomaly Detection Discussion Points cont.

- In practical applications, outlier detection models need to be updated frequently as the characteristics of an outlier changes over time and hence the relationship between outliers and normal records changes as well.

- Outlier detection one of the most profound applications of data mining that impacts the majority of the population through transaction monitoring, fraud prevention, and early identification of anomalous activity in the context of security.

Teaching materials ‘Data Science Concepts and Practice’ (Kotu, Deshpande, 2018)

Data Mining
Anomaly Detection H2O Isolation Random Forest

H2O Documentation (2023). Isolation Random Forest.
Available at:
https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/if.html
(Accessed: 12 October 2023).

https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/if.html (Accessed: 12 October 2023).

Data Mining
Anomaly Detection H2O Isolation Random Forest

There are multiple approaches to an unsupervised anomaly detection problem that try to exploit the differences between the properties of common and unique observations.

Build multiple decision trees such that the trees isolate the observations in their leaves. Ideally, each leaf of the tree isolates exactly one observation from your data set. The trees are being split randomly. We assume that if one observation is similar to others in our data set, it will take more random splits to perfectly isolate this observation, as opposed to isolating an outlier.

For an outlier that has some feature values significantly different from the other observations, randomly finding the split isolating it should not be too hard. As we build multiple isolation trees, hence the isolation forest, for each observation we can calculate the average number of splits across all the trees that isolate the observation. The average number of splits is then used as a score, where the less splits the observation needs, the more likely it is to be anomalous.

https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/if.html (Accessed: 12 October 2023).

References

- Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). *LOF: Identifying density-based local outliers*. In Proceedings of the ACM SIGMOD 2000 international conference on management of data (pp. 112).
- Haddadi, H. (2010). *Fighting online click-fraud using bluff ads*. ACM SIGCOMM Computer Communication Review, 40(2), 2125.
- Knorr, E. M., & Ng, R. T. (1998) *Algorithms for mining distance-based outliers in large datasets*. In Proceedings of the 24th VLDB conference (pp. 392403). New York, USA.
- RapidMiner Extension: Anomaly Detection. (2014). German research center for artificial intelligence. DFKI GmbH. Retrieved from ,http://madm.dfki.de/rapidminer/anomalydetection.

References

- Sadagopan, N., & Li, J. (2008) *Characterizing typical and atypical user sessions in clickstreams.* In Proceeding of the 17th international conference on World Wide Web—WWW ‘08 885. ,https://doi.org/10.1145/1367497.1367617.
- Sadik, S., & Gruenwald, L. (2013). *Research issues in outlier detection for data streams*. ACM SIGKDD Explorations Newsletter, 15(1), 3340.
- Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Anomaly detection*. Introduction to data mining (pp. 651676). Boston, MA: Addison Wesley.