Data Engineering

204426

Data Cleaning

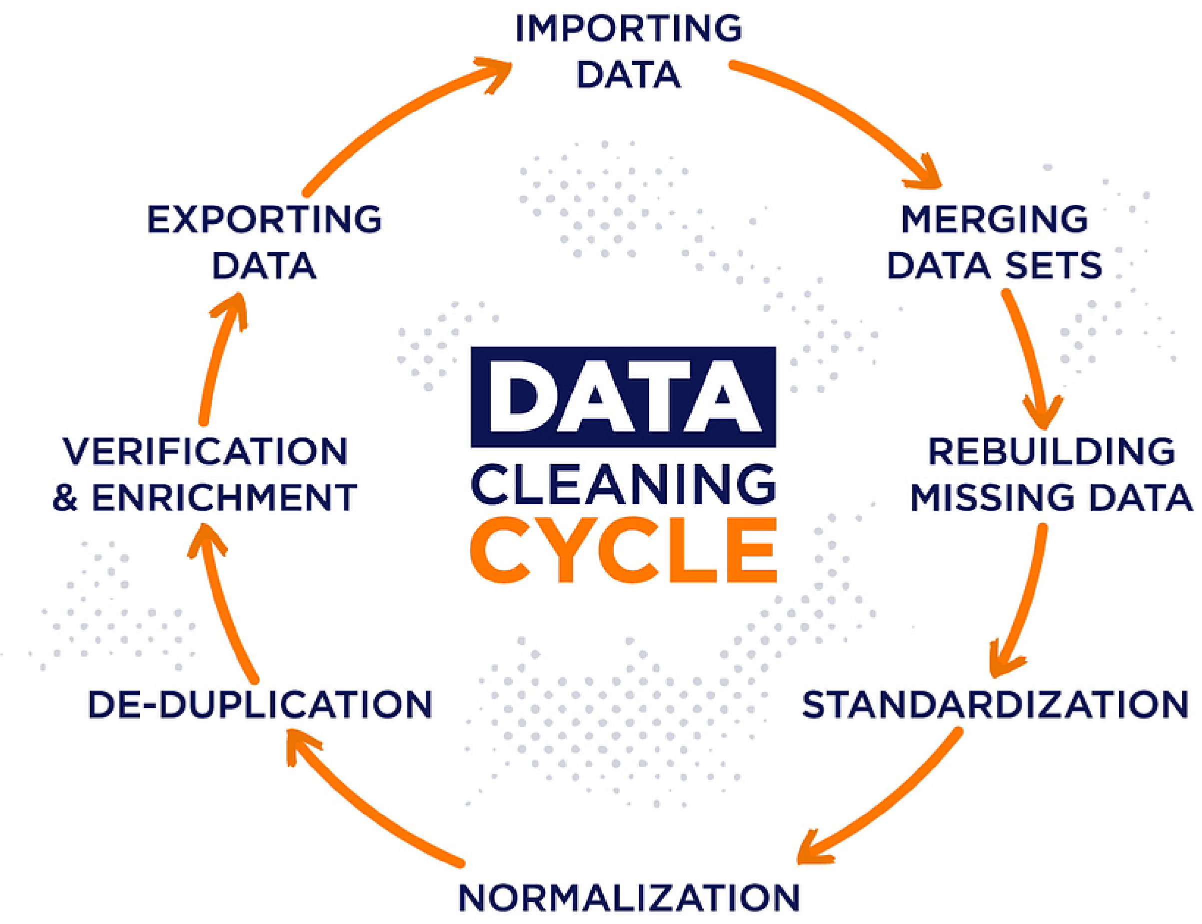
Outline

- Data Cleaning
- Data Cleaning Process
- Handle Missing Data
  - Listwise deletion
  - Pairwise deletion
  - Dummy variable adjustment
  - Mean, Mode, Median imputation
  - Regression imputation
  - Stochastic regression imputation
- Outlier Detection
  - Extreme Value Analysis
  - DBSCAN
  - Isolation forest

Data Cleaning

- The process of ensuring data is correct, consistent and usable.
- You can clean data by <u>identifying</u> errors or corruptions, <u>correcting</u> or <u>deleting</u> them, or manually processing data as needed to prevent the same errors from occurring.

Data Cleaning



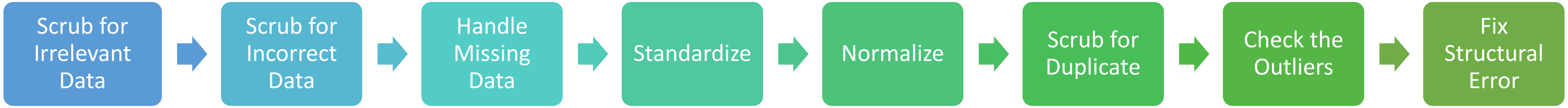
Data Cleaning Process – 5 Steps To Ensure Clean Data

1. **Data Audit**
    - Determine what kind of errors your data set contains and where they're located.
    - the use of statistical and database methods that help you detect anomalies and contradictions.

Data Cleaning Process – 5 Steps To Ensure Clean Data

2. Workflow Execution

- Specify what operations are a part of the sequence that cleans the data sets.
- A typical data cleaning workflow:



Data Cleaning Process – 5 Steps To Ensure Clean Data

3. Data Cleaning (Workflow Execution)

- The cleaning stage is the execution of operations specified in the workflow.
- The data cleaning process might feature different techniques relative to the project’s nature and the data type.
- But the final objective is always the same – removal or correction of data.

Data Cleaning Process – 5 Steps To Ensure Clean Data

4. Validation

- Audit the data again and make sure all the rules and constraints were in fact executed.
- You should consider the following questions:
  - What conclusions can you draw from the dataset?
  - Does it prove or disprove your hypothesis?
  - Are there any insights that help you form the next idea?

Data Cleaning Process – 5 Steps To Ensure Clean Data

5. Reporting

- Creating reports and summaries of the data cleaning is essential as far as streamlining and efficiency goes.
- Especially if you're processing a lot of data and working with many people.
- Reports allow you and your co-workers to compare findings and access the insights quickly and effortlessly.

Handle Missing Data

Listwise deletion

- Discards the data for any case that has one or more missing values.
- Advantages:
  - Can be applied to any statistical test (SEM, multi-level regression, etc.)
  - In the case of MCAR, both the parameters estimates and its standard errors are unbiased.
  - In the case of MAR among independent variables, listwise deletion parameter estimates can still be unbiased.
    - $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
    - the probability of missing data on X_1 is independent of y
    - but dependent on the value of X_1 and X_2
    - the model estimates are still unbiased.

Handle Missing Data

Listwise deletion

- Disadvantages:
  - It will yield a larger standard errors than other more sophisticated methods
  - If the data are not MCAR, but MAR, then your listwise deletion can yield biased estimates.
  - In other cases than regression analysis, other sophisticated methods can yield better estimates compared to listwise deletion.

Handle Missing Data

Pairwise deletion

- Only deletes cases when one of the variables in the particular model you are evaluating is missing.
- This method could only be used in the case of linear models such as linear regression, factor analysis, or SEM.
- The premise of this method based on that the coefficient estimates are calculated based on the means, standard deviations, and correlation matrix.

Handle Missing Data

Pairwise deletion

- Advantages:
  - If the true missing data mechanism is MCAR, pair wise deletion will yield consistent estimates, and unbiased in large samples
  - If the correlation among variables are low, pairwise deletion is more efficient estimates than listwise
- Disadvantages:
  - If the correlations among variables are high, listwise deletion is more efficient than pairwise.
  - If the data mechanism is MAR, pairwise deletion will yield biased estimates.
  - In small sample, sometimes covariance matrix might not be positive definite, which means coefficients estimates cannot be calculated.

Handle Missing Data

Dummy Variable Adjustment

- Add another variable in the database to indicate whether a value is missing.
- In a regression predicting Y, suppose there is missing data on a predictor X.
  - Create a new variable D=1 if X is missing and D=0 if X is present.
  - When X is missing, set X*=c where c is some constant (e.g., the mean of X).
  - Regress Y on both X* and D (and any other variables)

Handle Missing Data

y	x
11	0.3
15	1.0
10	
…	….
8	0.1

y	x*	d
11	0.3	0
15	1.0	0
10	0.6	1
…	….	…
8	0.1	0

Handle Missing Data

Dummy Variable Adjustment

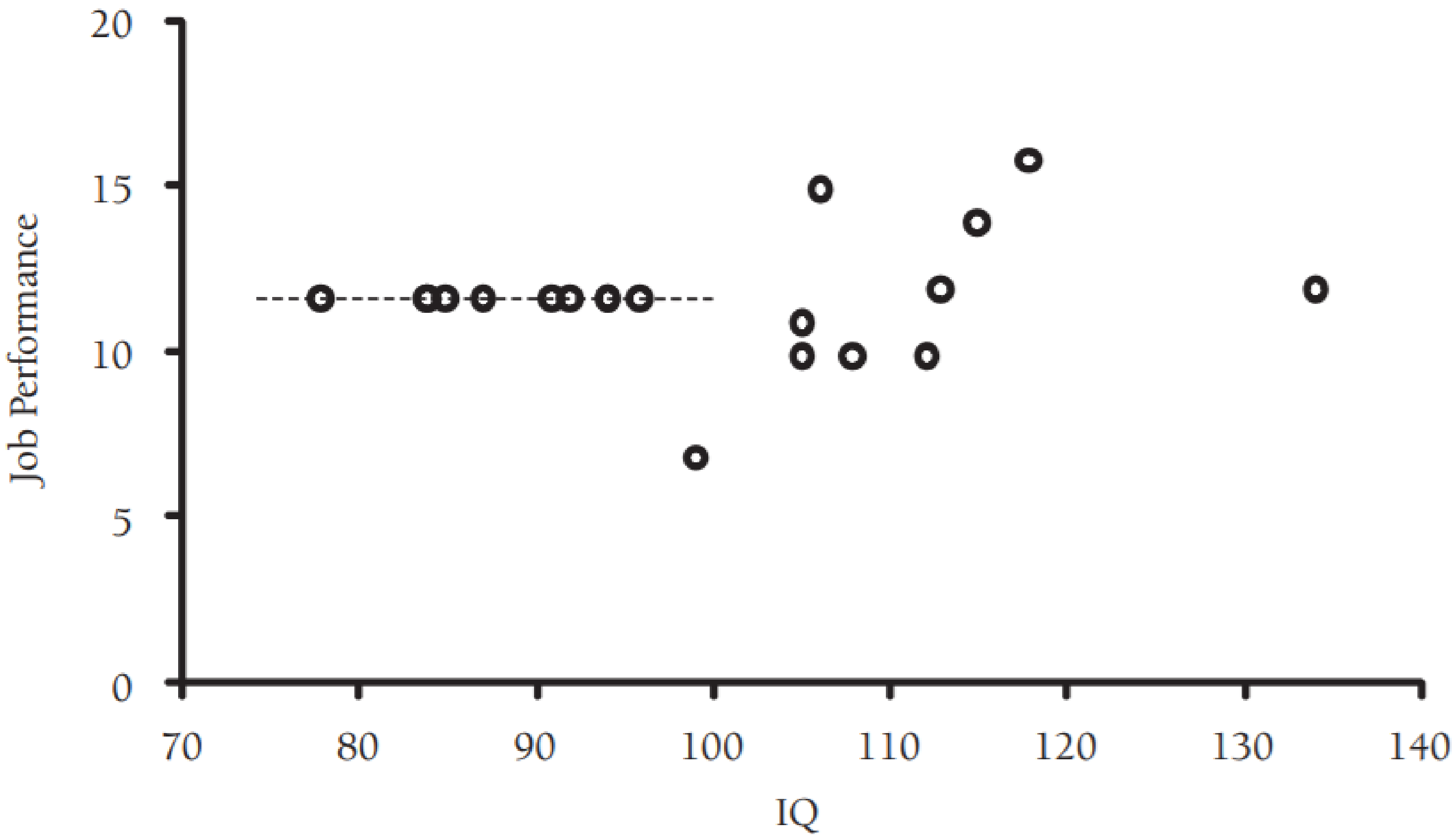
- Interpretation:
  - Coefficient of D is the difference in the expected value of Y between the group with data and the group without data on X.
  - Coefficient of X* is the effect of the group with data on Y
- Disadvantages:
  - This method yields bias estimates of the coefficient even in the case of MCAR

Handle Missing Data

Mean, Mode, Median Imputation

- Fill the missing values with the mean, mode or median of the available cases.
- Disadvantages:
    - Mean imputation does not preserve the relationships among variables
    - Mean imputation leads to An Underestimate of Standard Errors → you're making Type I errors without realizing it.
    - Biased estimates of variances and covariances

Handle Missing Data



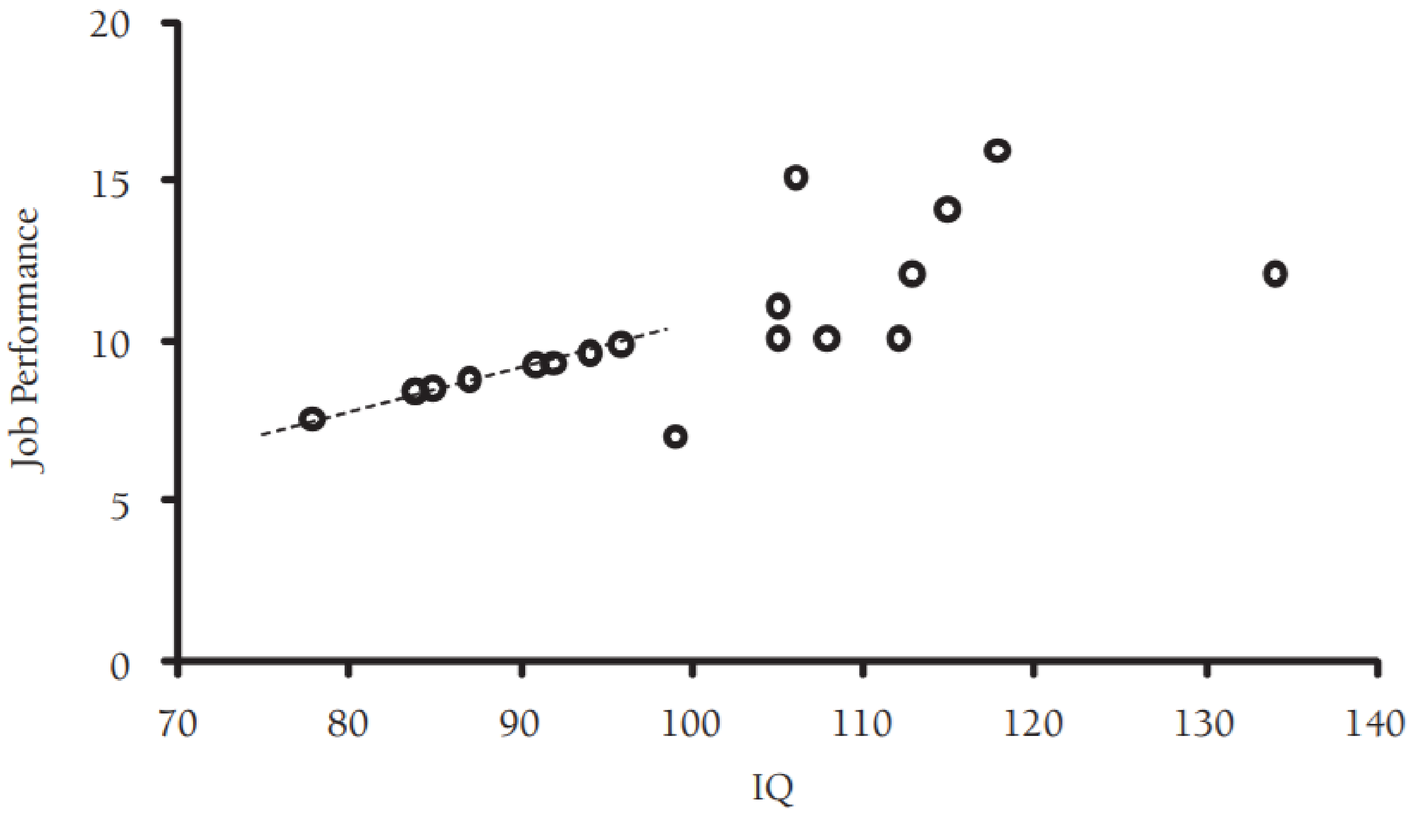
Mean imputation scatterplot of the IQ and job performance data

Handle Missing Data

Regression imputation

- Replaces missing values with predicted scores from a regression equation.
- Use information from the complete variables to fill in the incomplete variables.
- First step: estimate a set of regression equations that predict the incomplete variables from the complete variables.
- Second step: generate predicted values for the incomplete variables. These predicted scores fill in the missing values and produce a complete dataset.

Handle Missing Data



Regression imputation scatterplot of the IQ and job performance data

Handle Missing Data

Stochastic regression imputation

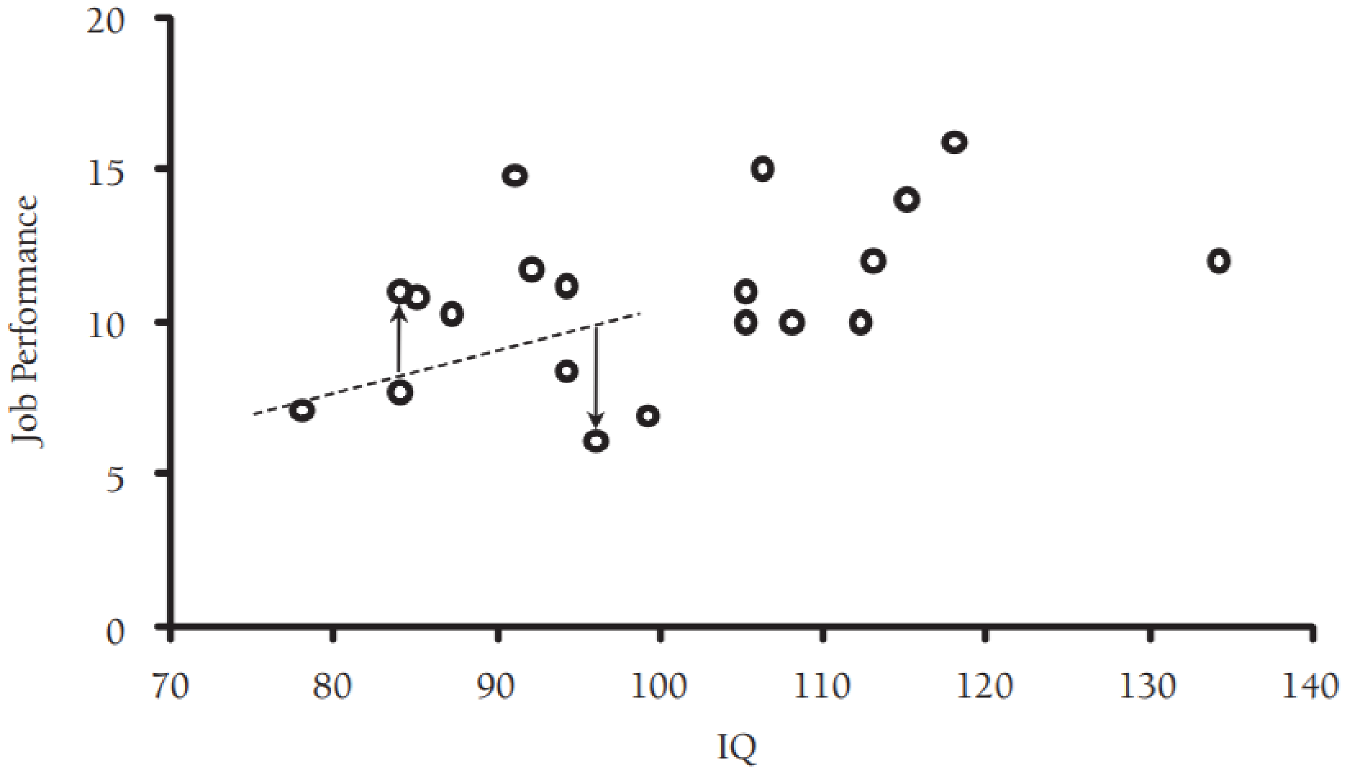
- Use regression equations to predict the incomplete variables from the complete variables
- But it takes the extra step of augmenting each predicted score with a normally distributed residual term.
- First step: estimate a set of regression equations that predict the incomplete variables from the complete variables.
- Second step: generate predicted values for the incomplete variables.
- Final step: restores lost variability to the data by adding a normally distributed residual term to each predicted score.

Handle Missing Data

Stochastic regression imputation

- The residual term is a random value from a normal distribution with a mean of zero and a variance equal to the residual variance from the regression of dependent variable on independent variables.

Handle Missing Data



Stochastic regression imputation scatterplot of the IQ and job performance data

การบ้าน

- ให้นักศึกษาจับคู่กับเพื่อน ทำการค้นคว้าวิธีการจัดการข้อมูลสูญหายมา 1 วิธีการ โดยเขียนสรุปรายงานการค้นคว้า 1 หน้ากระดาษ A4
- วิธีการจัดการข้อมูลสูญหายที่ทำการศึกษามาจะต้องไม่ซ้ำกับวิธีการที่อาจารย์สอน
- ภายในห้องเรียนจะต้องมีวิธีการจัดการข้อมูลสูญหายซ้ำกันไม่เกิน 5 กลุ่ม (ให้นักศึกษาวางแผนบริหารจัดการภายในห้องกันเอง)

Outlier Detection

Most common causes of outliers on a data set:

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation or data set unintended mutations)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

Outlier Detection

Extreme Value Analysis

- Z-score or standard score of an observation is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a gaussian distribution.
- Z-score of any data point can be calculated with the following expression:

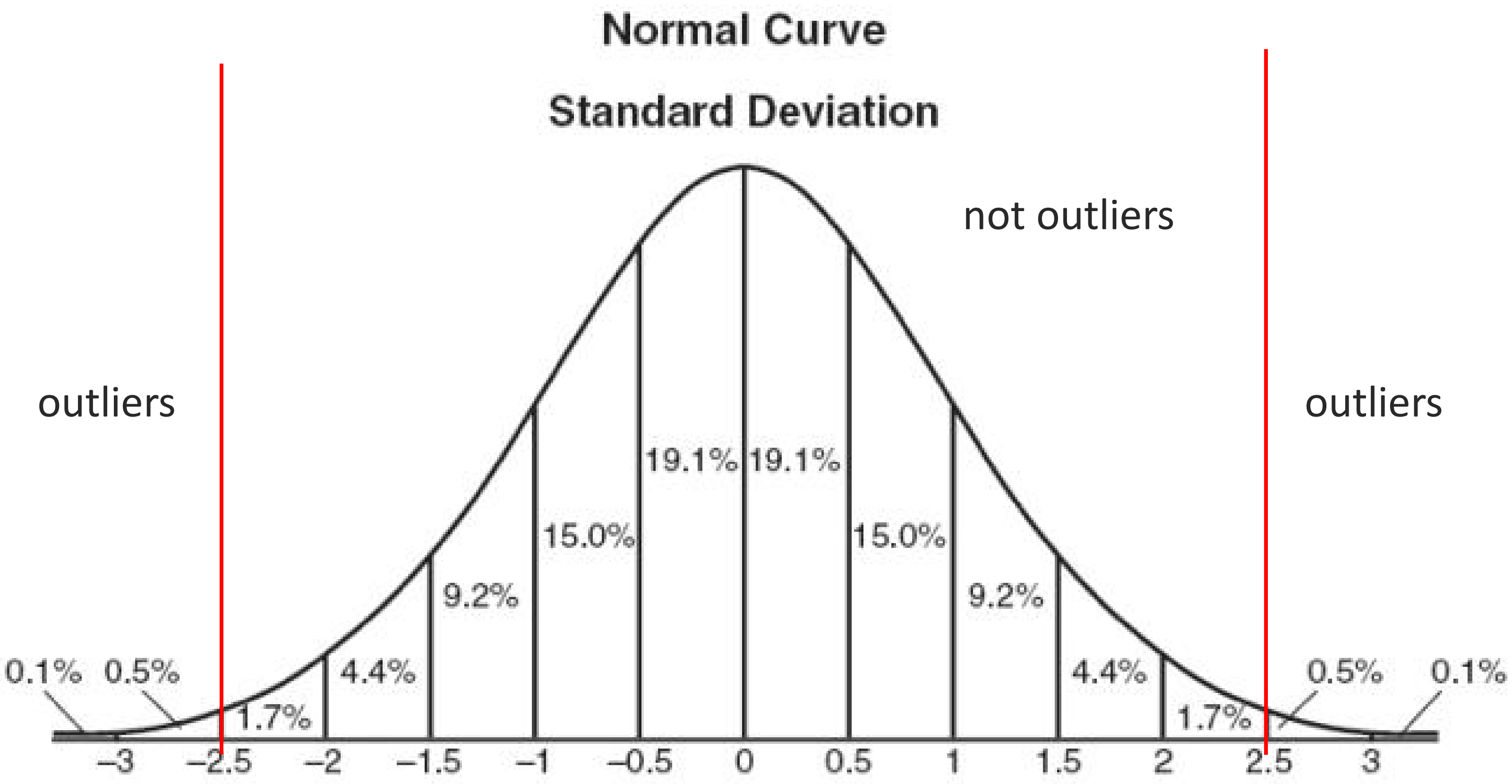
$$z = \frac{x - \mu}{\sigma}$$

- When computing the z-score for each sample, a threshold must be specified. Some good 'thumb-rule' thresholds can be: 2.5, 3, 3.5 or more standard deviations.

Outlier Detection

Extreme Value Analysis

By ‘tagging’ or removing the data points that lay beyond a given threshold we are classifying data into outliers and not outliers

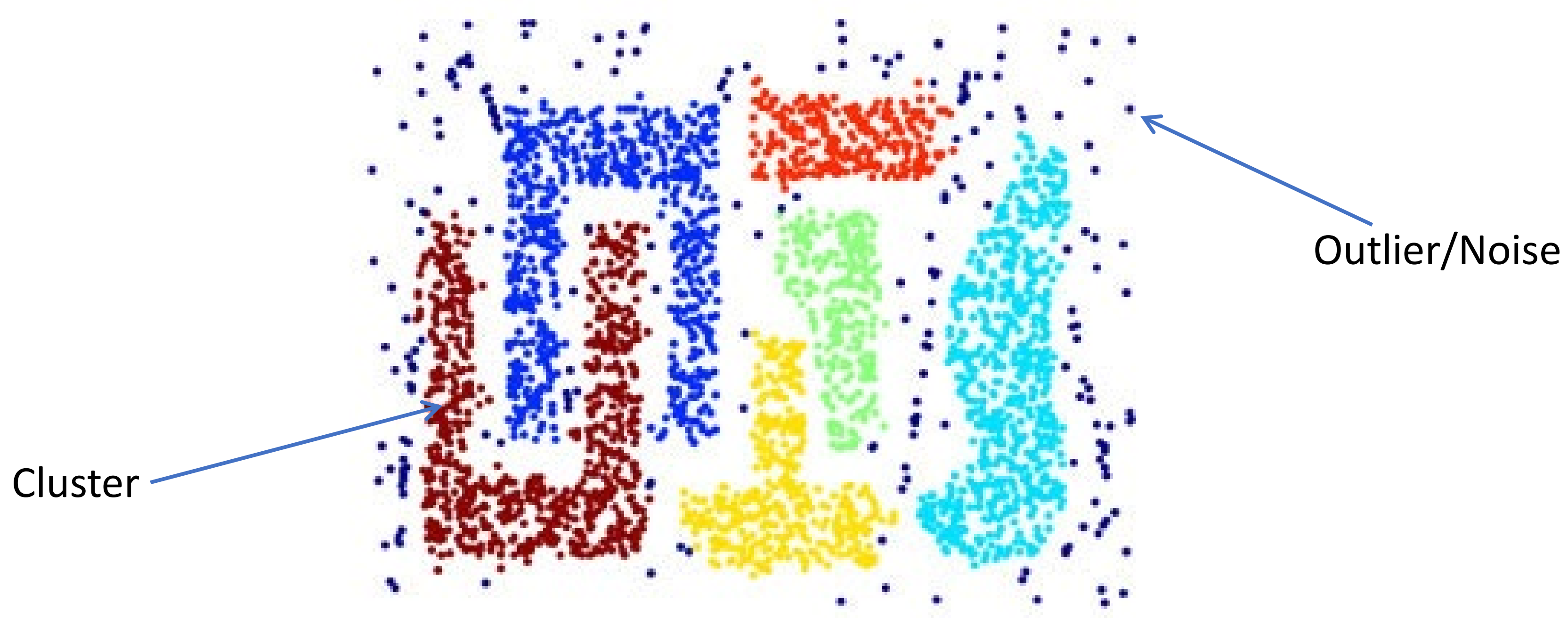


Outlier Detection

DBSCAN

Use the local density of points to determine the clusters.

- Groups together points that are closely packed together (point in high-density regions).
- Marking points that lie alone in low-density regions as outliers.



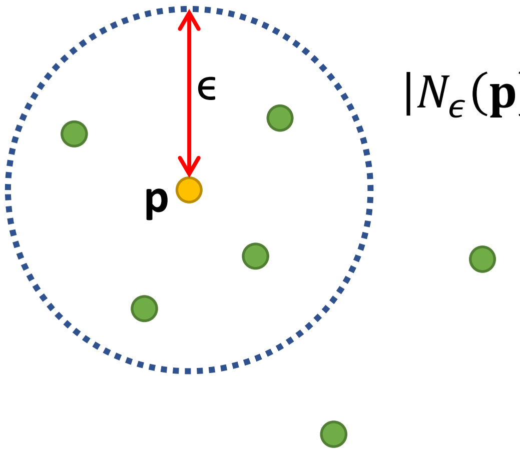
Outlier Detection

How do we measure density of a region?

- **Density at a point** - Number of points within a circle of Radius *Eps (ϵ)* from point **p**.

  ϵ-neighborhood: $N_\epsilon(\mathbf{p}) = \{\mathbf{q} \in \mathbf{D} |\ d(\mathbf{p},\mathbf{q}) \leq \epsilon\}$

- **Dense Region -** For each point in the cluster, the circle with radius ϵ contains at least minimum number of points (*MinPts*).



$|N_\epsilon(\mathbf{p})| = 4$

Outlier Detection

How do we measure density of a region?

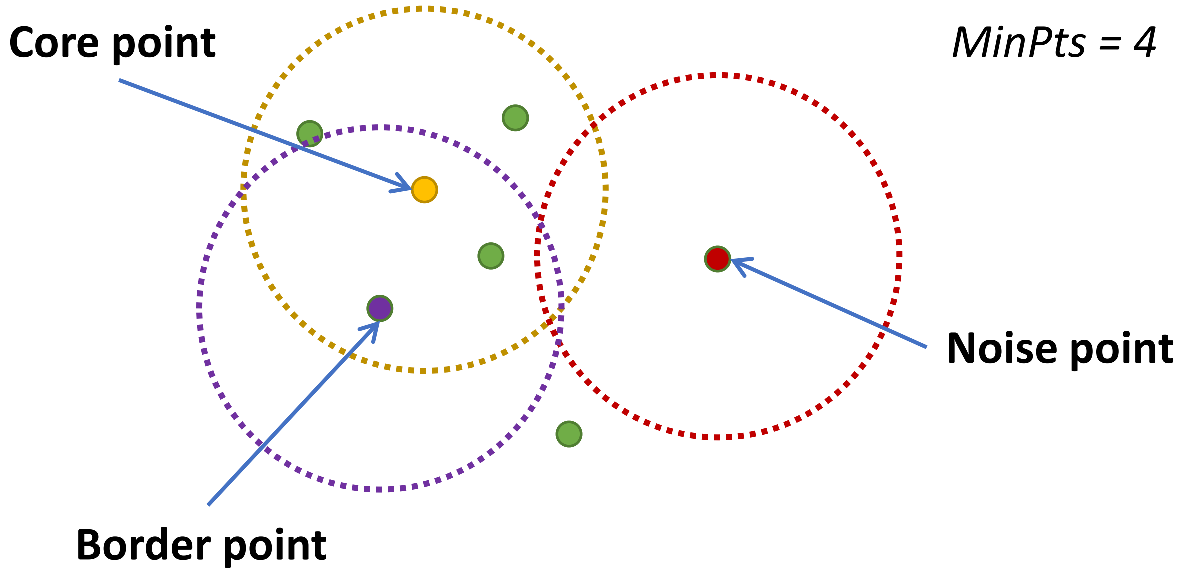
- **Density at a point** - Number of points within a circle of Radius *Eps (ϵ)* from point **p**.

  ϵ-neighborhood*:* $N_\epsilon(\mathbf{p}) = \{\mathbf{q} \in \mathbf{D} |\ d(\mathbf{p}, \mathbf{q}) \leq\ \epsilon\}$

- **Dense Region -** For each point in the cluster, the circle with radius ϵ contains at least minimum number of points (*MinPts*).

A point p can be classified as:

- **Core point** – if $|N_\epsilon(\mathbf{p})| \geq MinPts$
- **Border point** – if $|N_\epsilon(\mathbf{p})| < MinPts$ and **p** belong to ϵ-neighborhood of some core point
- **Noise point** – if **p** is neither a core nor a border point



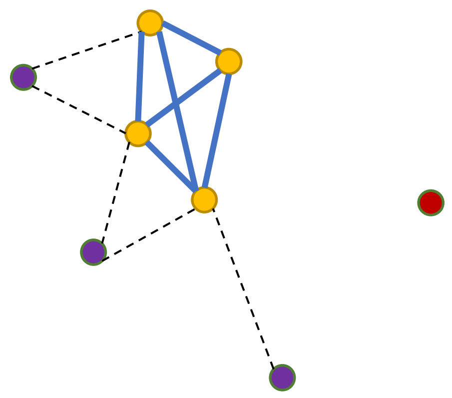
Outlier Detection

How the DBSCAN works

STEP 1: Find ϵ-neighborhood of every point, and identify the core points

STEP 2: Find the connected components of core points on the neighbor graph, ignoring all non-core points.

STEP 3: Assign each non-core point to a nearby cluster if the cluster is an ϵ - neighbor, otherwise assign it to noise.



MinPts = 4

core points

Connected Components -
There exists an edge between two core points

Outlier Detection

Isolation Forest

- Isolation forests are an effective method for detecting outliers or novelties in data.
- It is a relatively novel method based on binary decision trees.

Isolation forest’s basic principle is that outliers are few and far from the rest of the observations.

Outlier Detection

Isolation Forest

- **Build a tree (training)**
  - The algorithm randomly picks a feature from the feature space and a random split value ranging between the maximums and minimums.
  - This is made for all the observations in the training set.
  - To build the forest, a tree ensemble is made averaging all the trees in the forest.

Outlier Detection

Isolation Forest

- **Prediction**
  - Compares an observation against that splitting value in a “node”, that node will have two node children on which another random comparisons will be made.
  - The number of “splittings” made by the algorithm for an instance is named: “path length”.
  - As expected, <u>outliers will have shorter path lengths than the rest of the observations.</u>

Outlier Detection

Isolation Forest

- **Prediction**
  - An outlier score can be computed for each observation:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

Where $h(x)$ is the path length of the sample *x*,

$c(n)$ is the 'unsuccessful length search' of a binary tree (the maximum path length of a binary tree from root to external node)

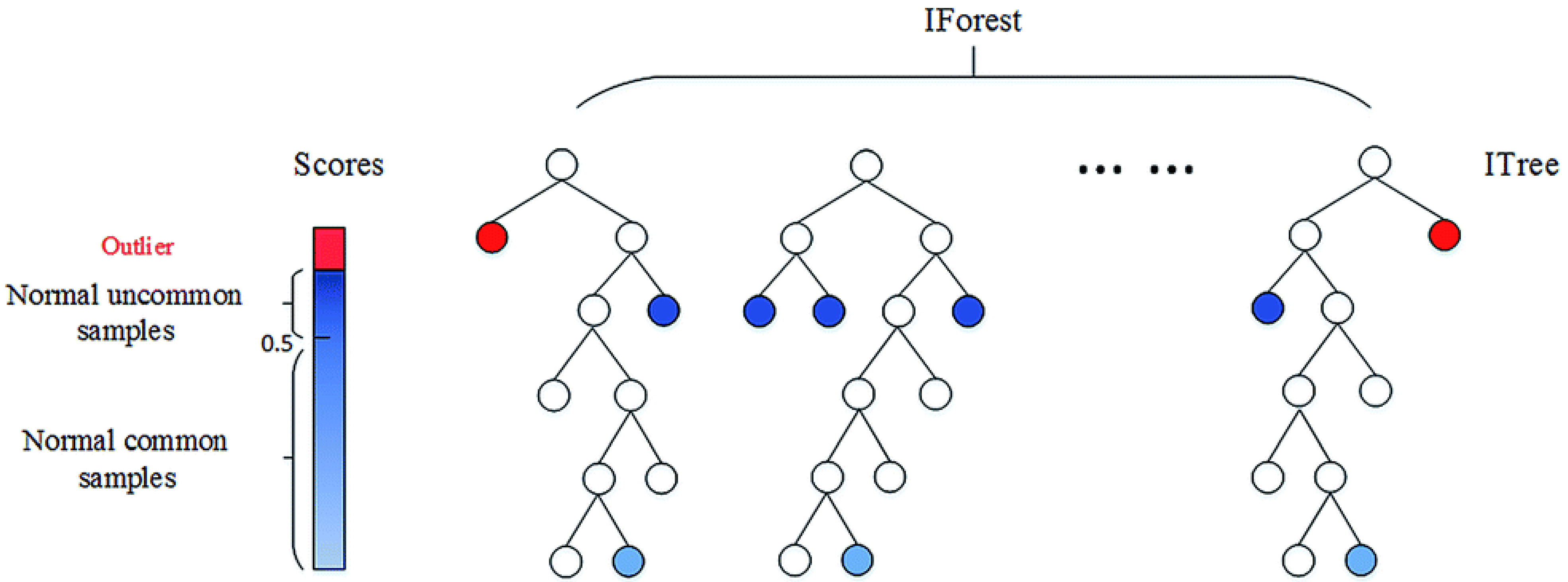
n is the number of external nodes.

After giving each observation a score ranging from 0 to 1; 1 meaning more outlyingness and 0 meaning more normality.

A threshold can be specified (ie. 0.55 or 0.60)

Outlier Detection

Isolation Forest



References

- Zaki, Mohammed J. and Wagner Meira Jr (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. USA: Cambridge University Press.
- Nguyen, M. (2021). A Guide on Data Analysis. url: https://bookdown.org/mike/data_analysis/
- Kowalewski, M. (2020). Data Cleaning In 5 Easy Steps + Examples. url: https://www.iteratorshq.com/blog/data-cleaning-in-5-easy-steps/
- Santoyo, S. (2017). A Brief Overview of Outlier Detection Techniques. url: https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561