Data Pre-Processing-I

(Introduction, Need, Data Cleaning)

Dr. JASMEET SINGH
ASSISTANT PROFESSOR, CSED
TIET, PATIALA

Data Pre-Processing

- **Data Pre-processing:** It is that phase of any machine learning process, which transforms, or encodes, the data to bring it to such a state where it can be easily interpreted by the learning algorithm.

“Data pre-processing is not a single standalone entity but a collection of multiple interrelated tasks”

“Collectively data pre-processing constitutes majority of the effort in machine learning process (approx. 90 %)”

Need of Data Pre-Processing

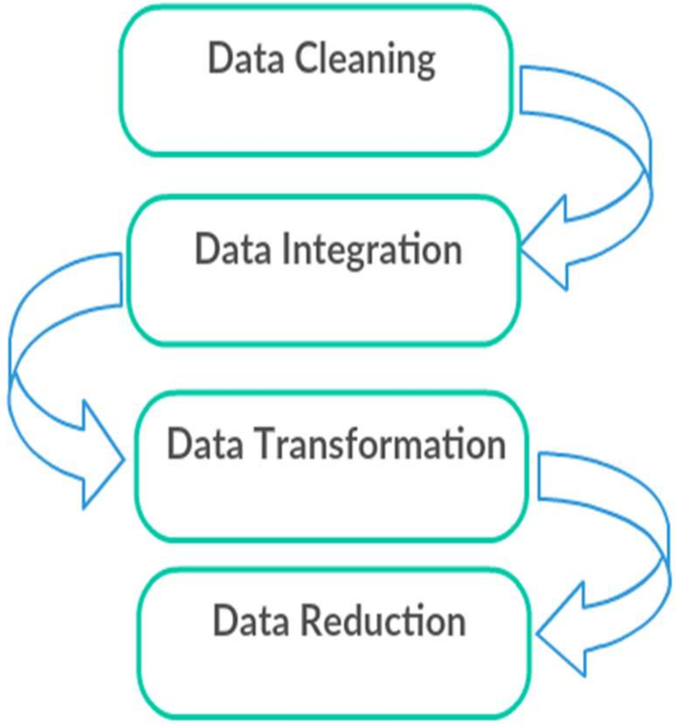
- Data in the real world is “quite messy”
  - **incomplete**: missing feature values, absence of certain crucial feature, or containing only aggregate data.
    - e.g. Height=“ ”
  - **noisy**: containing errors or outliers
    - e.g. Weight=“5000” or “-60”
  - **inconsistent**: containing discrepancies in feature values.
    - e.g. Age=“20” and dob=“12 july 1990”
    - e.g. contradictions between duplicate records

Need for data Pre-processing

- **Unstructured Data (Text)**
  - Lower Case
  - Normalization (remove punctuation, special symbols, urls)
  - Stopwords Removal (of, and, the,….)
  - Stemming/Lemmatization (plays, playing, played → play)
- **Unstructured Data (Images)**
  - Read image
  - Resize image
  - Remove noise(Denoise)
  - Segmentation
  - Morphology(smoothing edges)

Pre- Processing in Structured Data

- Major data pre-processing tasks
  - Data cleaning
  - Data integration
  - Data transformation
  - Data reduction



Data Cleaning

- **Data cleaning:** It is a procedure to "clean" the data by filling in missing values, smoothening noisy data, identifying or removing outliers, and resolving data inconsistencies.

- Data cleaning tasks
  - Fill missing values
  - Noise smoothening and outlier detection
  - Resolving inconsistencies

Data Cleaning- Missing Values

Missing values: data values are not available.

i.e. many data entities have no data values corresponding to a certain feature like BMI value missing for some persons in a diabetes dataset.

- Probable reasons for missing values:
  - faulty measuring equipment
  - reluctance of person to share certain detail
  - negligence on part of data entry operator
  - feature unimportance at time of data collection

Data Cleaning- Missing Values Contd…

- Missing data handling techniques
  - Removing the data entity
  - Manually filling the values
  - Imputation (process used to determine and assign replacement values for missing, invalid, or inconsistent data)

“Technique selection is specific to user’s preference, dataset or feature type or problem set”

Data Cleaning- Missing Values Contd…

- Sample dataset related to forest fires

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd…

Removing the data entity: Most easiest way directly to clean the data, but this is usually discouraged as it leads to loss of data, as you are removing the data entity or feature values that can add value to data set as well.

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
mar	89.3	102.2	11.4	99	1.8
aug	91.5	608.2	8	86	2.2
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd…

- Manually filing up of values : This approach is time consuming, and not recommended for huge data sets.

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	17	33	0.8
mar	91.6	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	380	22.2	92	1.8
aug	90	495.6	24.1	27	2
aug	91.5	608.2	8	86	2.2
sep	91	692.6	22	63	5.4
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd…

- Imputation : process used to determine and assign replacement values for missing, invalid, or inconsistent data. Various imputation methods include:

- Central Tendency Imputation
- Hot Deck Imputation
- Cold Deck Imputation
- Model Based Imputation
  - Nearest Neighbor Imputation
  - Tree-Based Imputation

Data Cleaning (Missing values)- Contd…

Central tendency Imputation : Replacing the missing value by central tendency (mean, median, mode) for a feature vector or belonging to same class of feature vector.

Mean

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Median

$$md = x_{\frac{(n-1)}{2}} \quad \text{for n is odd}$$

$$md = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right) \quad \text{for n is even}$$

Mode : Mode is the most frequent value corresponding to a certain feature in a given data set

Data Cleaning- Missing Values Contd…

- Replacing Mean Value:

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	NaN	33	NaN
mar	NaN	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	NaN	22.2	NaN	NaN
aug	NaN	495.6	24.1	27	NaN
aug	91.5	608.2	8	86	2.2
sep	91	692.6	NaN	63	5.4
sep	92.5	698.6	22.8	40	4

Month	FFMC	DC	temp	RH	wind
mar	86.2	94.3	8.2	51	6.7
oct	90.6	669.1	18	33	0.9
oct	90.6	686.9	15.3	33	3.57
mar	90.5	77.5	8.3	97	4
mar	89.3	102.2	11.4	99	1.8
aug	92.3	458.3	22.2	58.7	3.57
aug	90.5	495.6	24.1	27	3.57
aug	91.5	608.2	8	86	2.2
sep	91	692.6	15.3	63	5.4
sep	92.5	698.6	22.8	40	4

Data Cleaning- Missing Values Contd...

- “ Replacing by mean value: Not a suitable method if data set has many outliers”
- For example: weighs of humans   67, 78, 900,-56,389,-1 etc.

Outlier

Mean is 229.5

- Can be replaced with median in such cases.
- “Mode is a good option for missing values in case of categorical variables”

Data Cleaning- Missing Values Contd…

Hot Deck Imputation

- Computes how many number of features (other than feature with missing data) have same values in the entire training examples and choose it for replacement.
- Used mostly in categorical data.

Cold Deck Imputation

- Similar to hot deck imputation.
- In it missing observations are replaced by values from a source unrelated to the data set under consideration.

Data Cleaning- Missing Values Contd…

Nearest Neighbor-Based Imputation

- Rely on distance metrics
- evaluate the distance between recipients and donors.
- Used after converting all features to numerical (quantitative)

K-Nearest Neighbors (KNN) Imputation

- Nearest neighbor imputation algorithms efficiently fill in missing data by replacing each missing value with a value derived from similar cases within the entire dataset.

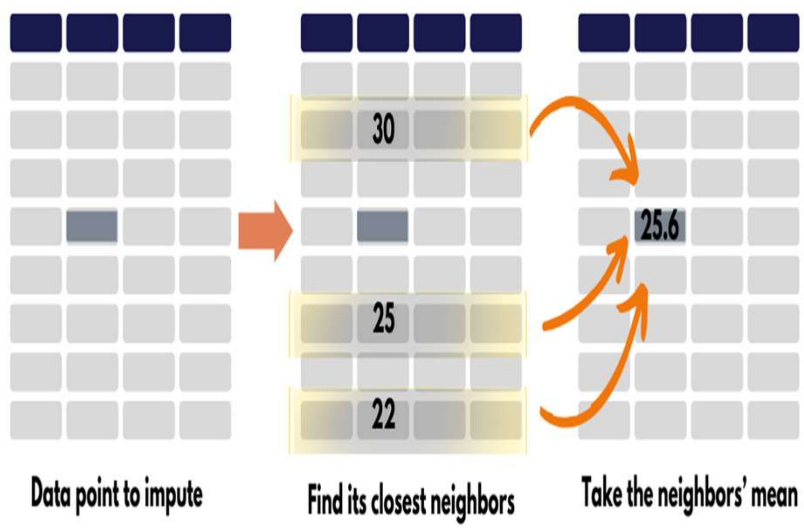
- Imputation with K-Nearest Neighbors (KNN) estimates missing values in a dataset by considering the values of the closest data points, determined by a distance metric like Euclidean distance.

- The missing value is then assigned the average of these nearest neighbors’ values, weighted by their proximity.

K-Nearest Neighbors (KNN) Imputation

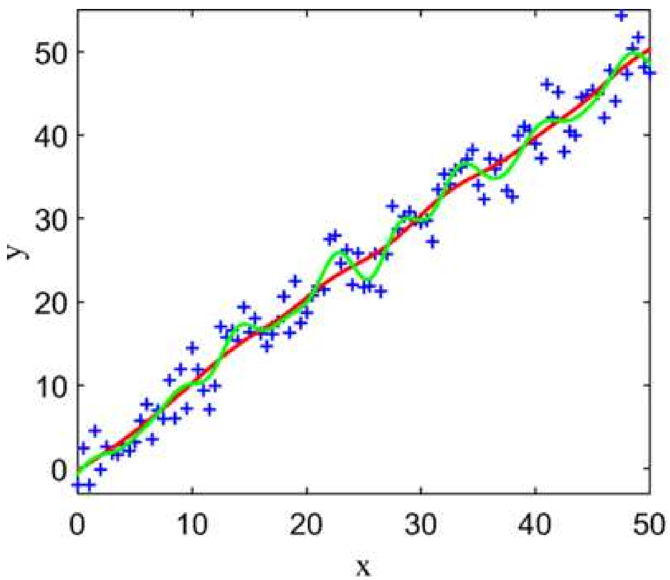
- Consider the following dataset with 5 variables and 11 observations. We aim to impute the missing value in the 5th row of variable 2. First, we identify the row's 3 closest neighbors (highlighted by the squared boxes) using a KNN algorithm. Then, we calculate the average of the values for variable 2 from these neighbors.

- The imputed value is calculated as (Value1 x w1 + Value2 x w2 + Value3 x w3) / 3, where w1, w2, and w3 are weights proportional to the distance of each neighbor from the data point being imputed.

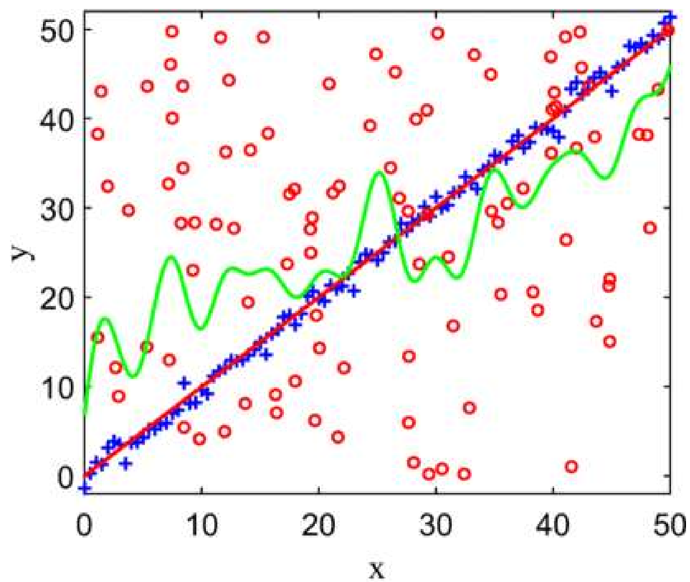


Data Cleaning- Noisy Data and Outliers

- Noise and outliers are both types of data anomalies, but they differ in their characteristics and impact on data analysis.



(a) noise



(b) outlier

Noise vs. Outliers

Noise	Outliers
1. Definition: Noise refers to random, unexplained variations in data that obscure the underlying pattern or signal. It can be thought of as irrelevant or extraneous data that does not convey useful information.	1. **Definition**: Outliers are data points that significantly deviate from the rest of the data. They appear as unusually high or low values that do not fit the general distribution of the data.
2. **Cause:** Noise can be introduced due to various factors such as measurement errors, environmental variations, or sensor inaccuracies.	2. **Cause:** Outliers can arise due to rare events, data entry errors, experimental errors, or natural variations.
3. **Frequency:** Noise is typically widespread and affects many data points.	**3. Frequency:** Outliers are typically rare and only affect a small number of data points.
4. **Pattern**: It does not follow any systematic pattern and appears as random fluctuations in the data.	4. **Pattern:** They stand out as being distinct from the rest of the data, often visible as isolated points on a graph.

Noise vs. Outliers

Noise	Outliers
5. **Impact**: Noise can make it difficult to detect the true signal or trends in the data, reducing the accuracy of models.	5. **Impact:** Outliers can have a significant impact on statistical analyses, such as skewing the mean or influencing the results of regression models.
6. **Handling:** Noise is often filtered out or smoothed using statistical methods like moving averages, smoothing techniques, or noise-reduction algorithms.	6. **Handling:** Outliers are often identified and treated separately, either by removing them, analyzing them as special cases, or applying robust statistical methods that are less sensitive to outliers.

Data Cleaning- Noisy Data

- Major reasons of random variations in data are:
  - Malfunctioning of collection instruments.
  - Data entry lags.
  - Data transmission problems

To deal with these anomalous values, data smoothing techniques are applied, some of the popular ones are

- Binning method
- Exponentially Weighted Moving Averages (EWMA)

Binning Method for Noisy Data

Binning method : performs the task of data smoothening.

Steps to be followed under binning method are:

Step 1: Sort the data into ascending order.

Step 2: Calculate the bin size (i.e. number of bins)

Step 3: Partition or distribute the data equally among the bins starting with first element of sorted data.

Step 4: perform data smoothening using bin means, bin boundaries, and bin median.

Last bin can have one less or more element!!

Binning Method for Noisy Data

Example : 9, 21, 29, 28, 4, 21, 8, 24, 26

Step1: sorted the data 4, 8, 9, 21, 21, 24, 26, 28, 29

Step 2 : Bin size calculation

$$\text{Bin size} = \frac{Max\ value - Min\ value}{data\ size}$$

$$= \frac{29-4}{9} = 2.777$$

But we need to take ceiling value, so bin size is 3 here

Binning Method for Noisy Data

- Step 3 : Bin partitioning (equi-size bins)

Bin 1: 4, 8, 9
Bin 2: 21, 21, 24
Bin 3: 26, 28, 29

Step 4 : data smoothening

- Using mean value : replace the bin values by bin average

Bin 1: 7, 7, 7
Bin 2: 22, 22, 22
Bin 3: 27, 27, 27

Binning Method for Noisy Data

- ➢ Using boundary values : replace the bin value by a closest boundary value of the corresponding bin.

  Bin 1: 4, 9, 9
  Bin 2: 21, 21, 24
  Bin 3: 26, 29, 29

  "Boundary values remain unchanged in boundary method"

  ➢ Using median values : replace the bin value by a bin median.

  Bin 1: 8, 8, 8
  Bin 2: 21, 21, 21
  Bin 3: 28, 28, 28

Exponentially Weighted Moving Averages

- Exponentially Weighted Moving Averages (EWMA) is a technique used to smooth out noisy data by giving more weight to recent observations.
- It's a popular method for time series data, especially when dealing with noisy or volatile data.
- Steps to perform EMWA:
  - Assign a weighting factor (α) between 0 and 1, which determines how quickly the average responds to new data.
  - Initialize the EWMA with the first data point.
  - For each subsequent data point, calculate the new EWMA as:

$$V_t = (1 - \beta)V_{t-1} + (\beta)\theta_t$$

Where V_t and θ_t are weighted and actual values on t^{th} day

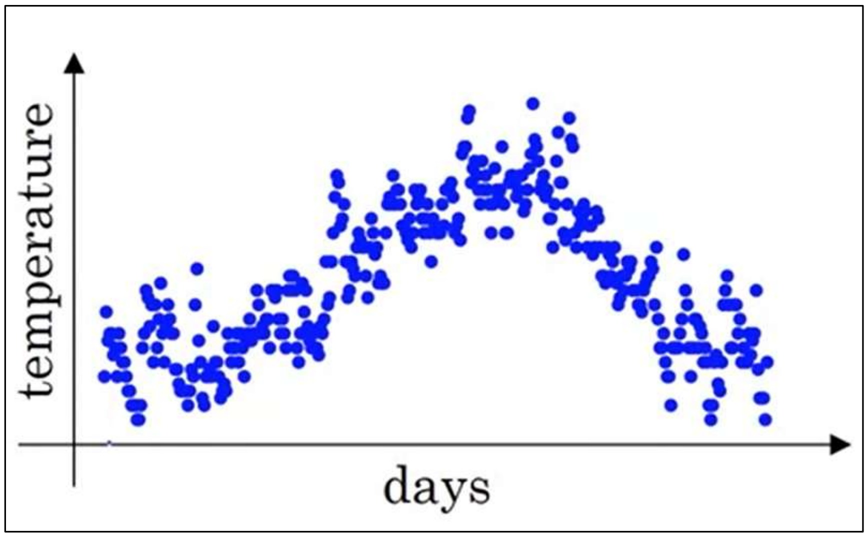
Exponentially Weighted Moving Averages

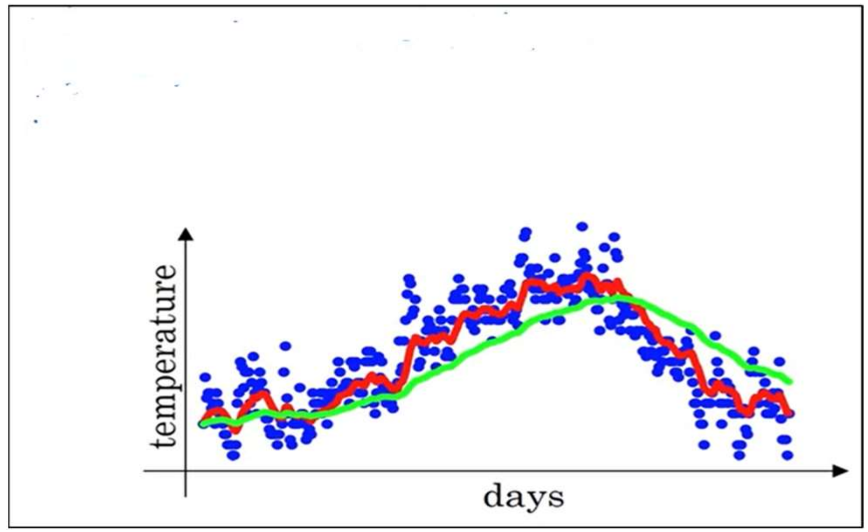
PLOT OF TEMPERATURE W.R.T EACH DAY OF AN YEAR

EFFECT OF WEIGHTED MOVING AVERAGES

$$V_t = (1-\beta)V_{t-1} + (\beta)\theta_t$$

WHERE V_t AND θ_t ARE WEIGHTED AND ACTUAL TEMPERATURES ON TTH DAY





Exponentially Weighted Moving Averages

Suppose we have daily stock prices with some noise:

Day 1: $100 Day 2: $105 Day 3: $102 Day 4: $110 Day 5: $108

Using EWMA smooth values with α = 0.3

Solution: EWMA_1 = $100

EWMA_2 = (0.3 * $105) + (0.7 * $100) = $101.5

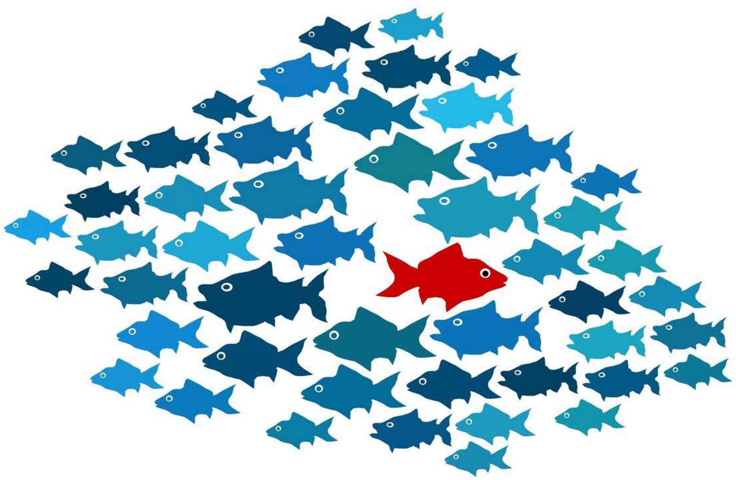
EWMA_3 = (0.3 * $102) + (0.7 * $101.5) = $101.85

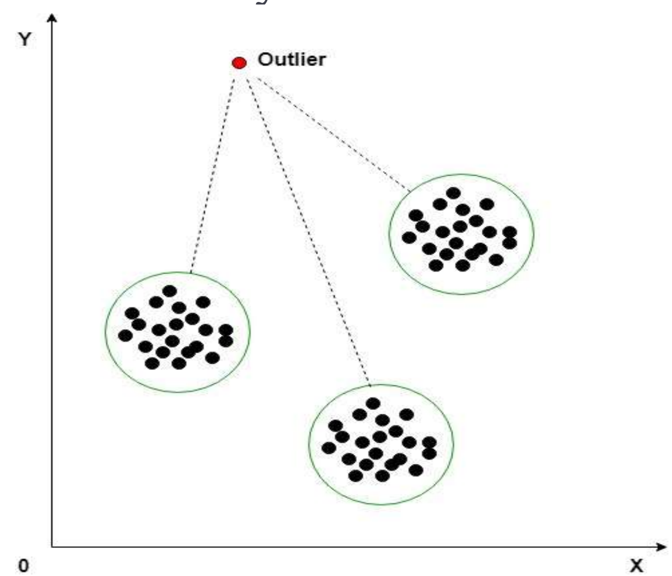
EWMA_4 = (0.3 * $110) + (0.7 * $101.85) = $104.095

EWMA_5 = (0.3 * $108) + (0.7 * $104.095) = $105.3665

Outlier Analysis

- An **outlier** is an object that deviates significantly from the rest of the objects.
- They can be caused by measurement or execution error.
- The analysis of outlier data is referred to as outlier analysis or outlier mining.





Outlier Analysis

- Types of Outliers:
  - Univariate: A univariate outlier is a data point that consists of an extreme value on one variable.
  - Multivariate: A multivariate outlier is a combination of unusual scores on at least two variables.
- Outlier Detection and Handling Methods
  - Extreme Value Analysis
  - Linear Models
  - Proximity-based Methods
  - Information Theoretic Methods
  - Isolation Forests Based Methods

Extreme Value Analysis- Outlier Analysis

- **Numeric Outlier**
- This is the simplest, nonparametric outlier detection method in a one dimensional feature space.
- Here outliers are calculated by means of the *IQR* (InterQuartile Range).
- The first and the third quartile (*Q1, Q3*) are calculated.
- An outlier is then a data point x_i that lies outside the interquartile range. That is:

$$x_i > Q_3 + k(IQR)\ and x_i < Q_1 - k(IQR)$$

where IQR=Q3-Q1 and k >=0

Assume the data 6, 2, 1, 5, 4, 3, 50. If these values represent the number of chapatis eaten in lunch, then 50 is clearly an outlier.

Sorted Values: 1, 2, 3, 4, 5, 6, 50

Q1 25 percentile of the given data is, 2

Q2 50 percentile of the given data is, 4.0

Q3 75 percentile of the given data is, 6

IQR= 6-2=4, k=1.5

Range: 2-1.5 X 4=-4 and 6+1.5X4=12

50 is Outlier

Extreme Value Analysis- Outlier Analysis

- **Z-score** is a parametric outlier detection method in a one or low dimensional feature space.
- This technique assumes a Gaussian distribution of the data.
- The outliers are the data points that are in the tails of the distribution and therefore far from the mean.
- How far depends on a set threshold z_{thr} for the normalized data points z_i calculated with the formula:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i is a data point, μ is the mean of all x_i and is the standard deviation of all x_i.

An outlier is then a normalized data point which has an absolute value greater than z_{thr}.

$$|z_i| > z_{thr}$$

Commonly used z_{thr} values are 2.5, 3.0 and 3.5.

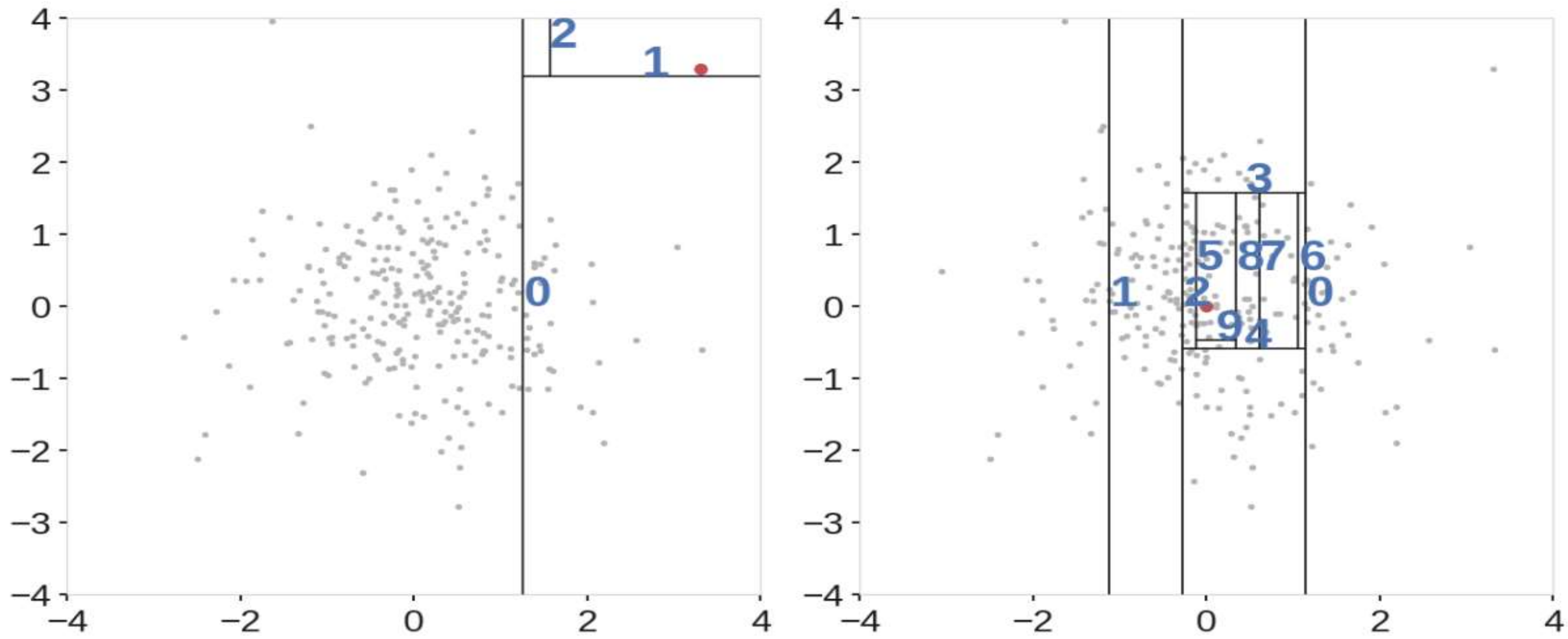
Outlier Analysis

- **Linear Models**:
  - ➢Projection methods that model the data into lower dimensions using linear correlations.
  - ➢For example, principle component analysis and data with large residual errors may be outliers.
- **Proximity-based Models**:
  - ➢Data instances that are isolated from the mass of the data as determined by cluster, density or nearest neighbor analysis.
- **Information Theoretic Models**:
  - ➢Outliers are detected as data instances that increase the complexity (minimum code length) of the dataset.

Outlier Analysis-Isolation Forest Based Methods

- Isolation Forests(IF), similar to Random Forests, are build based on decision trees. And since there are no pre-defined labels here, it is an unsupervised model.
- In an Isolation Forest, randomly sub-sampled data is processed in a tree structure based on randomly selected features.
- The samples that travel deeper into the tree are less likely to be anomalies as they required more cuts to isolate them.
- Similarly, the samples which end up in shorter branches indicate anomalies as it was easier for the tree to separate them from other observations.

Outlier Analysis-Isolation Forest Based Methods (Contd...)



(a) Anomaly point (b) Nominal point

Outlier Analysis-Isolation Forest Based Methods (Contd...)

- The algorithm starts with the training of the data, by generating Isolation Trees:

1. When given a dataset, a random sub-sample of the data is selected and assigned to a binary tree.
2. Branching of the tree starts by selecting a random feature (from the set of all N features) first. And then branching is done on a random threshold (any value in the range of minimum and maximum values of the selected feature).
3. If the value of a data point is less than the selected threshold, it goes to the left branch else to the right. And thus a node is split into left and right branches.
4. This process from step 2 is continued recursively till each data point is completely isolated or till max depth(if defined) is reached.
5. The above steps are repeated to construct random binary trees.

Outlier Analysis-Isolation Forest Based Methods (Contd…)

- After an ensemble of iTrees(Isolation Forest) is created, model training is complete.
- During scoring, a data point is traversed through all the trees which were trained earlier.
- Now, an ‘anomaly score’ is assigned to each of the data points based on the depth of the tree required to arrive at that point. This score is an aggregation of the depth obtained from each of the iTrees.

Outlier Analysis-Isolation Forest Based Methods (Contd...)

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

Where x: data point under consideration

n: total data points

E(h(x)): Mean (expected value) of depth of data point x in all the isolation trees

c(n): mean depth of all data points in all the trees.

- If x is outlier, E(h(x)) << c(n) then anamoly score is closer to 1
- If x is not an outlier E(h(x))> c(n) then anamoly score lies in between 0 and 0.5

Data Cleaning – Inconsistent Data

Inconsistent Data: discrepancies between different data items.

e.g. the "Address" field contains the "Phone number"

To resolve inconsistencies

- Manual correction using external references
- Semi-automatic tools
  - To detect violation of known functional dependencies and data constraints
  - To correct redundant data

To avoid inconsistencies, perform data assessment like

knowing what the data type of the features should be and whether it is the same for all the data objects."

Data Cleaning-Summary