Data Exploration/Preparation

Agenda

Key Takeaways-

- Mean, median and mode
- Imputing the missing values with the right statistic
- Measure of dispersion - Variance and Standard deviation
- Percentiles and Quartiles
- Assessing the linear relationship
- Data Encoding and methods
- Data Normalization

Measures of Central tendency

- A measure of central tendency is a single value that attempts to describe/summarize a set of data by identifying the central position.
- This value need not to be always present in the data.
- The most commonly used measure of central tendency are
  - Mean
  - Median
  - Mode
- They are also known as summary statistics.

Mean

The mean (or average) is equal to the sum of all the values in the dataset divided by the total number of values in the dataset. So, if there are n values in the dataset such as x1, x2, x3…….xn, then the mean is

$$mean\ (\bar{x}) = \frac{x_1 + x_2 + x_3 + \ldots\ldots\ldots\ldots\ldots x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Problems with mean :

- Not robust to outliers (influenced by the presence of extremely large/smaller values)

Median

- Median is the value that lies at the center of the data when the data is sorted.
- Median is less affected by the outliers.
- The method for finding the median depends on whether our dataset has an even or odd number of values.
- Finding the median -
  - Sort the data in ascending order
  - If there are odd number of values in the dataset, then the median is the center value that divides the dataset in two equal halves.
  - If there are even number of values in the dataset, then median is the average of two middle values.

- E.g.
  Case 1 : [23,17,18,21, 16,14, 2] → Sort it → [2,14,16,17,18,21,23], median = 17
  Case 2 : [4,6,3,5,9,70] → Sort it → [3,4,5,6,9,70], median = 4

Mode

- Mode is the value that occurs most frequently in the dataset.
- Mode is generally used if the data is non-numeric(discrete).
- A dataset can have multiple modes. Dataset with one value of mode is unimodal, with two modes is bimodal and so on.

- **E.g.**

  [“Hero”, “Honda”, “TVS”, “Hero”, “Suzuki”, “Hero”, “Honda”, “Hero”] → mode = Hero

Measure of dispersion

- A measure of dispersion is a statistic that tells us how dispersed, or spread out, data values are.
- Measure of dispersion helps us in identifying the overall spread of the data.
- Most commonly used measure of dispersion are
  - Range
  - Variance
  - Standard Deviation
  - Interquartile Range

Measure of dispersion

- A measure of dispersion is a statistic that tells us how dispersed, or spread out, data values are.
- Measure of dispersion helps us in identifying the overall spread of the data.
- Most commonly used measure of dispersion are
  - Range
  - Variance
  - Standard Deviation
  - Interquartile Range

Variance

- Variance is defined as the average of squared difference from the mean.
- Variance is a better indicator of dispersion.

$$variance\ (\sigma^2) = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots\ldots\ldots\ldots\ldots\ldots+(x_n-\mu)^2}{n} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

- NOTE - Variance is expressed in the squared units.

Standard Deviation

- Standard deviation represents the actual variation of the data from the mean and is represented in the same unit as that of the data.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

Percentiles and Quartiles

- Percentile represents the value below which a given percentage of observations fall. For e.g. 80th percentile of the data tells the value below which 80% of the data lies (or the value above which 20% of the data lies).
- Quartiles are the 3 points that divide the data into 4 equal parts.
    - First Quartile(Q1) : 25th percentile of the dataset i.e the value below which 25% of the values lies.
    - Second Quartile(Q2) : 50th percentile of the dataset i.e the value below which 50% of the values lies.
    - Third Quartile(Q3) : 75th percentile of the dataset i.e the value below which 75% of the values lies.
- **InterQuartile Range**
    - IQR = (Q3 - Q1)
    - IQR is used to indicate the spread of the data.

Assessing the linear relationship

- Covariance and Correlation are the two important statistical terms to check the linear relationship/association between two numeric variables.

Covariance

- Covariance suggests the direction of linear relationship between two variables.

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_x) * (y_i - \mu_y)}{N}$$

- The covariance values lie anywhere between -∞ to +∞
- A positive value of covariance indicates directly proportional relationship b/w the variables, negative value indicates inversely proportional, whereas zero value of covariance indicates no relationship b/w the variables.
- The covariance values changes with scaling of variables and their unit of measurement.

.

Correlation

- Correlation measures both the magnitude(strength) and direction of the linear relationship between two variables.

$$Correlation = \frac{Cov(x,y)}{\sigma_x * \sigma_y}$$

- Correlation value ranges from -1 to 1.
- Therefore, its magnitude has direct significance and can be used to compare how strong or weak the relationship is.
- Correlation of -1 means perfect inversely proportional relationship.
- Correlation of 0 means no relationship.
- Correlation of 1 means perfect directly proportional relationship.

.

Data Encoding

- Converting the categorical data to numeric values so that they are easily ingestible to ML algorithms and helps them in learn better.
- Some ML algorithms can directly support categorical features whereas many others don't.
- So data encoding is a way to transform the categorical values into suitable numeric values.
- Several types of categorical data -
    - Binary : Having only two values. E.g. Yes/No, True/False, Man/Woman
    - Ordinary : The categories have an inherent order. E.g. Service ratings - Very unsatisfied, Unsatisfied,OK, Satisfied, Excellent. Debit card types - Classic, Silver, Gold, Platinum
    - Nominal : The categories don't have any inherent order. E.g. city names

- Most commonly used Data Encoding techniques -
    - Label Encoding
    - One Hot Encoding

Label Encoding

- In Label Encoding, each label is converted to a numeric/integer value.
- It is suitable when the categorical data is ordinal in nature so that it retains the order.

Actual	Encoded
Very unsatisfied	1
Unsatisfied	2
OK	3
Satisfied	4
Excellent	5

Manufacturer	Encoded
Hero	1
Honda	2
TVS	3
Suzuki	4

- Downside - Using Label Encoding, Nominal data can be misinterpreted by the algorithms.

One Hot Encoding

- In One Hot Encoding, for each label, a new dummy variable is defined. Here, each value is mapped to a binary vector having either 0 or 1, where 0 indicates the absence and 1 indicates the presence of a category.
- In the binary vector, only one value is hot (1) and rest are cold (0), therefore the name is one hot encoding.

Manufacturer
Hero
Honda
TVS
Suzuki

Hero	Honda	TVS	Suzuki
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

Data Normalization/Standardization

Why Data Normalization is needed?

- The real time data may contain attributes with varying scales. E.g. in loan data, the age generally varies from 18 to 100, whereas loan amount varies from 1 to any maximum threshold.
- Any distance based algorithm like kNN can easily gets affected by such variations in scale.

Age	Loan amount
26	200500
55	200500
28	210000
45	210000

Euclidean distance of (25, 200000) w.r.t others
500.000999999
500.89919145472777
10000.00044999999
10000.01999998