Sneha Rao
1032170779

Camlin Page
Date / /

LAB ASSIGNMENT - 1

AIM - To study data collection, data preparation, data handling and perform exploratory data analysis

OBJECTIVES - 1) To learn ~~data~~ python programming of different modules/ libraries
2) To understand the concept of EDA (Exploratory data analysis)

THEORY -

Data collection →
Collection of data is the most fundamental step in any data science process. There are many different ways of collecting data

1) Building dataset from scratch
Advantages -
- Features are included based on the purpose of the research question or task. Not vice versa.
- This helps to only use meaningful data
- It is traceable how the variables were created.

Disadvantages
- It can be challenging to find suitable sources.
- It takes a lot of time to gather data.

- Transforming the data features into right format can be lot of effort.

2) Using govt websites.

Advantages.

- Commonly high data quality. Data used by other researchers or practitioners.
- Often used data is well documented. Therefore, one can understand how the variables were created.

Disadvantages

- It can take a lot of time to gather & transform features
- It can be hard to understand the datasets
- Access is sometimes only given on request basis.

3) Accessing private datasets

↳ Need to substantiate quality of data before using it in any professional projects.

→ Various types of data

- Numerical

It represents quantitative measurement.

Ex - Height of people, stock prices.

- Discrete data.

Integer based, often counts of something.

Ex - How many times did I toss "Heads"?

— Continuous data.
It has an infinite ~~data~~ no of possible values. Ex - How much rainfall on given day!

— Categorical data
Qualitative data, Ex - genders, Yes/No, etc.
Assign some no to categorical data but they don't have any mathematical meaning.

— Ordinal data
Mixture of ~~non~~numerical & categorical data. Categorical data has mathematical meaning. for ex - movie rating of on scale of 1-5. Rating must be 1 2 3 4 5!
They have mathematical meaning, like rating 1 movie is worse than rating 2 movie.

Label encoding →.
In label encoding, each category is mapped to a no or a label. The label chosen for the categories have no relationship. So categories that have some ties or are close to each other lose such info after encoding. It supports the pandas dataframe as i/p & can transform data.

Camlin Page
Date / /

Onehot encoding →
A one hot Encoding allows the representation of categorical data to be more expressive. Many machine learning algos can't work w/ categorical data directly. The categories must be converted into numbers.

Operations to be performed on dataset →

Steps in preprocessing of data
1) Importing python modules/libraries
2) Importing data
3) Displaying data
4) Creating the independent & dependent vars
5) Replacing missing value of meaningful value
6) Encoding categorical data
7) Splitting the data into training & testset
8) Doing feature scaling on data
9) Use only 3-4 graphs/plots

Dataset used →

Titanic - A kaggle competition dataset that is used for ML competition aimed at predicting who might have survived the sinking. It includes passenger info like class embarkment port, fare, cabin etc.

Camlin Page
Date / /

FAQs

Q1) List two common libraries for data manipulation. Give an example for each library.

→ The two libraries are-

(1) Pandas

Ex- dataset = pd.read_csv("train.csv")
dataset.dropna()

(2) Scikit learn

Ex- X = sklearn.preprocessing.StandardScaler()
fit_transform(X)

Q2) Give an example on how ordinal data is handled in ML algorithm.

→ To handle ordinal categorical data, we assign them numbers tantamount to the significance of it. - This can be done using a technique called LabelEncoding.

Ex-

Spiciness = ['low', 'high', 'medium', 'low', 'low', 'high']

Encoded = sklearn.preprocessing.LabelEncoding(Spiciness)

Encoded = [0, 2, 1, 0, 0, 2]

Q3) Can one hot encoding be used for continuous data. If yes, give an example

→ The main purpose of OneHotEncoder is to encode categorical data, which therefore can't be directly applied to continous data. Although if we can convert the continous data into nominal data, we can practically use one hot encoding on it. One method of doing this is called binning.

Ex -

Age = [9, 25, 27, 17, 15, 76, 14]

Binned ages = ["1-10", "20-30", "~~30-50~~", "20-30", "10-20", "10-20", "70-80", "10-20"]

This one can now be one hot encoded.

Q4) Why is it necessary to encode str?

→ Most ML models are mathematical, i.e they perform mathematical calculations to predict the o/p. Therefore, all the ip(s) are need to be numeric. Therefore, we need to be encode the data & convert them to numbers before passing them through the model.

Q5) State the significance of EDA?

→ Although ML models can learn most of patterns themselves, there is some amount of processing & manipulation required before passing it through the process. We need some info like -

- What columns are more significant
- How many classes does the dataset have
- Is there is a statistically significant relationship b/w columns & the target is noticed trend statistically significant.
- Are there any missing value or outliers.
- Are there any problematic trends like multi-collinearity. or dummy-variable trap.

To get ans to questions like these, we need EDA. This includes methods like
- Graphs & plots
- Ad Hoc testing & comparison
- Clustering
- Chi-squared testing.

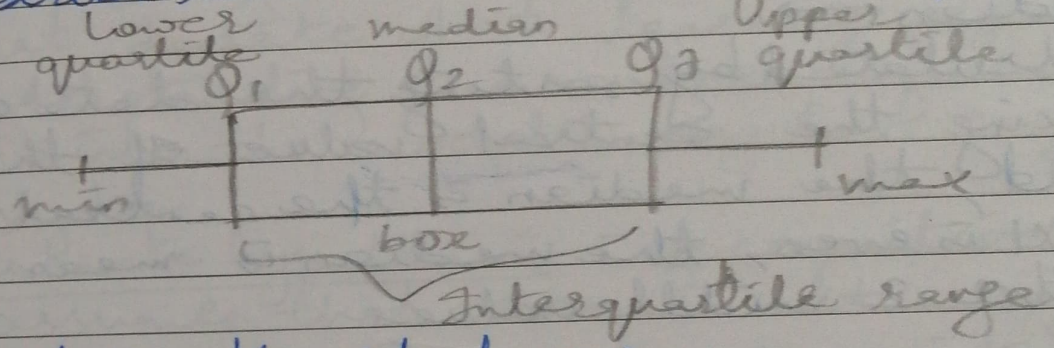
Q6). 'Handling missing values of data is an imp step in data preprocessing'. Comment on the statement.

→ Missing data in the dataset is represented using NaN. This value can't be processed by any type of ML model. Therefore these need to be handled in order to proceed w/ creating a model. The most common techniques to drop the records w/ a NaN value in any column. However this method then leads to loosing a lot of data in some cases. This can then be handled using

methods like -
- Imputing w/ mean, median or other central tendency
- Interpolating the data
- Filling the missing slots w/ 0 or missing category.

Q7) State any 4 techniques/plots used for EDA?
→ Exploratory data analysis charts.
- Correlation Heatmap
- Boxplot
- Histogram
- Barplot

Q8) Describe the box-and-whisker plot.
→



→ Making the plot →
The whiskers are first drawn at the min & max points of the data. The first split (Q2) is made at the median of the data column. The next 2 splits are made at the median b/w the min & Q2 (Q1) & Q2 and the max (Q4). The difference b/w Q4 and Q1 is called interquartile range. The range of the data is the diff b/w Maxm & Minm

Interpreting the plot →
Values lying in the interquartile range are considered normal. If value lies more than one & a half times the length of the box from either end, the values is deemed to be outliers.

Q9) Explain central tendency funct
→ Mean →
Central value of a discrete set of numbers

$$\text{mean} = \frac{\Sigma X}{|X|}$$

Median →
Calculated by sorting the data and taking the central value. If the |X| is odd the median is the central value. If |X| is even, the median is mean of the 2 middle values.

Mode →
Value in a distribution that occurs most

CONCLUSION

Data collection, data preparation, handling various data types was studied & EDA was performed.