Monday, 8 December 2025 6:01 PM

Recap.

- Python
  - problem solving skills.
  - OOPs. (revision).
  - Insight OOPs in real time.

- EDA

Data Analysis
- Univariate & Bi-variate —— why?
  - Univariate: Non-visual, visual
  - Bi-variate: Non-visual, visual.

Why?
- To understand the data
- To gain insights from the data } Uni-variate.
- find relationship b/w two - variables } bi-variates

Purpose → Decision → Data driven decision

Data Driven decision
- Analysis of the data
- Data distribution. — aid (helping tool)

Univariate
- Feature selection
- Feature engineering.
- Feature transformation
- Missing value detection
- Outliers.
- Relation b/w variables. → bivariate

* Data Distribution

skewness → left skewed or right skewed.
→ property of the distribution.

- Gaussian distribution
- Normal distribution } most common term.
  → continuous data

Data

- Discrete data
  - Binomial distribution
  - Bernoulli distribution
  - Poisson distribution
  - Multi model distribution
- Continuous data
  - Log normal distribution
  - Pareto distribution
  - Uniform data
  - Normal data.
  - Exponential distribution
  - Gamma distribution

2-3 distributions

- **Bernoulli Distribution**

My data follows Bernoulli's distribution
→ it exactly contains 2 possible values.
→ one value represent a success
→ other value represent a failure

Eg.
Yes / No
0 / 1
True / False

} Cardinality. → no. of unique value present in a column
→ In bernoulli distribution → 2.

→ bernoulli distribution
In logistic regression
→ by default → we cannot perform multi-class classification.
→ it can only perform binary classification

why?
Assumption of vanilla (basic) logistic regression
- Target variable has a cardinality = 2.

Why to learn data distribution.
- Better understanding of ML.
- Stronger at data analysis.

The more you understand distribution, the deeper your understanding of data science becomes.

- **Applications**
  - Better data driven decision. ✓
  - Better outlier handling. ✓
  - Better feature engineering. ✓
  - Stronger ML modelling.

22 - age
220 - 30 → 190
50 - 45 → clean

- Data distribution is not optional. → they are foundational.

In classroom

Logistic regression (basic) → Binary classification

→ multi-class classification.

addition strategies

- → OVR (One v/s Rest).
- → OVO (One v/s One).

Target

2 unique value

Ist half. → One v/s Rest

1 v/s 2+3

→ 2 | 2

Basic logistic regression inherently relies on mapping the target variable to Bernoulli distribution, which requires only two classes. It fails to do multi-class classification is used directly.

We have to use the OVR and OVO mechanism along with logistic regressions to do multi-class classification.

- Real time Example

Pizza Store dataset.

→ cardinality = 2.
(Bernoulli's distri

Order-id	Pizza.	Soft drink
1	Margherita	Yes → Success
2	Veggie Delight	No → Failure
3	Paneer Special	Yes
4	Margherita,	No
5	Farm house	No

5

2

- What is the probability of people buying a soft-drink with their pizza.

$P(\text{soft drink} = \text{Yes})$

$$P(\text{event}) = \frac{\text{No. of favourable outcomes}}{\text{Total no. of outcomes.}}$$

Favourable outcomes = 2.
Total no. of outcomes = 5.

$P(\text{soft drink} = \text{Yes}) = \frac{2}{}$

5

- Probability of failure (soft drink = No)

$$P(\text{Failure}) = 1 - P(\text{Success})$$
$$= 1 - \frac{2}{5}$$
$$= \frac{3}{5}.$$

In the next 3 orders

what is the probability that exactly 2 customers will order a soft drink with their pizza.

→ what is the probability that 2 out of 3 customers will order a soft drink with their pizza. → combination → Bernoulli distribution → Binomial distribution

- Binomial distribution

  - A collection of Bernoulli trials.

Bernoulli's trial → One event with two outcomes (success | failure).

Formula

$$P(\text{exactly } i^{th} \text{ success in a } n \text{ trial}) = {}^nC_x\, p^x (1-p)^{n-x}.$$

where n = no. of trials
x = no. of successful trials.
p = probability of success
$(1-p)$ = probability of failure.

$${}^nC_x = \frac{n!}{x!\,(n-x)!}$$

$n = 3$
$x = 2.$
$p = \frac{2}{5}.$
$(1-p) = \frac{3}{5}.$

$${}^3C_2 \cdot \left(\frac{2}{5}\right)^2 \left(\frac{3}{5}\right)^{3-2}.$$

$$\frac{3!}{2!\,(1)!} \cdot \left(\frac{2}{5}\right)^2 \left(\frac{3}{5}\right)^1$$

$$\frac{3 \times \cancel{2!}}{\cancel{2!}} \times \frac{4}{25} \times \frac{3}{5}$$

$$\Rightarrow \frac{3 \times 4 \times 3}{25 \times 5} = \frac{36}{125}. \approx 0.288$$

→ 22 nd → 1 now.