Foundation of Data Science using Programming Language

Finalizing Data Science Problems

Dr. Abid Sohail

Associate professor

abidbhutta@gmail.com

Data Science Problems classification

Live Generative AI:

- **Interactive Content Generation:**
  - *Definition:* Creating content that responds to user input or changes dynamically.
  - *Applications:* Video game environments, dynamic website content, and interactive art installations.
- **Conversational Agents:**
  - *Definition:* Real-time generation of human-like responses in natural language.
  - *Applications:* Chatbots, virtual assistants, and customer support systems.

Predictive Analytics:

- **Time Series Forecasting:**
  - *Definition:* Predicting future values based on historical data that is ordered chronologically.
  - *Applications:* Stock price forecasting, weather predictions, and energy consumption forecasts.
- **Anomaly Detection:**
  - *Definition:* Identifying unusual patterns or outliers in data.
  - *Applications:* Fraud detection, network security monitoring, and equipment failure prediction.

Classification:

- **Imbalanced Classification:**
  - *Definition:* Dealing with datasets where one class is significantly underrepresented.
  - *Applications:* Fraud detection (where fraud cases are rare), rare disease diagnosis.
- **Multi-label Classification:**
  - *Definition:* Assigning multiple labels or categories to each instance.
  - *Applications:* Image tagging, topic categorization, and news article classification.

Regression:

- **Non-linear Regression:**
  - *Definition:* Modeling relationships between variables that cannot be represented by a straight line.
  - *Applications:* Growth modeling, predicting complex biological processes.
- **Time-to-Event Prediction:**
  - *Definition:* Estimating the time until a specific event occurs.
  - *Applications:* Predicting customer churn time, time until system failure.

Clustering:

- **Density-Based Clustering:**
  - *Definition:* Grouping data points based on their density in the feature space.
  - *Applications:* Identifying regions of high and low population density in spatial data.
- **Hierarchical Clustering:**
  - *Definition:* Creating a tree of clusters to represent relationships at different levels.
  - *Applications:* Taxonomy creation, organizing documents in a hierarchical structure.

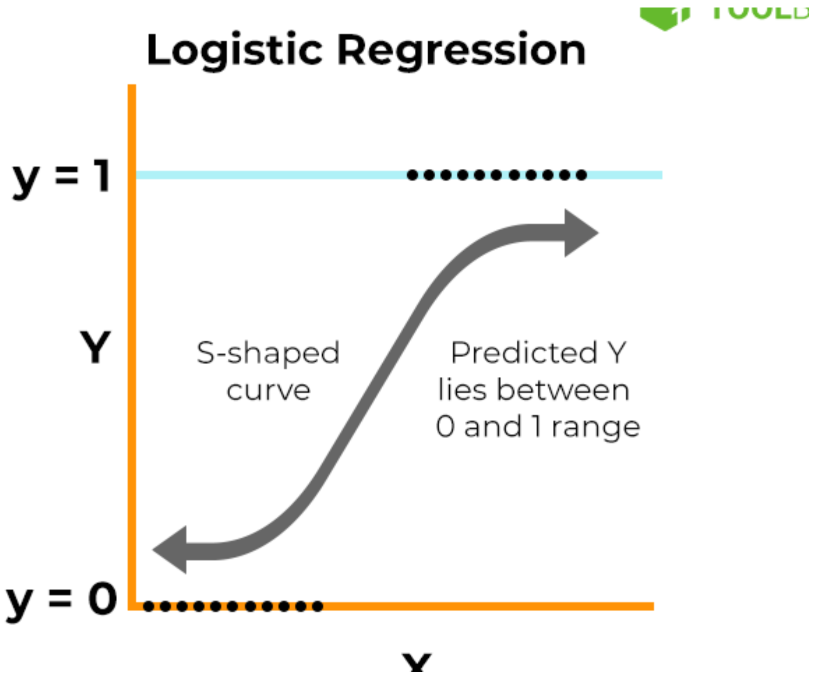
Logistic Regression

- Logistic Regression is a fundamental algorithm in data science, particularly in the field of classification. Let's break down Logistic Regression in
- **Type:** Supervised learning algorithm.
- **Task:** Classification. (GOOD for Binary)
- **Output:** Probability of belonging to a particular class.

Logistic Regression

- x = input value
- y = predicted output
- b0 = bias or intercept term
- b1 = coefficient for input (x)

$$y = \frac{e^{(b_0 + b_1X)}}{1 + e^{(b_0 + b_1X)}}$$



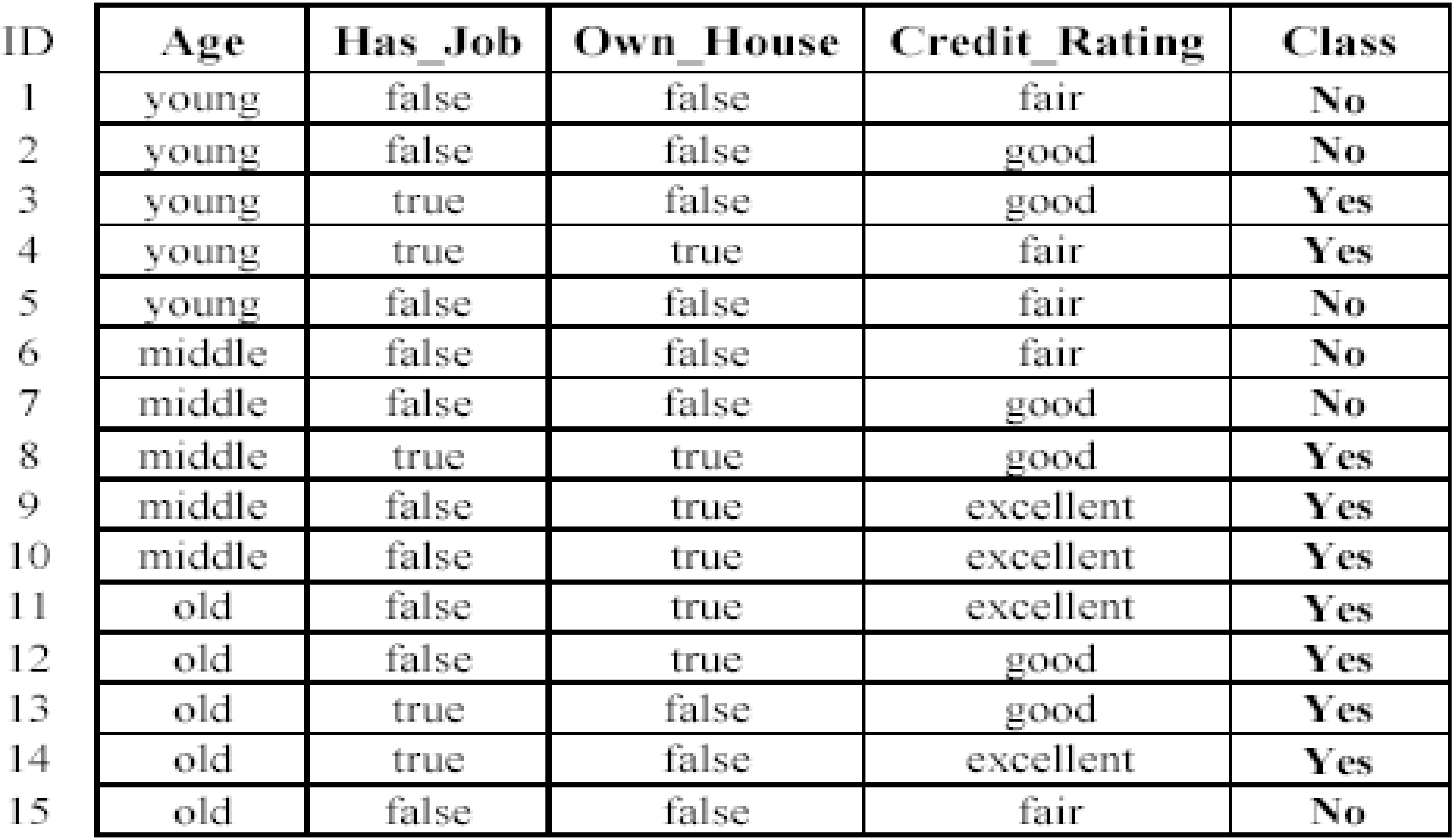
Application in Data Science

- **Binary Classification:**
  - Logistic Regression is commonly used for binary classification problems (two classes).
  - Examples include spam detection, whether a customer will buy a product (yes/no), etc.
- **Multiclass Classification:**
  - Logistic Regression can be extended to handle multiple classes using techniques like one-vs-rest or one-vs-one.
  - Applications include handwritten digit recognition, sentiment analysis with multiple classes, etc.
- **Probability Interpretation:**
  - The output of Logistic Regression is interpreted as the probability of the input belonging to a particular class.
  - A threshold (commonly 0.5) is chosen to classify instances into one of the classes.

An example: data (loan application)

Binary

Approved or not



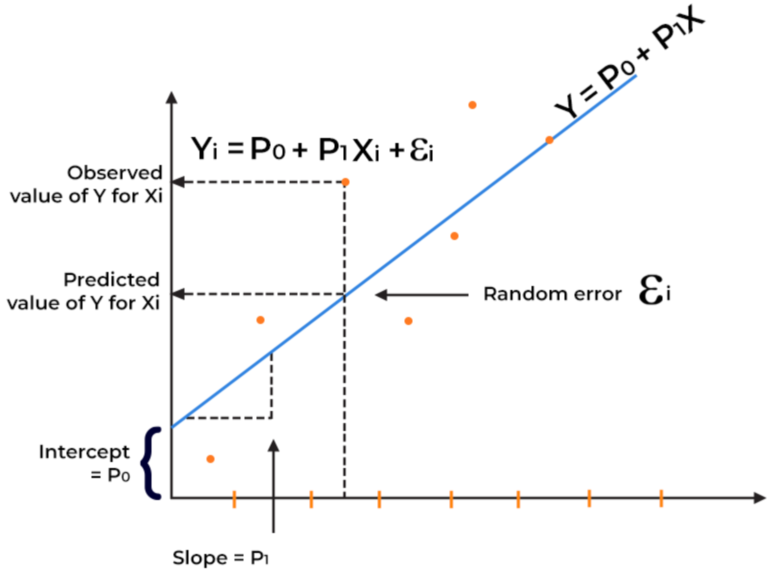
ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Linear Regression

- **Type:** Supervised learning algorithm.
- **Task:** Regression (predicting a continuous outcome).(Co-relation based)
- **Output:** Continuous numerical values.

Linear Regression

- The formula for simple linear regression is
  - Y = mX + b,
  - where Y is the response (dependent) variable, X is the predictor (independent) variable, m is the estimated slope, and b is the estimated intercept.


$Y = P_0 + P_1X$
$Y_i = P_0 + P_1X_i + \varepsilon_i$
Observed value of Y for Xi
Predicted value of Y for Xi
Random error ε_i
Intercept = P_0
Slope = P_1

Application in Data Science:

- **Prediction of Continuous Values:**
  - Linear Regression is used for predicting a continuous outcome based on input features.
  - Examples include predicting house prices, stock prices, or student exam scores.
- **Relationship Analysis:**
  - It helps in understanding the linear relationship between independent and dependent variables.
- **Feature Importance:**
  - Coefficients in the linear equation can indicate the importance of each feature in predicting the target variable.

Example Dataset: Salary Prediction

Years of Experience	Education Level (1-10)	Salary (in thousands of dollars)
1	3	40
2	4	45
3	3	50
5	5	60
7	6	70
8	7	80
10	8	95
12	9	110
15	10	130
20	10	150

- **Years of Experience:** The number of years of professional experience of the individuals.
- **Education Level (1-10):** A numerical rating representing the education level of the individuals, where 1 is the lowest and 10 is the highest.
- **Salary (in thousands of dollars):** The salary of the individuals in thousands of dollars, representing the dependent variable.

Market basket Analysis

Implementing Market Basket Analysis in Python

Clustering problem

Member_number	Date	itemDescription
1808	21-07-2015	tropical fruit
2552	5/1/2015	whole milk
2300	19-09-2015	pip fruit
1187	12/12/2015	other vegetables
3037	1/2/2015	whole milk
4941	14-02-2015	rolls/buns
4501	8/5/2015	other vegetables
3803	23-12-2015	pot plants
2762	20-03-2015	whole milk
4119	12/2/2015	tropical fruit
1340	24-02-2015	citrus fruit
2193	14-04-2015	beef
1997	21-07-2015	frankfurter
4546	3/9/2015	chicken

Member_number	**Date**	**itemDescription**	
1631	2222	08-01-2015	yogurt
3796	2222	21-07-2015	berries
4881	2222	28-12-2015	whole milk
8433	2222	28-12-2015	sausage
10571	2222	13-02-2014	grapes
11296	2222	03-04-2014	pork
14695	2222	31-07-2014	sugar
15709	2222	21-07-2015	other vegetables
17780	2222	08-01-2015	dental care
19945	2222	21-07-2015	butter
21030	2222	28-12-2015	pork
24582	2222	28-12-2015	coffee
26720	2222	13-02-2014	chewing gum
27445	2222	03-04-2014	rolls/buns
30844	2222	31-07-2014	seasonal products
31858	2222	21-07-2015	newspapers
33460	2222	22-02-2014	cling film/bags
35609	2222	13-02-2014	curd cheese
38011	2222	22-02-2014	canned vegetables