Logistic Regression

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Definition

Logistic regression is defined as a supervised machine learning algorithm that accomplishes **binary classification** tasks by predicting the probability of an outcome, event, or observation.

The model delivers a binary or discrete outcome limited to two possible outcomes: yes/no, 0/1, or true/false.

Logical regression analyzes the **relationship between one or more independent variables and classifies data into discrete classes.**

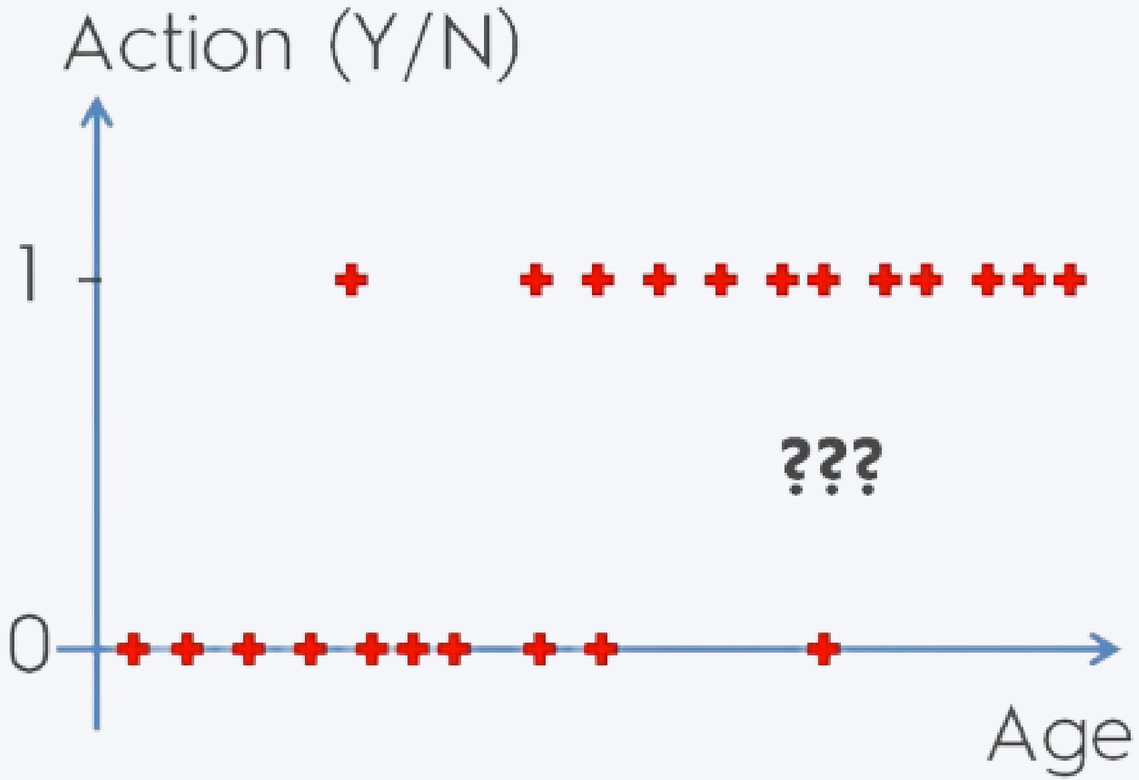
It is extensively used in predictive modeling, where the model estimates the mathematical probability of whether an instance belongs to a specific category or not.

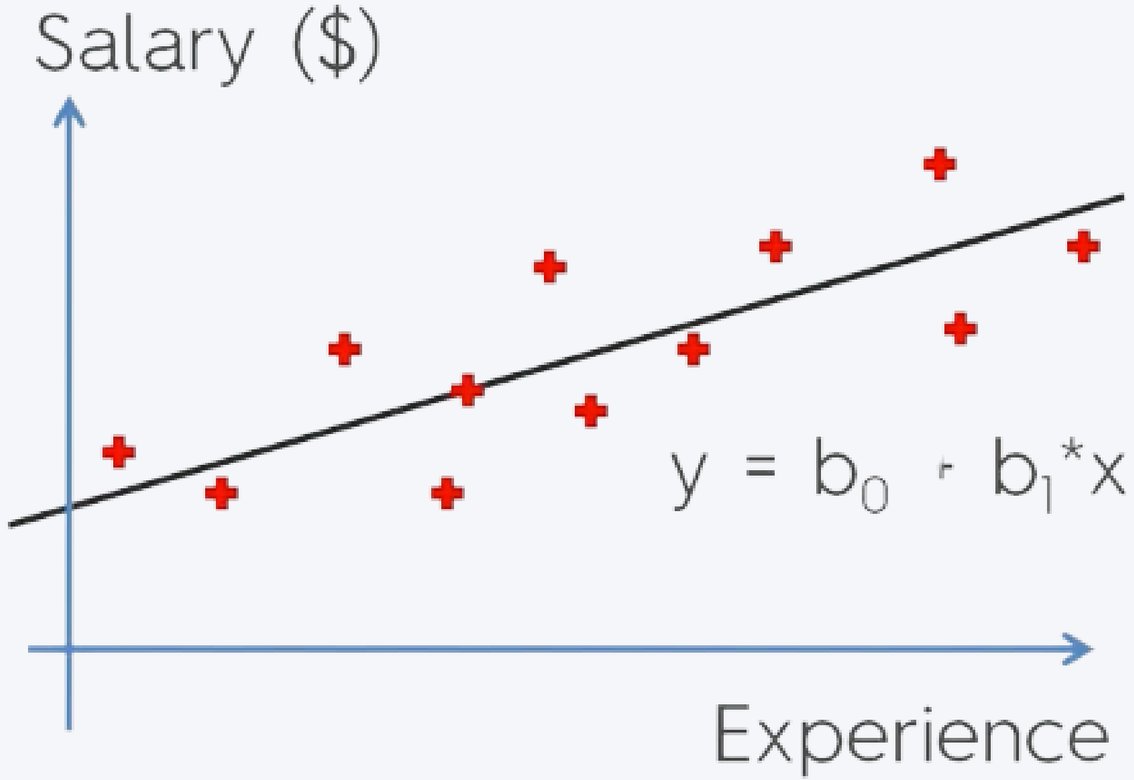
https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Understanding Logistic Regression





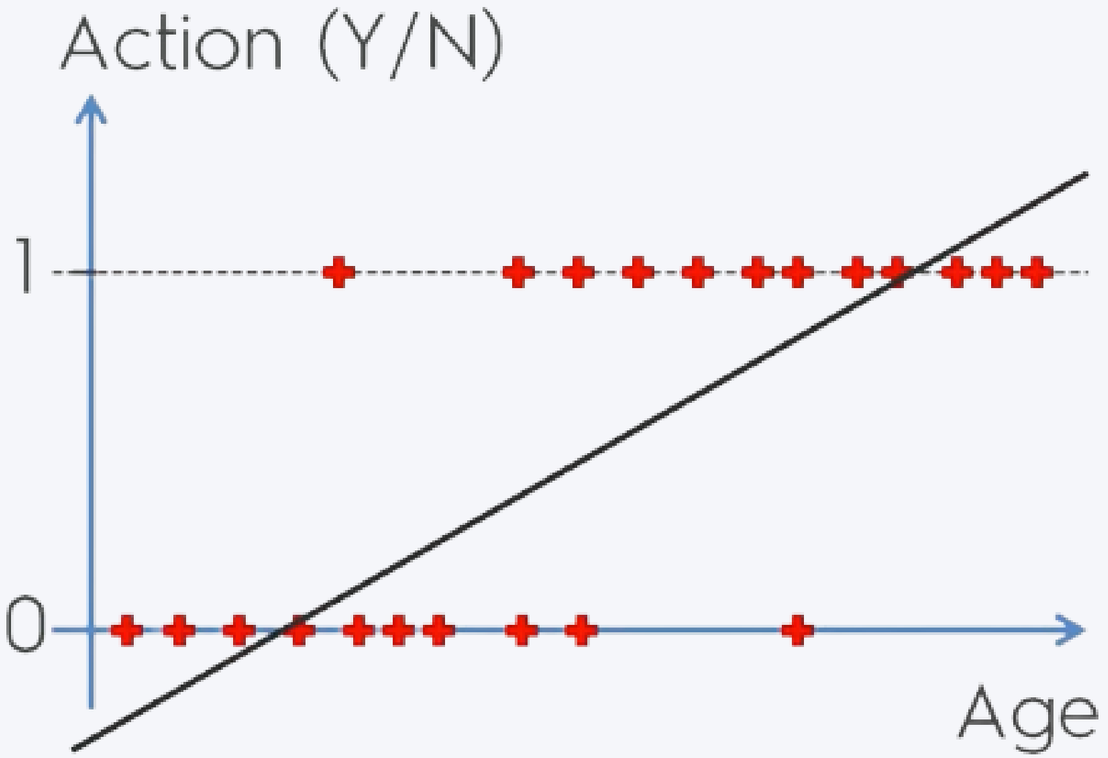
https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Understanding Logistic Regression

- Linear Regression Fails for Classification – It predicts continuous values instead of clear 0 or 1 labels.
- **Invalid Probability Outputs** – Predictions can go below 0 or above 1, which makes no sense for binary outcomes.
- **No Clear Decision Boundary** – It lacks a proper mechanism to classify points, unlike logistic regression, which uses a sigmoid function.

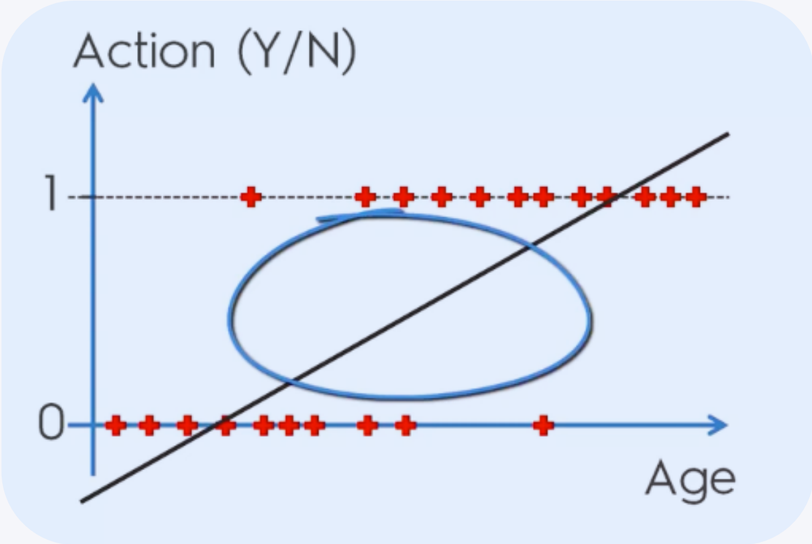


https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

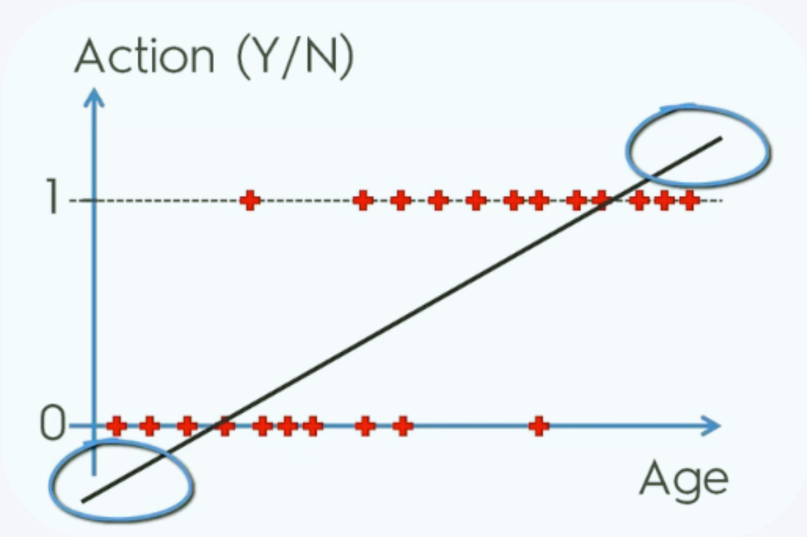
Logistic Regression

أكاديمية طويق
Tuwaiq Academy

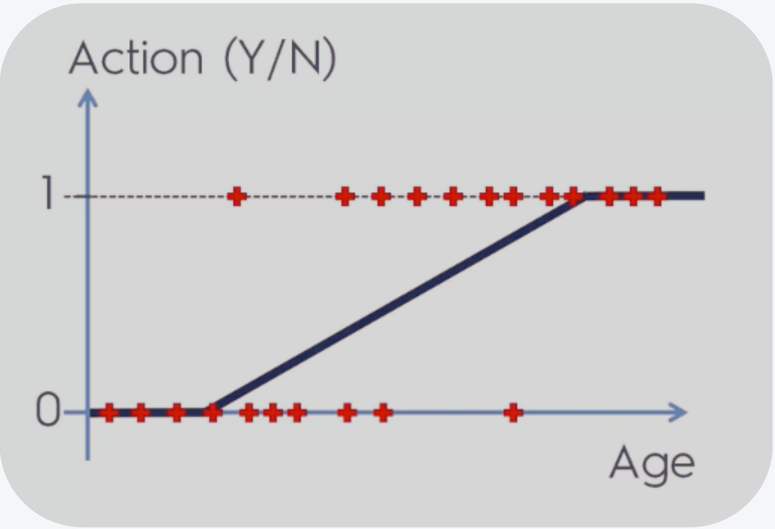
Understanding Logistic Regression



- Makes Sense – Some points in the middle could lean toward 0 or 1 but aren't strictly one or the other.
- Those in-between cases should have a probability, not a fixed value.



- Doesn’t Make Sense – Probabilities must stay between 0 and 1.
- Invalid Outputs – Values above 1 should be 1, and below 0 should be 0.



- So we kind of saying this modeling works

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

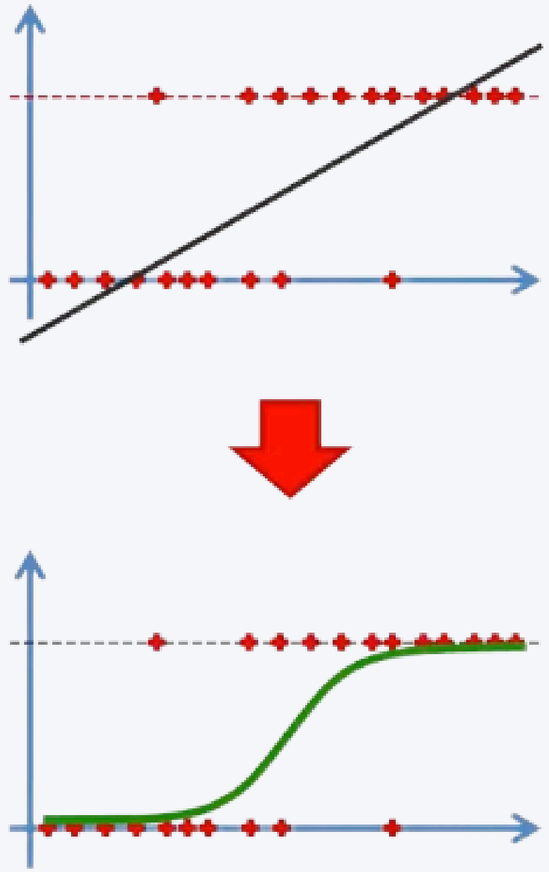
Understanding Logistic Regression

$$y = b_0 + b_1 * x$$

Sigmoid Function

$$p = \frac{1}{1 + e^{-y}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 * x$$



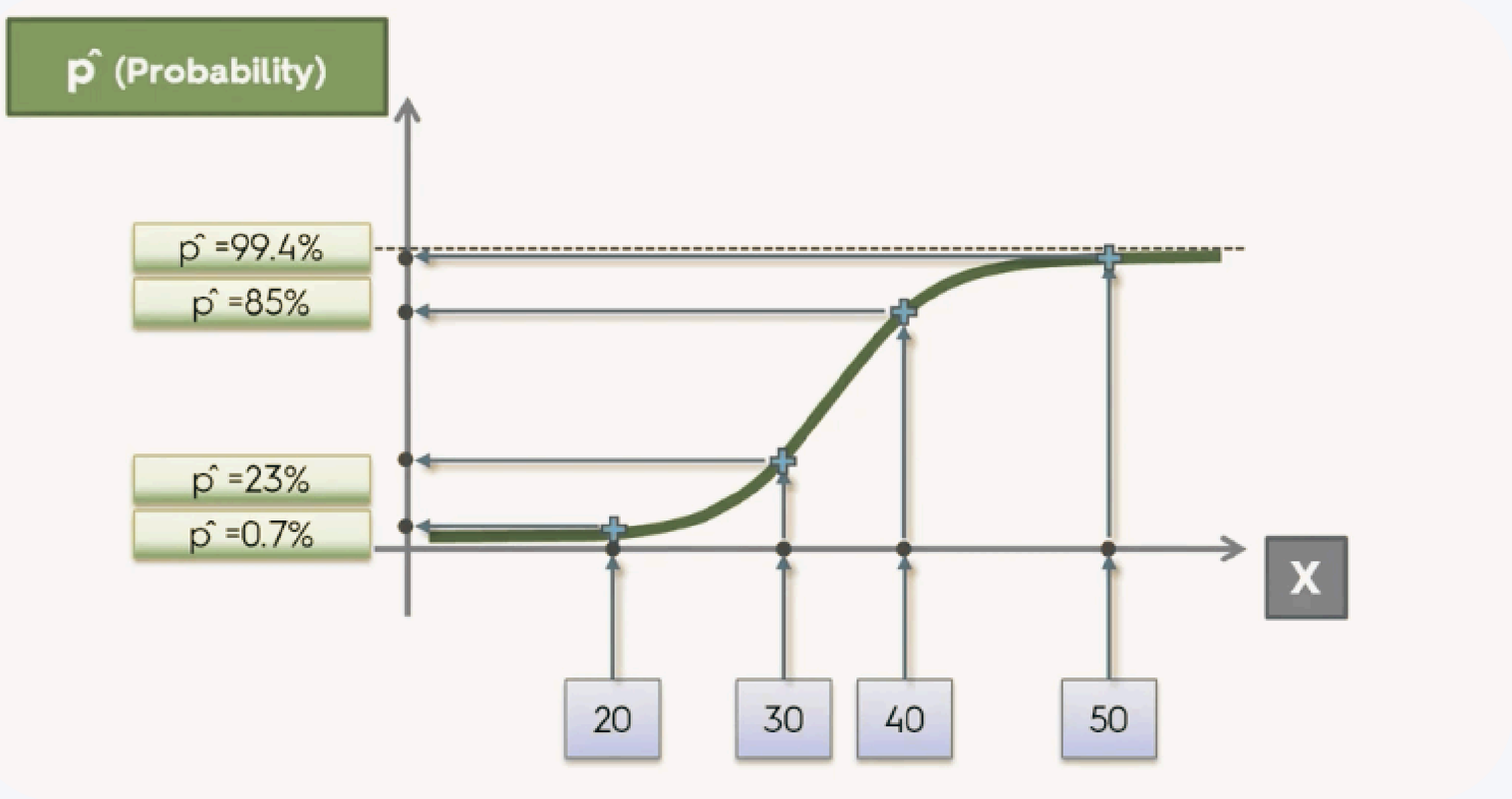
slope

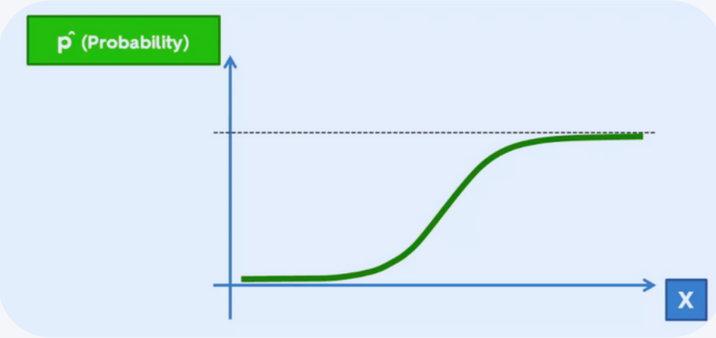
best fitting line

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

Understanding Logistic Regression



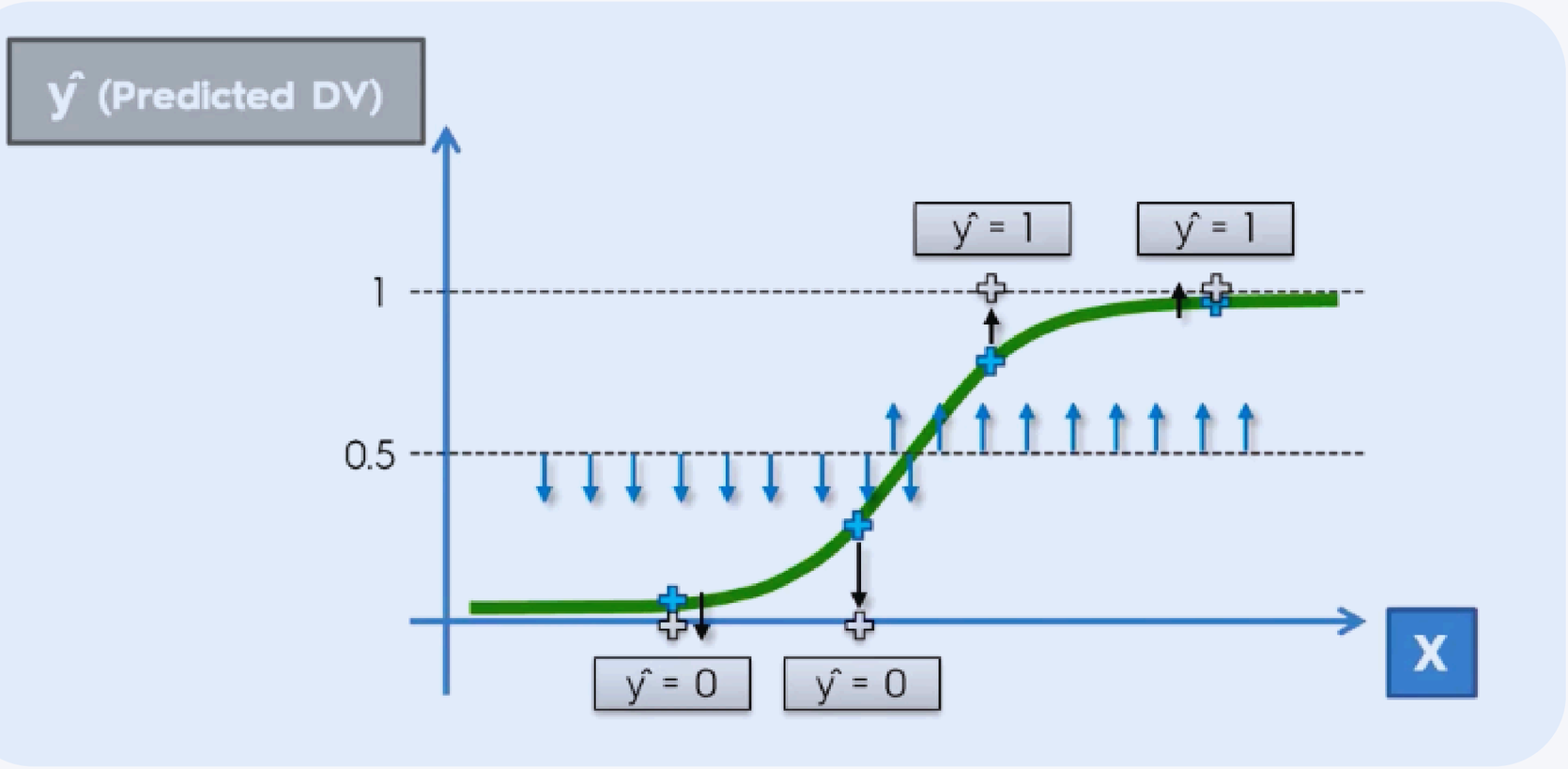


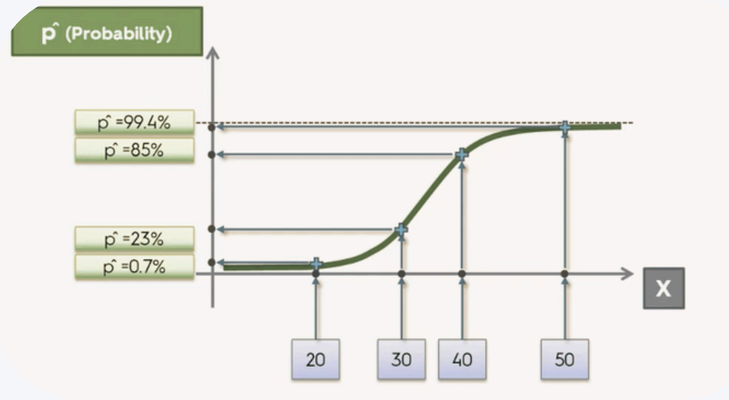
- This could represent probabilities since it is between 0 and 1
- Projecting blue values to each probabilities based on the training data
- The line represents the logistic regression fitting line (slope)

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

Understanding Logistic Regression





- Threshold at 0.5 – Anything below 0.5 is predicted as 0, and anything above is 1.
- Mismatch with Data – Predictions fall on 0 and 1, but not exactly where the data points are.
- Loss Optimization – The model tries to find the best-fitting line with minimal loss.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Key Assumptions For Applying Logistic Reg.

The dependent/response variable is binary or dichotomous

Little or no multicollinearity between the predictor/explanatory variables

Variance Inflation Factor (VIF), which determines the correlation strength between the independent variables in a regression model.

Linear relationship of independent variables to log odds

Log odds express probabilities in terms of the ratio of success to failure, while probability measures success out of total events. For example, if you win 5 out of 12 tennis games, your probability of winning is 5/12, but your odds of winning are 5 to 7 (wins to losses).

Requires sufficiently large sample size

No extreme outliers

Should have independent observations

Observations should be independent, meaning they shouldn't be influenced by or repeated from other observations in the dataset.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

Log Odds in Logistic Regression

Odds represent the ratio of success to failure. It is calculated as:

$$\text{Odds} = \frac{p}{1-p}$$

where:

- p is the probability of success (event happening).
- 1−p is the probability of failure (event not happening).

- If the probability of passing an exam is **0.75 (75%)**, the odds of passing are:

$$\frac{0.75}{1-0.75} = \frac{0.75}{0.25} = 3$$

  Meaning: **The odds of passing are 3 to 1** (3 successes for every 1 failure).

- If the probability of passing is **0.25 (25%)**, the odds are:

$$\frac{0.25}{1-0.25} = \frac{0.25}{0.75} = 1/3$$

  Meaning: **The odds of passing are 1 to 3** (1 success for every 3 failures).

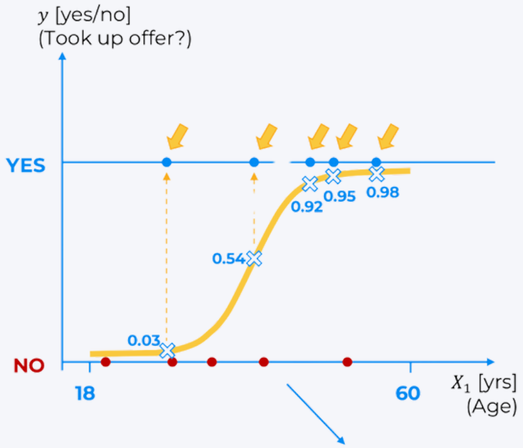
https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

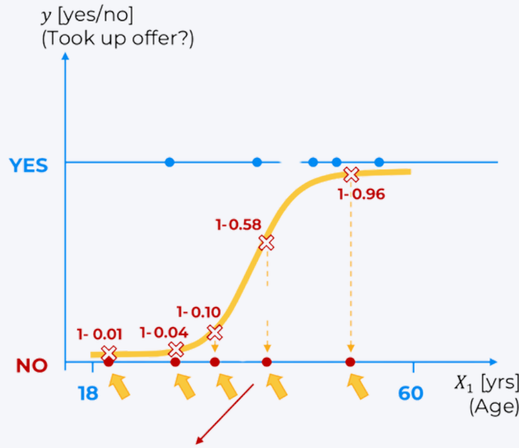
Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Maximum Likelihood

- **Logistic regression** calculates likelihood by assigning probabilities to observations and maximizing the likelihood of correct classification.
- The logistic curve (yellow) **aligns well** with the actual data points.
- Each observation has an assigned probability based on the curve.
- Since the probabilities for correct classifications are higher, the overall likelihood is higher.





- With (**Lower Likelihood Model)**

Likelihood = 0.03 x 0.54 x 0.92 x 0.95 x 0.98 x (1 – 0.01) x (1 – 0.04) x (1 – 0.10) x (1 – 0.58) x (1 – 0.96)

Likelihood = **0.00019939**

- The logistic curve does not fit the data as well.
- Incorrect probabilities are assigned to observations, resulting in a lower likelihood value.
- The likelihood is much smaller (0.00019939) compared to the left model.

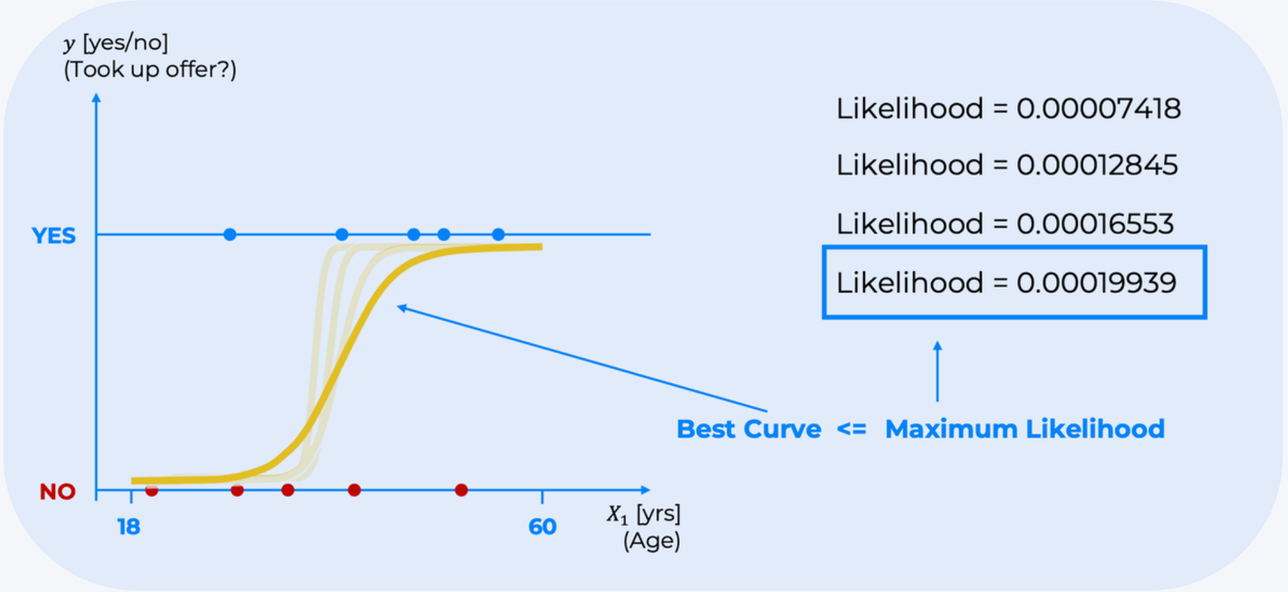
https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Maximum Likelihood

- The process of optimizing the logistic regression model to find the **best-fit curve.**
- Different logistic curves are **tested**, each corresponding to a different **likelihood** value.
- The goal is to find the curve with the **highest likelihood.**
- The highlighted likelihood 0.00019939 represents the best model, meaning this curve best separates the data into "YES" and "NO" categories.
- **Maximum Likelihood Estimation (MLE)** selects the best logistic regression model by finding the curve that maximizes the likelihood of correct classification.



https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression: Strengths and Limitations

- **Strength: Interpretability:**
  - Logistic regression is **easy** to understand because its coefficients directly show how each feature affects the log odds of the outcome.
  - Each coefficient represents the **change in log odds** for a one-unit change in the feature.
  - This makes it useful for domains like healthcare, finance, and social sciences, where explainability is crucial.

In a model predicting heart disease risk:

- A coefficient of +0.5 for cholesterol level means higher cholesterol increases the odds of heart disease.
- A coefficient of -0.3 for exercise frequency means more exercise decreases the odds.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression: Strengths and Limitations

Limitation: Struggles with Non-Linear Relationships

- Logistic regression assumes a linear relationship between **features** and **log odds**, which means it can't easily capture complex patterns in the data
- If the relationship between input variables and the outcome is non-linear, logistic regression won't perform well.

- Logistic regression performs poorly when **classes are overlapping or not easily separable.**
- If the data points of two classes mix heavily, logistic regression struggles to find a clear decision boundary.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

Logistic Regression

أكاديمية طويق
Tuwaiq Academy

Handling Multiple Categories in Logistic Regression

Logistic regression is primarily designed for binary classification (0 or 1). However, when dealing with multiple classes(e.g., classifying emails as Primary, Social, or Promotions), we need extensions of logistic regression.

There are two main approaches:

Multinomial Logistic Regression (Softmax Regression)

- This method directly extends logistic regression to multiple classes.
- Instead of using a single sigmoid function, it uses the softmax function to calculate the probability of each class.
- The model predicts one class out of many based on the highest probability.

One-vs-All (OvA) Approach

- Instead of using a single model, OvA trains multiple binary logistic regression models.
- Each model separates one class from the rest.
- The final prediction is made by choosing the model with the highest probability.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

أكاديمية طويق
Tuwaiq Academy

Hands-On Code

Logistic Regression Implementation

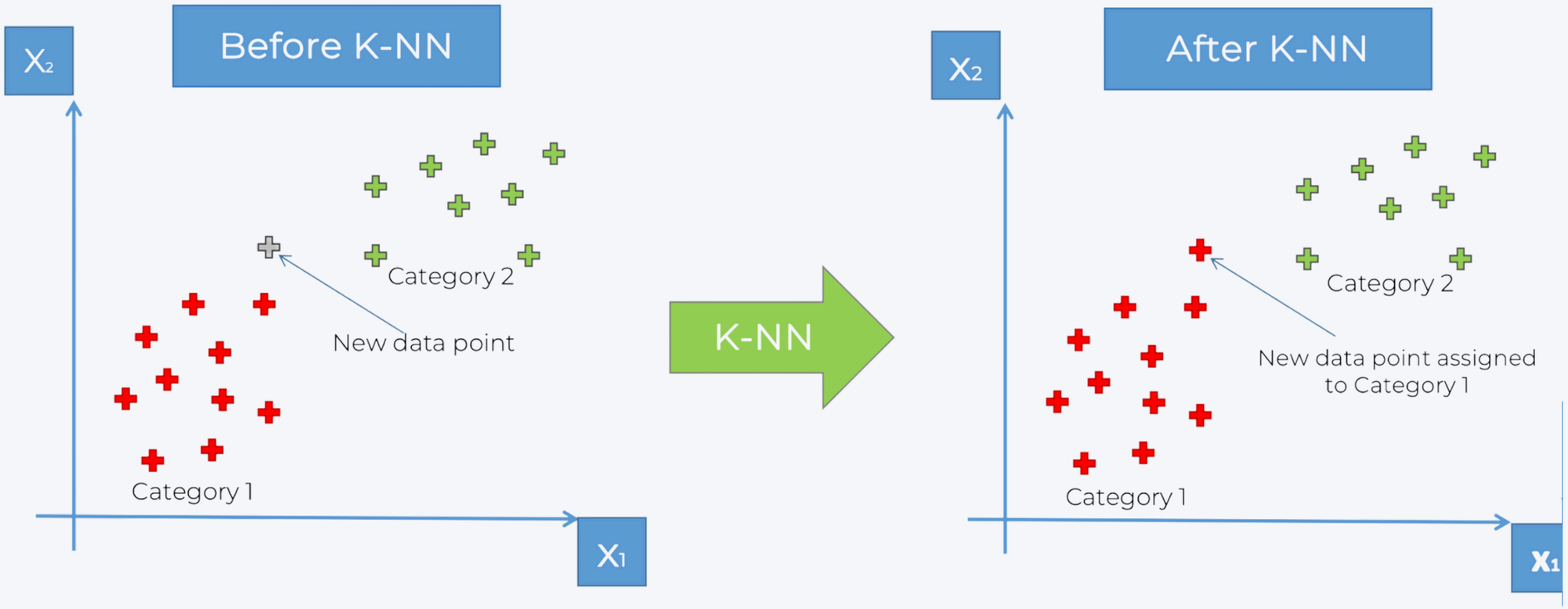
K-Nearest Neighbors (K-NN)

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

Definition

K-Nearest Neighbors (K-NN) is a non-parametric, instance-based learning algorithm used for classification. It's simple, intuitive, and often works surprisingly well for many datasets.



https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

How does it work !

K-NN makes predictions based on **similarity**. Instead of learning explicit patterns from the training data, it stores all the training data and makes decisions based on the **most similar examples** when a new data point arrives.

🔹 Steps for Classification:

1. Choose a value for **K** (the number of neighbors to consider).
2. Calculate the **distance** between the new data point and all existing data points.
3. Select the **K** closest points based on the chosen distance metric.
4. **Assign** the most common class (majority vote) among these K neighbors to the new data point.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

How does it work !

K-NN makes predictions based on **similarity**. Instead of learning explicit patterns from the training data, it stores all the training data and makes decisions based on the **most similar examples** when a new data point arrives.

🔹 Steps for Classification:

1. Choose a value for **K** (the number of neighbors to consider).
2. Calculate the **distance** between the new data point and all existing data points.
3. Select the **K** closest points based on the chosen distance metric.
4. **Assign** the most common class (majority vote) among these K neighbors to the new data point.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

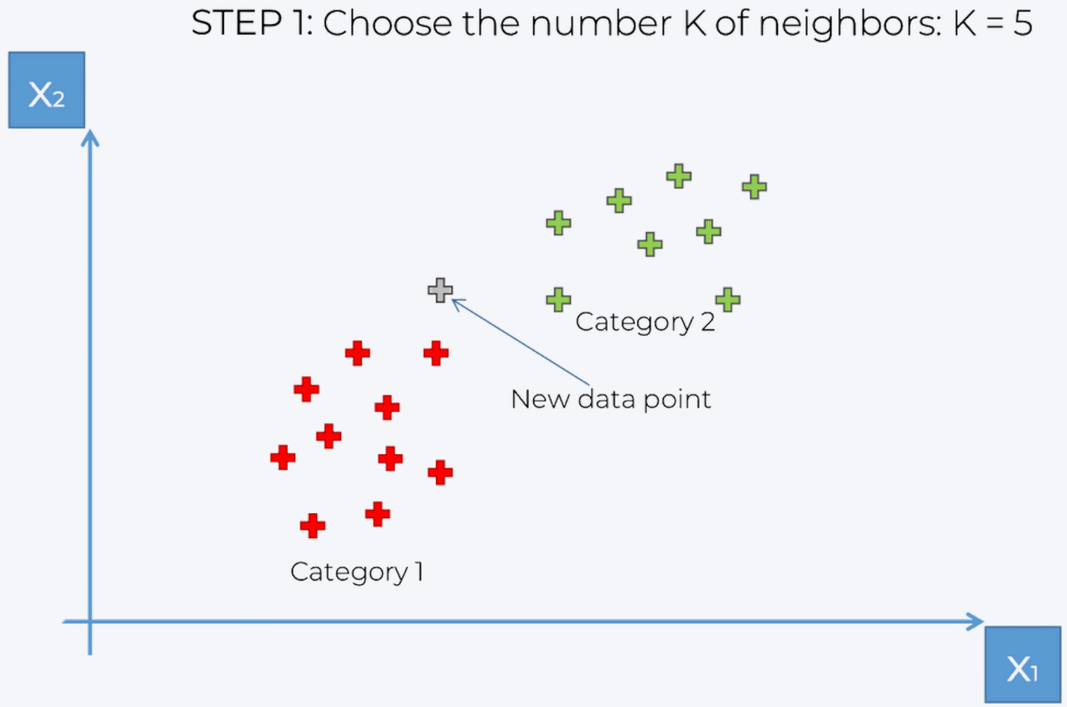
أكاديمية طويق
Tuwaiq Academy

How does it work !

Choosing the Value of K:

- **Small K (e.g., 1 or 3):**
  - More sensitive to noise.
  - High variance, meaning it can change significantly with small dataset variations.

- **Large K (e.g., 10 or 20):**
  - More generalized but may ignore local patterns.
  - Reduces overfitting but can smooth out important details.



A common approach is to try different K values and use cross-validation to find the best one.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

How does it work !

Since K-NN finds the "nearest" neighbors, we need a way to measure distance.

- **Euclidean Distance (Most Common)**

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- Used for continuous numerical data.

- **Manhattan Distance**

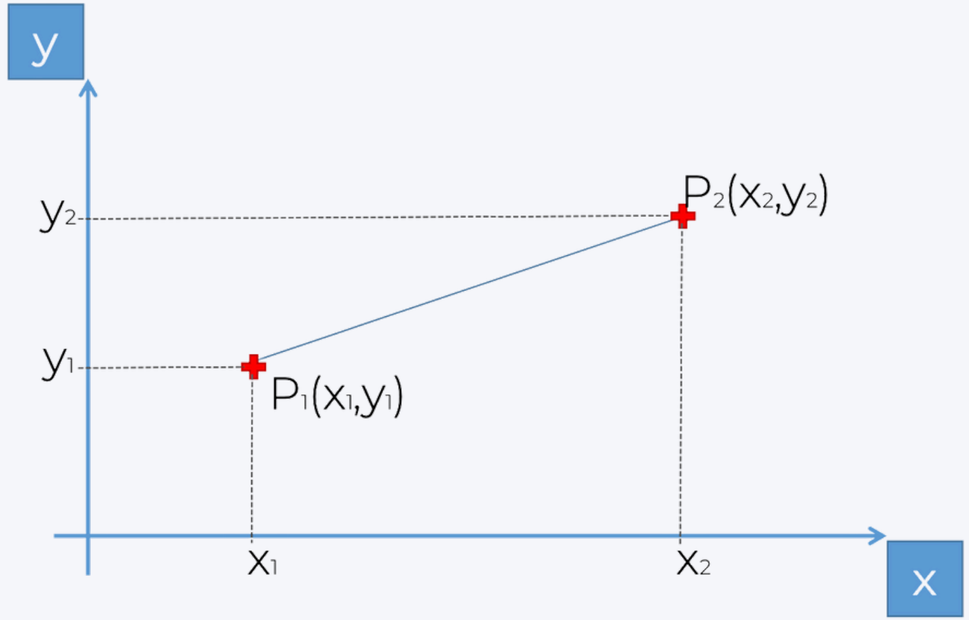
$$d = |x_2 - x_1| + |y_2 - y_1|$$

- Used when movement is restricted to horizontal and vertical paths.

- **Hamming Distance**

- Used for categorical data (e.g., DNA sequences or text classification).

Choosing the right distance metric depends on the type of data.



https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

How does it work !

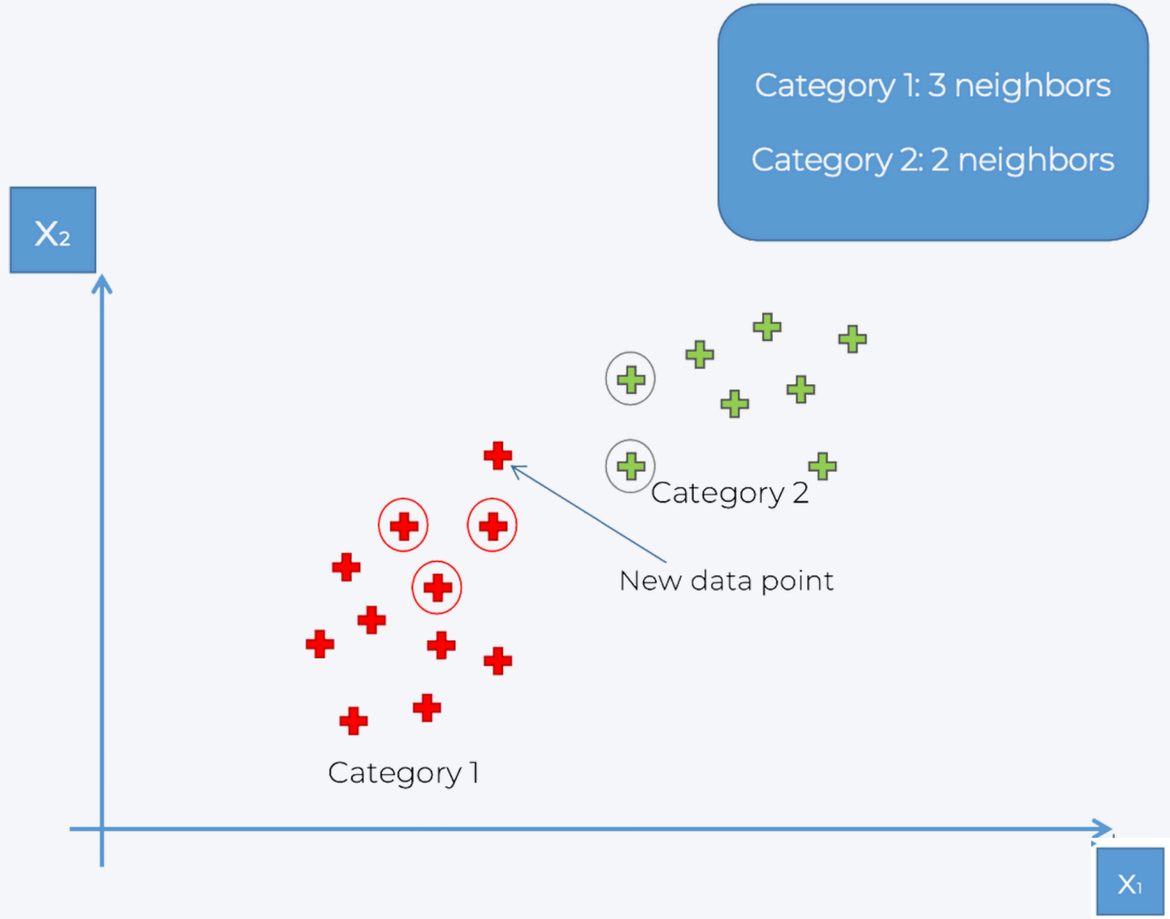
3. Find the Nearest Neighbors (K Closest Points):

- **W**e are using **K = 5** (since we see 5 nearest neighbors considered).
- The algorithm selects the 5 closest points to the new data point.

Majority Voting (Classification Decision):

- Among the 5 neighbors:
  - 3 belong to Category 1 (red crosses).
  - 2 belong to Category 2 (green plus signs).

Since Category 1 has the majority (3 out of 5), the new data point is classified as Category 1 (red cross).



https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

Strengths of K-NN

- Simple & Easy to Understand – No training phase, just storing data and comparing distances.
- Works Well with Small Data – Effective for datasets with clear separation.
- Non-Parametric – Makes no assumptions about the data distribution.
- Can Handle Multi-Class Problems – Works for problems with multiple categories.

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

K-Nearest Neighbors (K-NN)

أكاديمية طويق
Tuwaiq Academy

Limitations of K-NN

- **Computationally Expensive for Large Datasets**
  - Since K-NN stores all data, it can become slow for large datasets.
  - Requires calculating distances for all points at prediction time.

- **Sensitive to Irrelevant Features**
  - If some features are not useful, they can mislead K-NN.
  - Feature selection and normalization (scaling data properly) are important.
  -

- **Struggles with Imbalanced Data**
  - If one class is much larger than another, K-NN may always favor the majority class.

- **Not Good for High-Dimensional Data**
  - In high-dimensional spaces, all points start looking equally distant (Curse of Dimensionality).

https://www.researchgate.net/publication/344389401_DS-Regression-04-SVR

أكاديمية طويق
Tuwaiq Academy

Hands-On Code

K-Nearest Neighbors