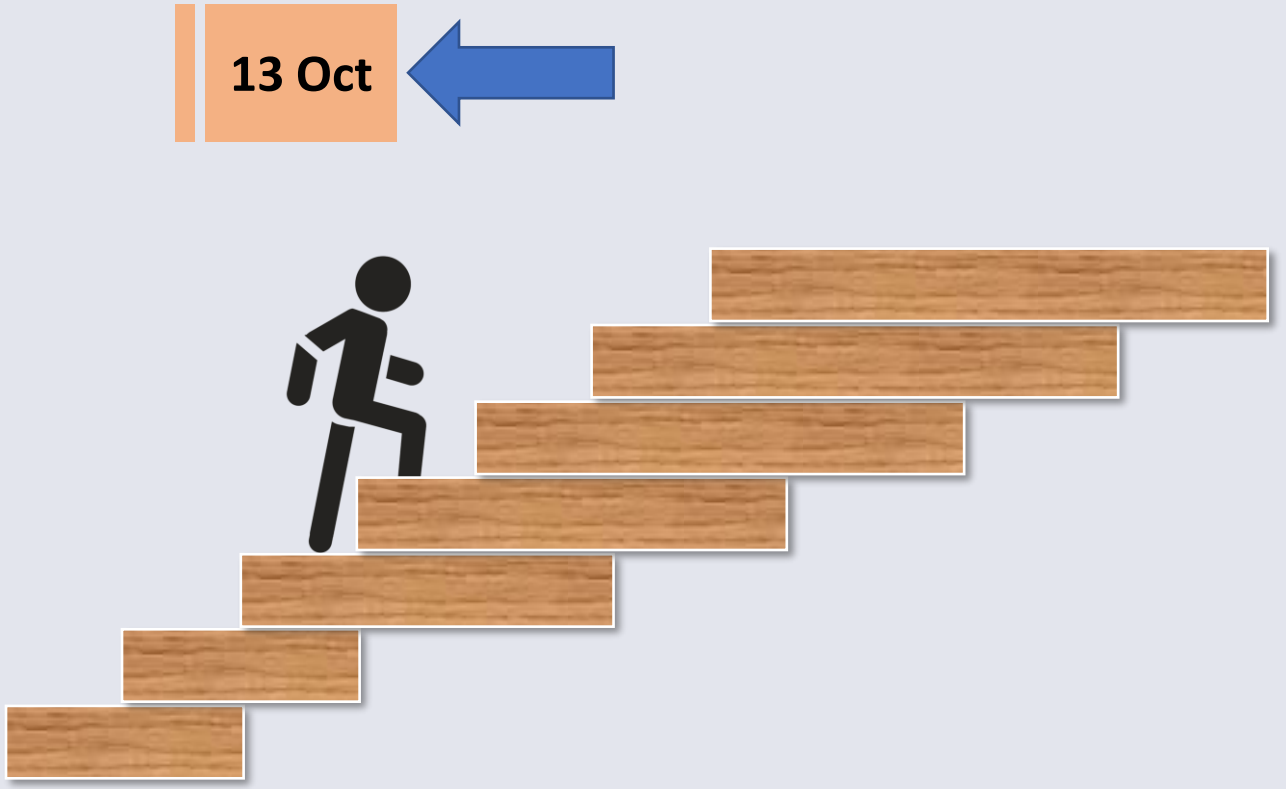
University of Lancashire

Lecture 4

Machine Learning-II

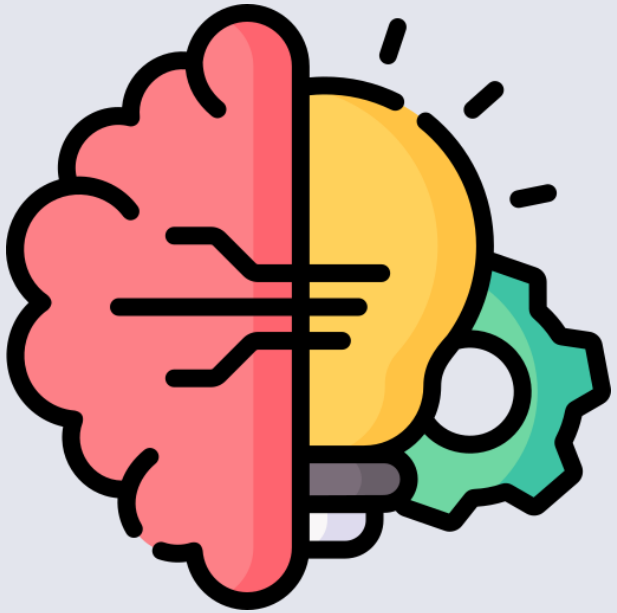
Previous Week Recap (..)

- Machine Learning ML
- Types of Machine Learning
- Relation between ML and AI
- Key steps in developing ML model



Today's Contents

- Key Challenges/Concepts
  - Overfitting vs Underfitting
  - Bias-Variance Tradeoff
- Supervised Learning Algorithms



Today's Contents

Learning Objectives of Today's Lecture

- Learning about the key challenges while training a machine learning model.
- Understanding supervised machine algorithms (Linear Regression, K-nearest Neighbour)
- Building practical understandings of Linear Regression and KNN

- **Important Directions**

Key Challenges/Concepts

Overfitting vs Underfitting

❑ Overfitting

- The model performs well on the training data but fails to generalize to new data.
- This usually happens when the model learns noise or irrelevant details.

Key Challenges/Concepts

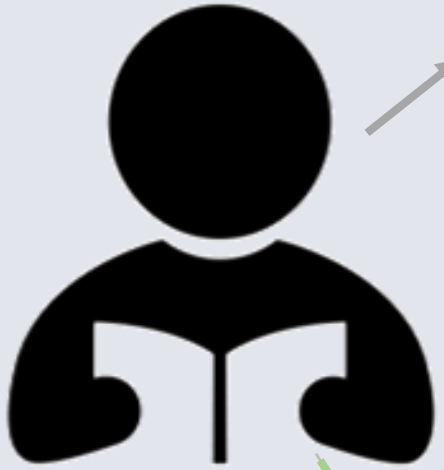
Overfitting vs Underfitting

❑ Underfitting

- The model is too simple and fails to capture the patterns in the data.
- Leading to poor performance even on the training data.

Key Challenges/Concepts

A student studying for exam



- Memorizes exact questions answers from past exam
- Without truly understanding it.

Overfitting

Final exam
Same questions: Does great
Different questions (new data): will struggle

- Barely studies might only learn very basic concept

Underfitting

Final exam
Will struggle even if exam has same questions. Not enough to anwser

Key Challenges/Concepts

In case of training a model

❑ **Overfitting**

- Model is too complex (Massive number of layers, or DT with many trees)
- Perfectly fit, capturing even smallest variations/noise but fails to generalize.

(Memorizing instead of learning)

Key Challenges/Concepts

In case of training a model

❑ **Overfitting**

- **Training accuracy** is very high, but the **test accuracy** is much lower.
- A decision tree that splits the data repeatedly until each leaf node has only one data point

Key Challenges/Concepts

In case of training a model

- **Underfitting**

- Model is too simple (Very few layers, or DT with few trees)

- Happen if you a linear model for a dataset with non-linear patterns is used

Key Challenges/Concepts

In case of training a model

❑ **Underfitting**

- **Training** and **testing** accuracy are low because the model fails to learn the relationships between the features and the target.

Key Challenges/Concepts

Algorithms

- Different algorithms are suited for different types of problems.

Classification algorithms: Discrete outputs

Regression algorithms: Continuous outputs.

- Understanding the strengths and weaknesses of each algorithm

  helps in choosing the right one for a given task.

Supervised Learning Algorithms – Regression

Linear Regression

- Statistical method used to model the relationship between a dependent variable and one or more independent variables.

Goal: Find the best-fitting line (or hyperplane in higher dimensions) that predicts the dependent variable based on the independent variable(s).

Dependent (Target): Outcome we want to predict (e.g., house prices)

Independent (Predictor): Variable used to make predictions (e.g., Square foot, No. of bedrooms, etc.)

Supervised Learning Algorithms – Regression

Linear Regression

Size (Sq. ft)	Price (currency)
1000	200,000
1500	300,000
1600	320,000
1700	340,000
1800	360,000
1900	380,000

Supervised Learning Algorithms – Regression

Linear Regression

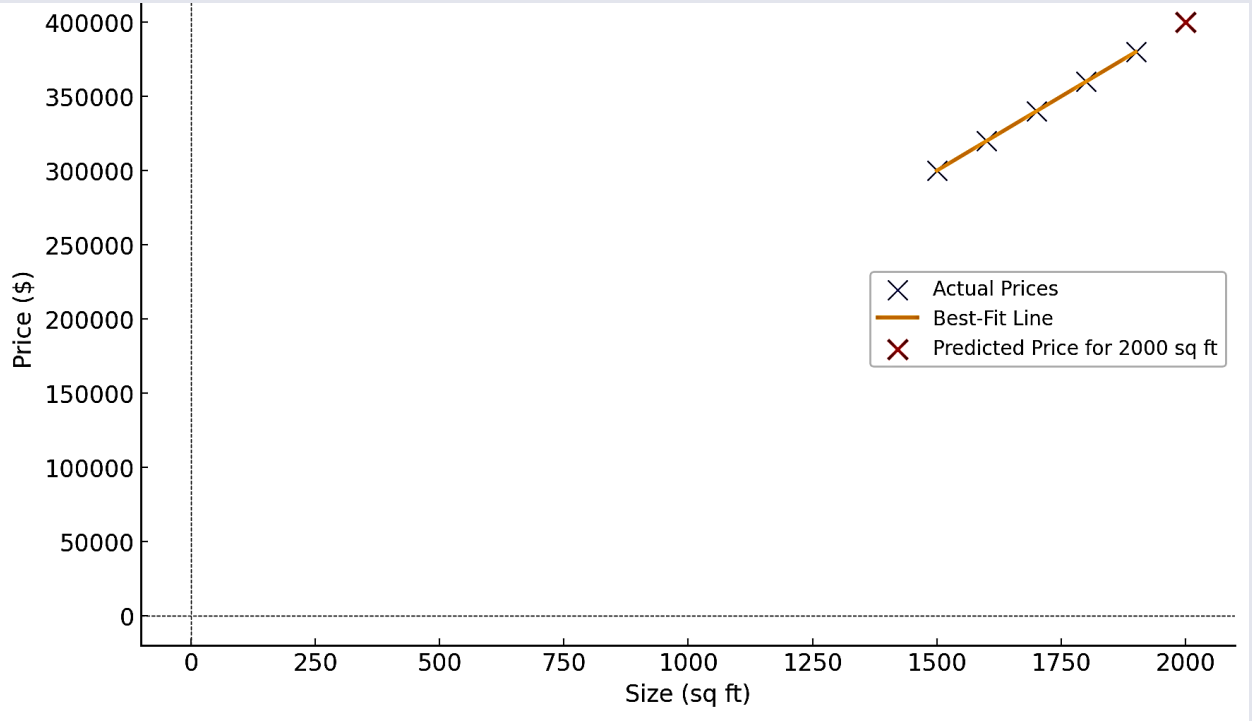
Equation of a Line: y = mx + b

:y is the dependent variable.

:m is the slope of the line (change in y for a unit change in x).

:x is the independent variable.

:b is the y-intercept (value of y when x=0).



Supervised Learning Algorithms – Classification

K-Nearest Neighbors (KNN)

- Supervised learning algorithm used for both regression and classification
- Based on the majority class of its nearest 'k' neighbors.

Applicable: Simple classification, pattern recognition, recommended system

Supervised Learning Algorithms – Classification

K-Nearest Neighbors (KNN)

- KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points.
- Based on the majority class of its nearest 'k' neighbors.
- Select the K number of points which is closet to the test data.

Supervised Learning Algorithms – Classification

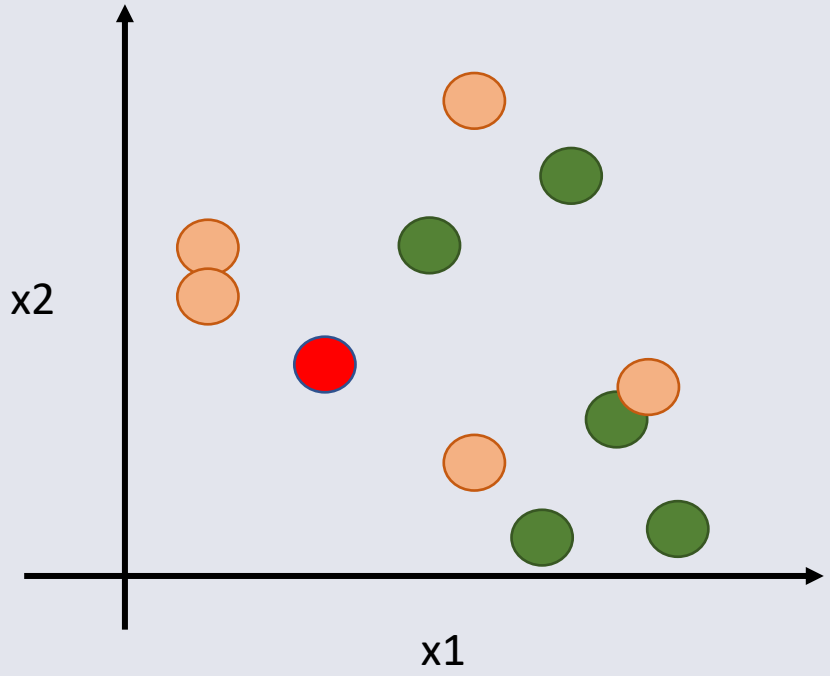
Steps in KNN

- ❑ Select the number K of the neighbors
- ❑ Calculate the Euclidean distance of K number of neighbors
- ❑ Take the K nearest neighbors as per the calculated Euclidean distance.
- ❑ Among these k neighbors, count the number of the data points in each category.
- ❑ Assign the new data points to that category for which the number of the neighbor is maximum.

Supervised Learning Algorithms – Classification

K-Nearest Neighbors (KNN)

Imagine we have a dataset of distinguishing between two points.

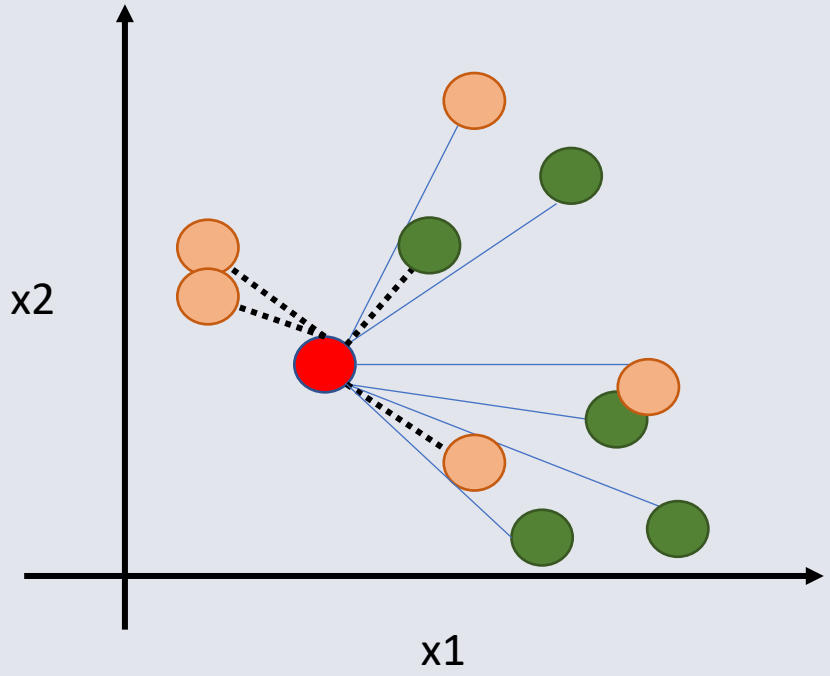


You want to classify the red-point (Test_class)

Supervised Learning Algorithms – Classification

K-Nearest Neighbors (KNN)

Imagine we have a dataset of distinguishing between two points.

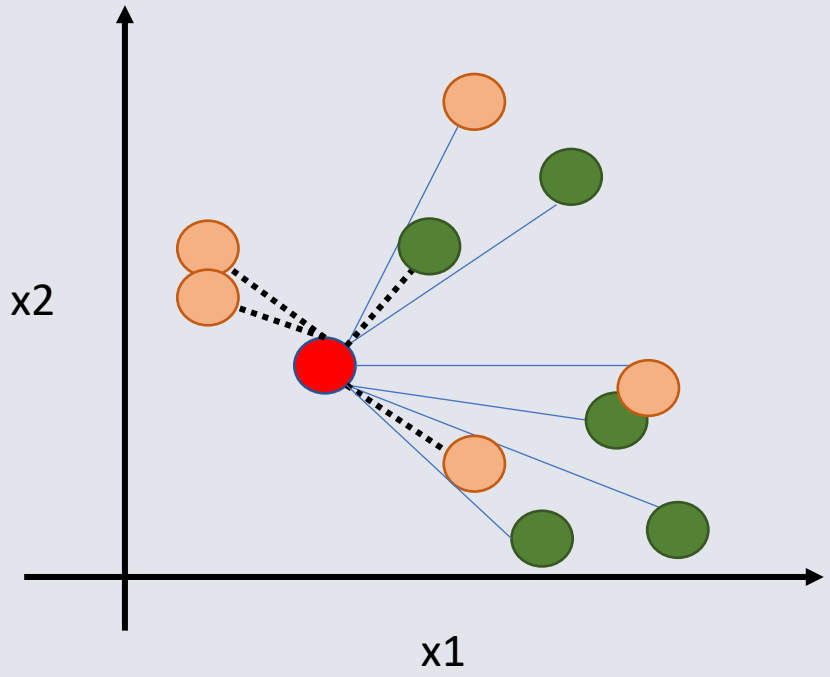


Calculate distances between the test point and all other neighbors

Supervised Learning Algorithms – Classification

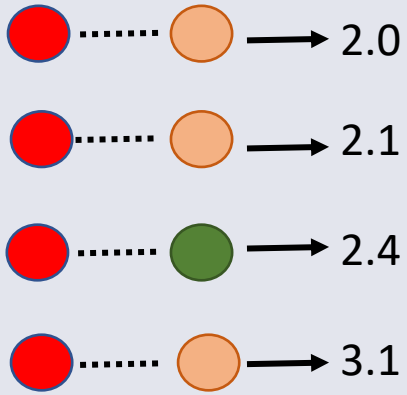
K-Nearest Neighbors (KNN)

Imagine we have a dataset of distinguishing between two points.



Finding Neighbors

Point Distance



Ranking the points by increasing distance

1st NN

2nd NN

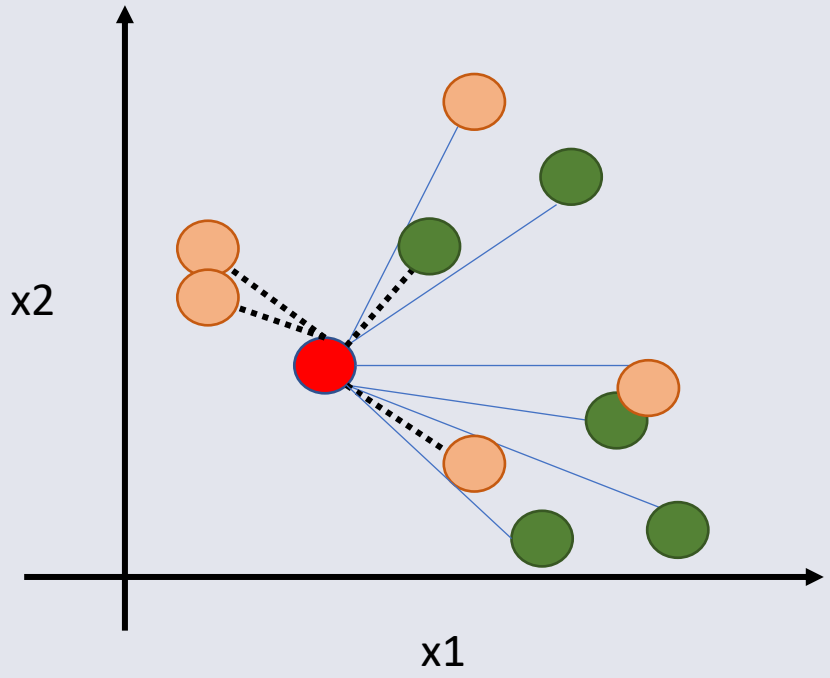
3rd NN

4th NN

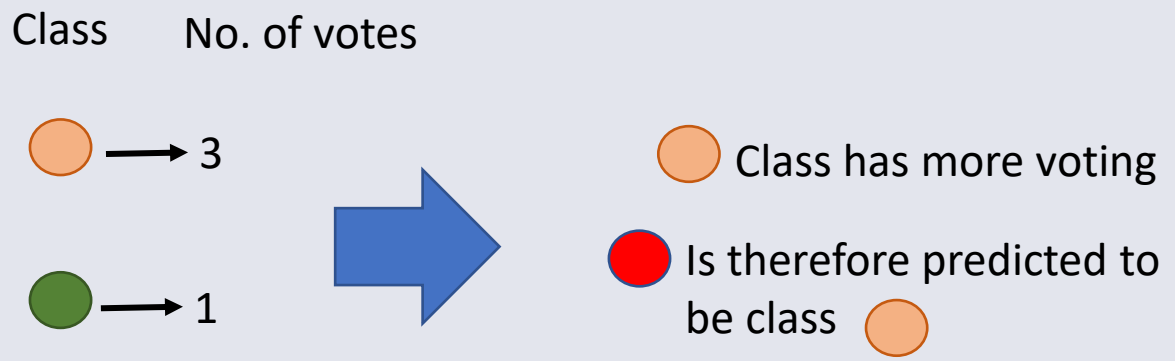
Supervised Learning Algorithms – Classification

K-Nearest Neighbors (KNN)

Imagine we have a dataset of distinguishing between two points.



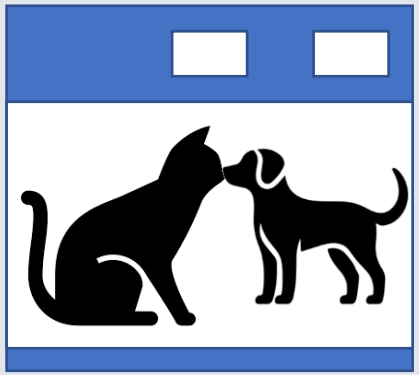
Voting on Label



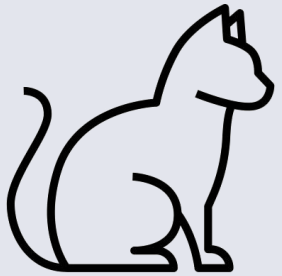
Supervised Learning Algorithms – Classification

Another Example

Input image



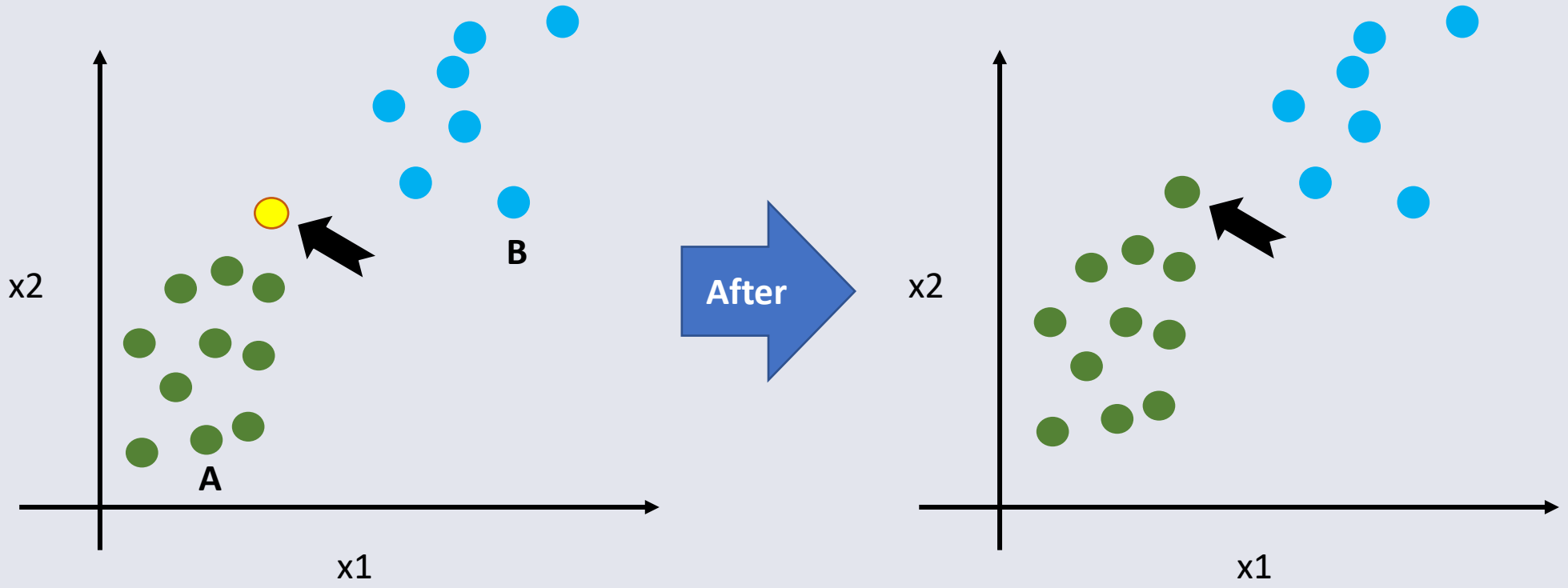
KNN classifier



Predicted output

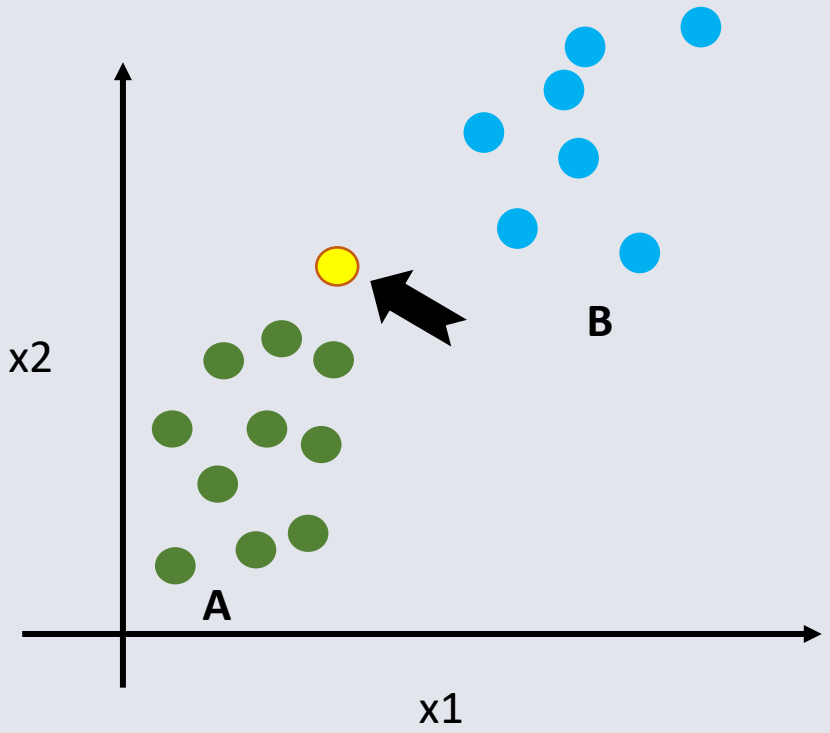
Supervised Learning Algorithms – Classification

Another Example

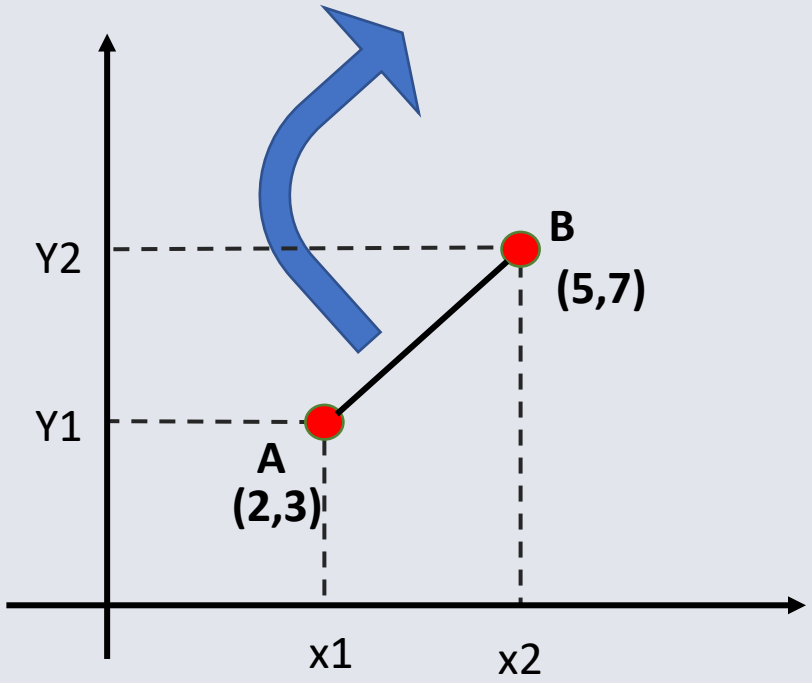


Supervised Learning Algorithms – Classification

Another Example

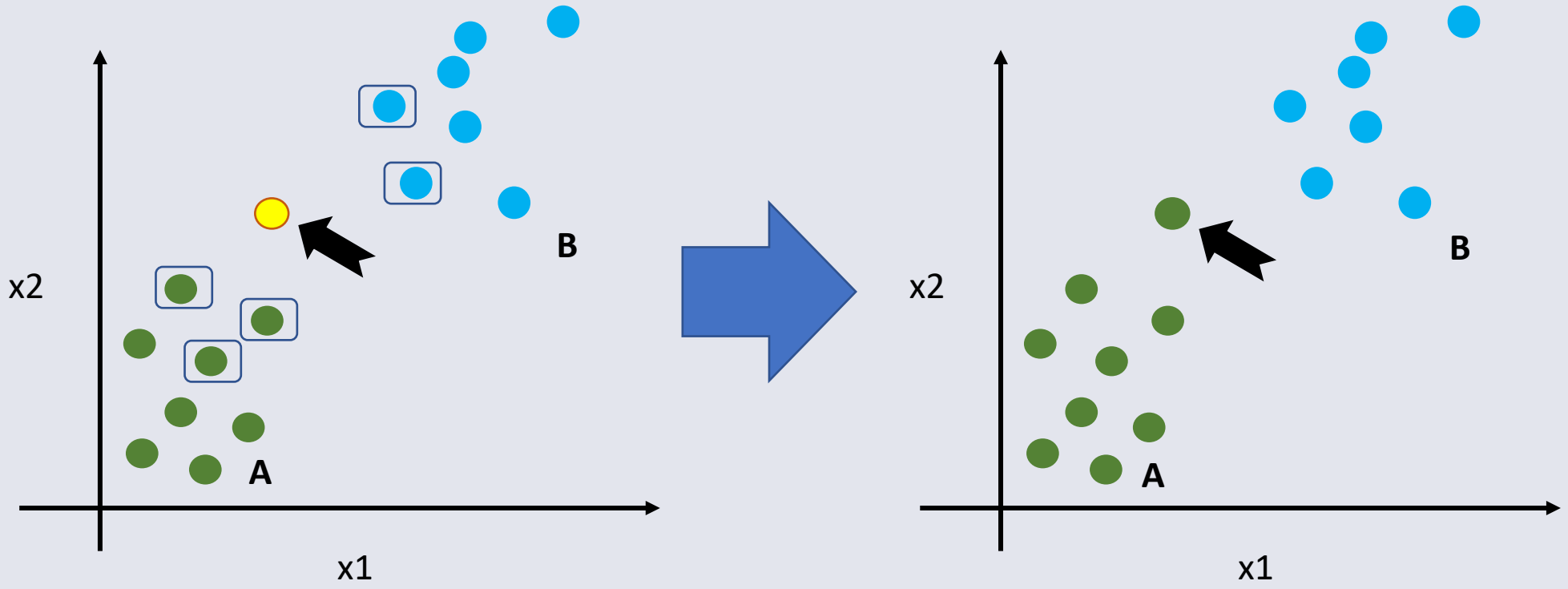


$$D = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$



Supervised Learning Algorithms – Classification

Another Example



Practical example whether a fruit is an Apple or an Orange

Classify a new fruit as either an Apple or an Orange using K-Nearest Neighbors (KNN).

We'll use two simple features for each fruit:

1. Weight (grams)

2. Texture score = (1 = Smooth, 2 = Rough)

Fruit	Weight (g)	Texture	Category
F1	150	1	Apple
F2	170	1	Apple
F3	140	1	Apple
F4	130	2	Orange
F5	120	2	Orange
F6	110	2	Orange

Practical example whether a fruit is an Apple or an Orange

New Data Point

A **new fruit (F7)** has:

Weight = **145g**

Texture = **1.5** (somewhat between smooth and rough)

We want to classify: Apple or Orange?

Practical example whether a fruit is an Apple or an Orange

Step 1: Choose K

Let's choose **K = 3** (we'll look at the 3 closest fruits).

Step 2: Calculate Euclidean Distance

Formula:

$$D = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category
F1	(150, 1)	Apple
F2	(170, 1)	Apple
F3	(140, 1)	Apple
F4	(130, 2)	Orange
F5	(120, 2)	Orange
F6	(110, 2)	Orange

(145, 1.5)

The new datapoint

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	
F2	(170, 1)	Apple	
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$(150-145)^2$
F2	(170, 1)	Apple	
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$(150-145)^2 + (1-1.5)^2$
F2	(170, 1)	Apple	
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)}$
F2	(170, 1)	Apple	
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F2	(170, 1)	Apple	
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F2	(170, 1)	Apple	$\sqrt{((170-145)^2 + (1-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F3	(140, 1)	Apple	
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F2	(170, 1)	Apple	$\sqrt{((170-145)^2 + (1-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F3	(140, 1)	Apple	$\sqrt{((140-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F4	(130, 2)	Orange	
F5	(120, 2)	Orange	
F6	(110, 2)	Orange	

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F2	(170, 1)	Apple	$\sqrt{((170-145)^2 + (1-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F3	(140, 1)	Apple	$\sqrt{((140-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F4	(130, 2)	Orange	$\sqrt{((130-145)^2 + (2-1.5)^2)} = \sqrt{(225 + 0.25)} =$ **15.01**
F5	(120, 2)	Orange	$\sqrt{((120-145)^2 + (2-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F6	(110, 2)	Orange	$\sqrt{((110-145)^2 + (2-1.5)^2)} = \sqrt{(1225 + 0.25)} =$ **35.01**

Practical example whether a fruit is an Apple or an Orange

Fruit	(Weight, Texture)	Category	Distance to (145, 1.5)
F1	(150, 1)	Apple	$\sqrt{((150-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F2	(170, 1)	Apple	$\sqrt{((170-145)^2 + (1-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F3	(140, 1)	Apple	$\sqrt{((140-145)^2 + (1-1.5)^2)} = \sqrt{(25 + 0.25)} =$ **5.02**
F4	(130, 2)	Orange	$\sqrt{((130-145)^2 + (2-1.5)^2)} = \sqrt{(225 + 0.25)} =$ **15.01**
F5	(120, 2)	Orange	$\sqrt{((120-145)^2 + (2-1.5)^2)} = \sqrt{(625 + 0.25)} =$ **25.00**
F6	(110, 2)	Orange	$\sqrt{((110-145)^2 + (2-1.5)^2)} = \sqrt{(1225 + 0.25)} =$ **35.01**

Apple = 2

Orange = 1

The new fruit (F7) is classified as an Apple

Exercise (10 Minutes~)

Predict if a Vehicle is a **Car** or a **Truck**

Vehicle	Weight	Engine	Type
V1	1	1	Car
V2	2	1	Car
V3	4	3	Truck
V4	5	4	Truck

Classify a new vehicle as **Car** or **Truck** using KNN.

New vehicle (V5) has:

Weight = 3 tons, Engine = 2 liters

Evaluating Machine Learning Algorithms

Cross Validation:

Estimate the model's performance by splitting the dataset into training and testing parts multiple times.

Metrics for Classification:

Accuracy: The proportion of correct predictions.

Precision: The number of true positives divided by the sum of true positives and false positives.

Recall: The number of true positives divided by the sum of true positives and false negatives.F1-Score: The harmonic mean of precision and recall, useful when class distribution is imbalanced.

Evaluating Machine Learning Algorithms

Metrics for Regression

Mean Squared Error (MSE): Measures the average squared difference between predicted and actual values.

R-Squared: Describes how well the model explains the variance in the target variable.

Next Week

- **Supervised Learning Algorithms – Classification**
  - Support Vector Machines (SVM)
  - Decision Trees
  - Random Forest
- **Unsupervised Learning Algorithms – Clustering**
  - K-Means Clustering
  - Principal Component Analysis (PCA)

Any Question?