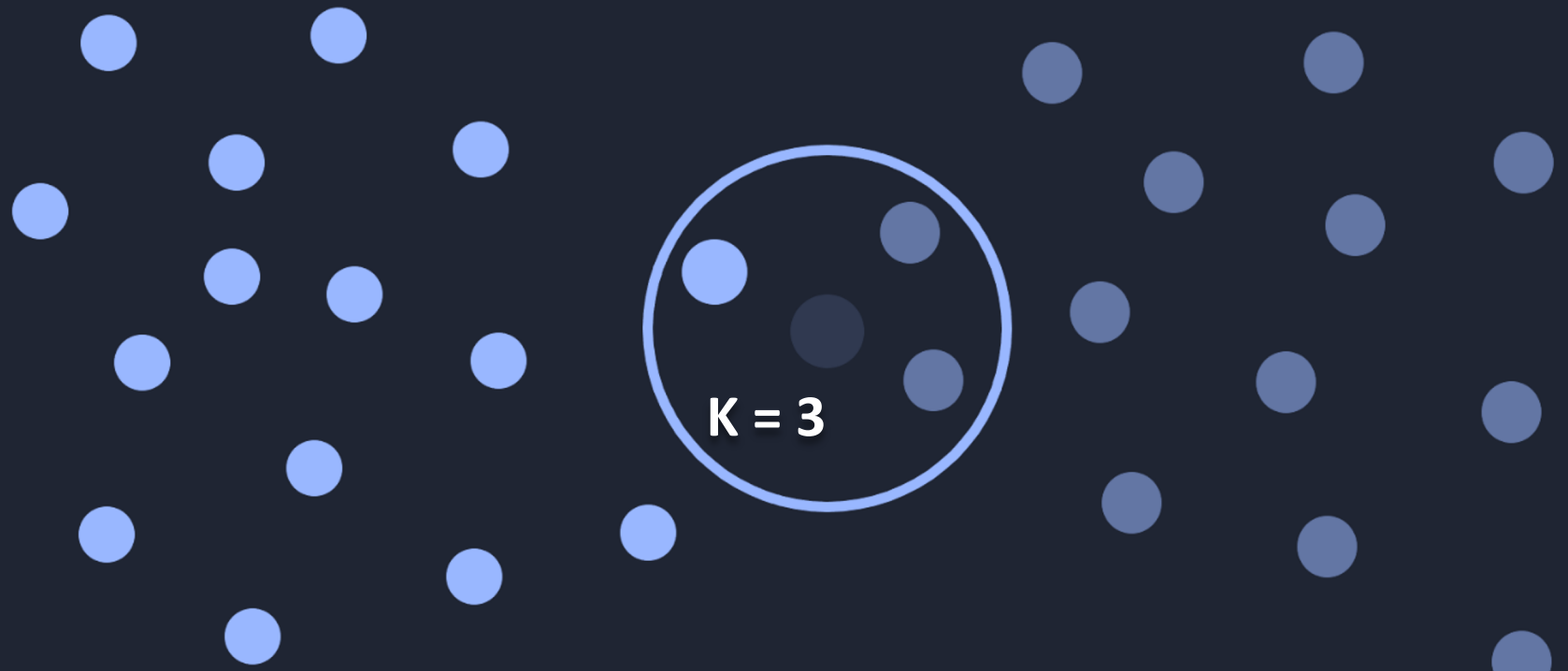


K – Nearest Neighbor

INSAID

Course Overview

You are here...

Term	CDF	GCD	GCDAI	PGPDSAI
Term 1	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python
Term 2	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques
Term 3	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling
		Minor Project	Minor Project	Minor Project
Term 4		Machine Learning Foundation	Machine Learning Foundation	Machine Learning Foundation
Term 5		Machine Learning Intermediate	Machine Learning Intermediate	Machine Learning Intermediate
Term 6		Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)
		Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)
		Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)
		Capstone Project	Capstone Project	Capstone Project
Term 7		Bonus: Industrial ML (ML – 4 & 5)	Basics of AI, TensorFlow, and Keras	Basics of AI, TensorFlow, and Keras
Term 8			Deep Learning Foundation	Deep Learning Foundation
Term 9			NPL – I/CV – I	CV – I
Term 10			NLP – II/CV – II	NLP – I
			Capstone Project	Capstone Project
Term 11				CV – II
Term 12				NLP – II
				NLP – III + CV – III
				AutoVision & AutoNLP
				Building AI product

INSAID

Term Context

- **K – Nearest Neighbor**  You are here…
- K-means Clustering
- Ensemble Learning
- Optimization

INSAID

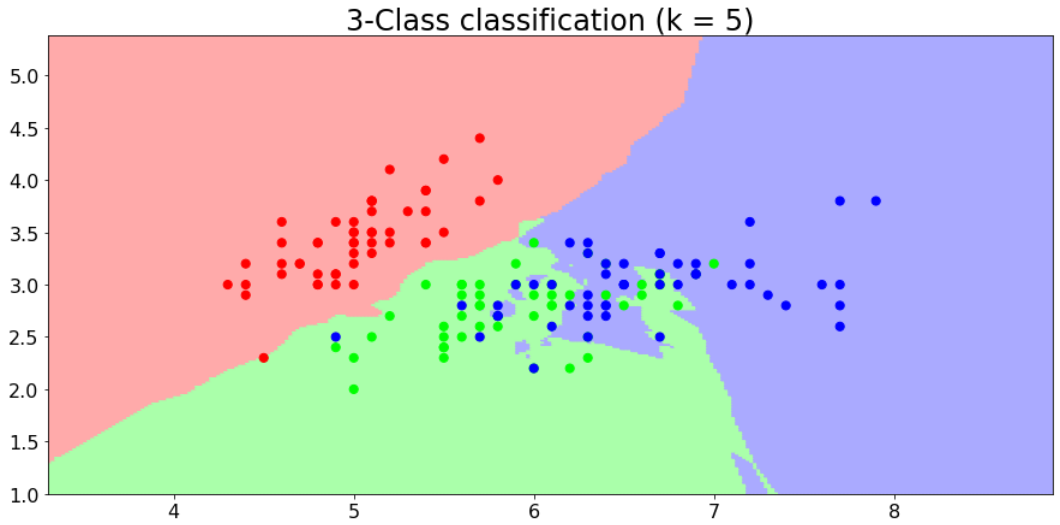
Agenda

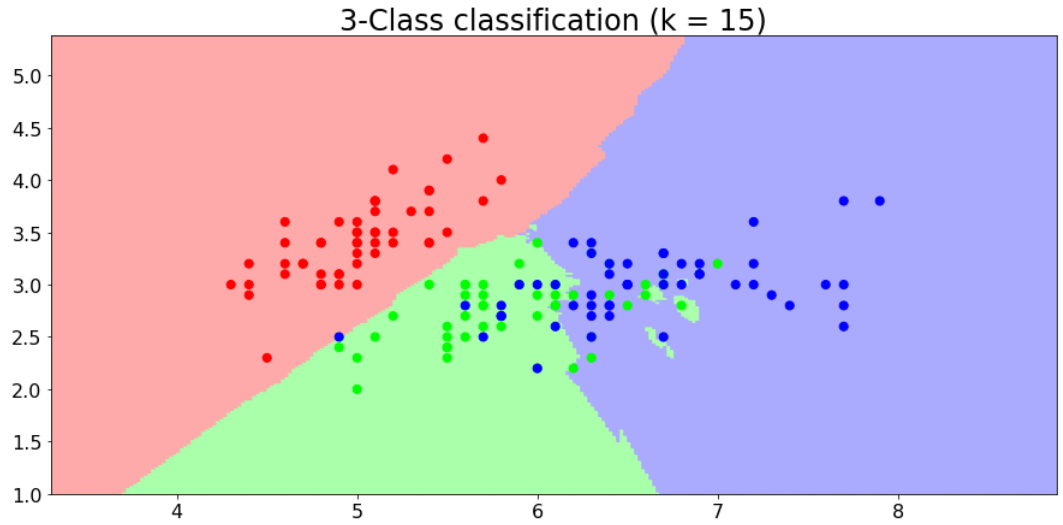
1. **What is K Nearest Neighbor algorithm?**
2. When to use KNN?
3. Similarity Measures
4. Estimation of Similarity
5. Types of KNN
6. KNN for Regression
7. KNN for Classification
8. The K-Factor
9. Other Distance Measures
10. Advantages
11. Limitations
12. Applications

INSAID

The K-Factor

- As the value of **k increases**, the **ability** of KNN to **generalize increases**.
- But we see that the model **fails** to find a good decision boundary when **k** is equivalent to **number of samples**.
- Conversely, **k=1 overfits** the model.

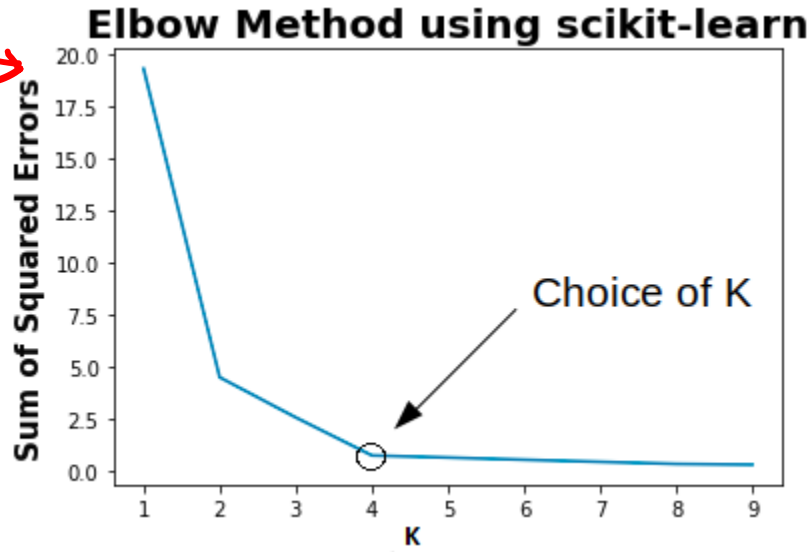




INSAID

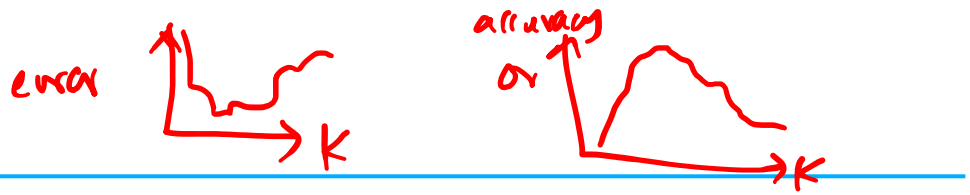
The K-Factor – Optimal value of K

- **Sqrt(n)**, where n is the total number of data points.
- **Odd value of k** is selected to **avoid any confusion** between even number of classes of data.
- We can also apply **elbow method** to find out k.

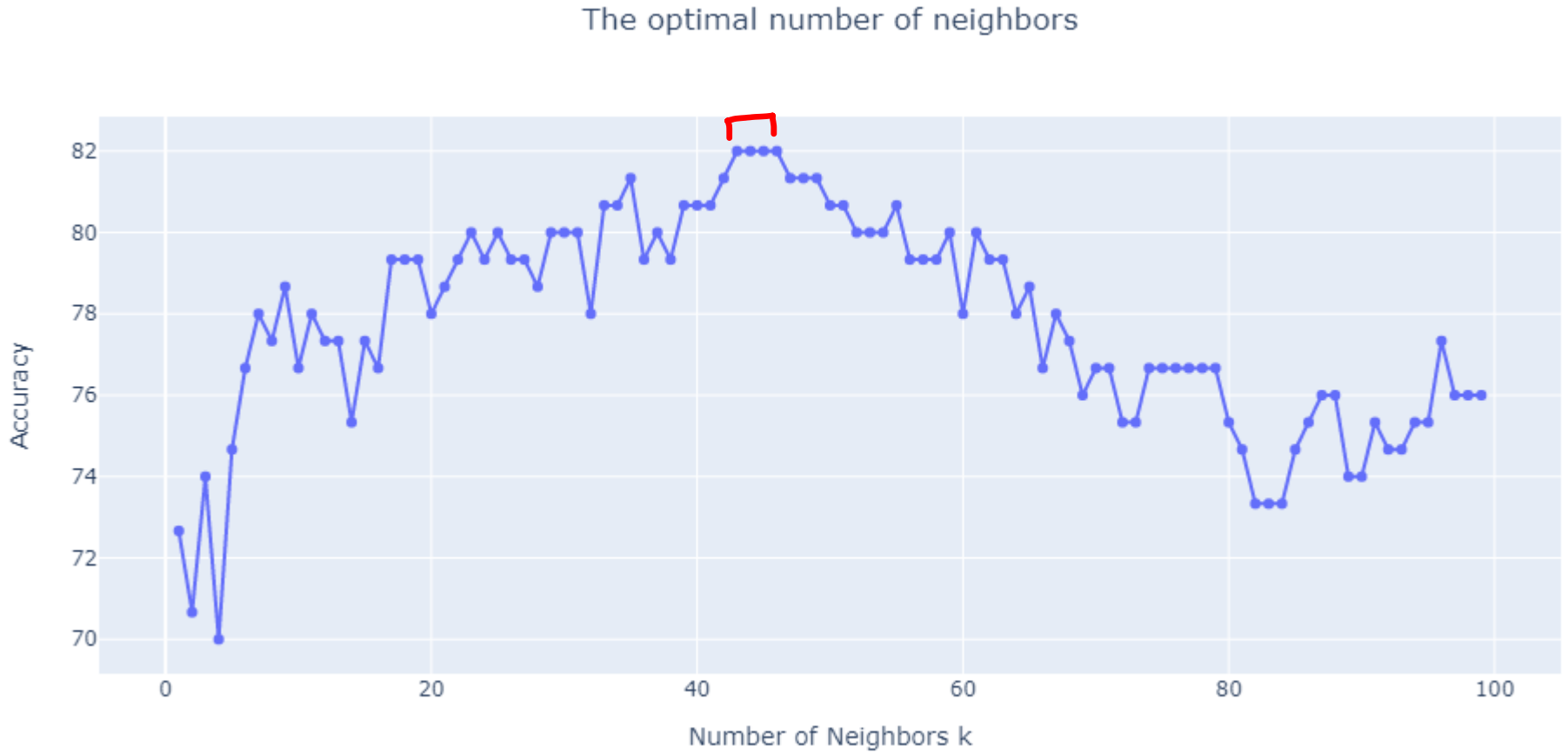


INSAID

The K-Factor – Optimal value of K



- At **k=43**, we can see that the model fits the **best decision boundary** before the accuracy starts to plummet.



INSAID

Agenda

1. What is K Nearest Neighbor algorithm?
2. When to use KNN?
3. Similarity Measures
4. Estimation of Similarity
5. Types of KNN
6. KNN for Regression
7. KNN for Classification
8. The K-Factor
9. **Other Distance Measures**
10. Advantages
11. Limitations
12. Applications

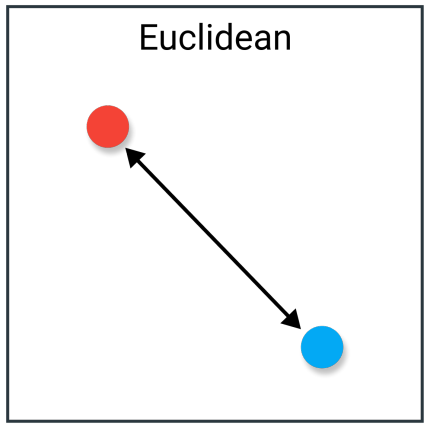
INSAID

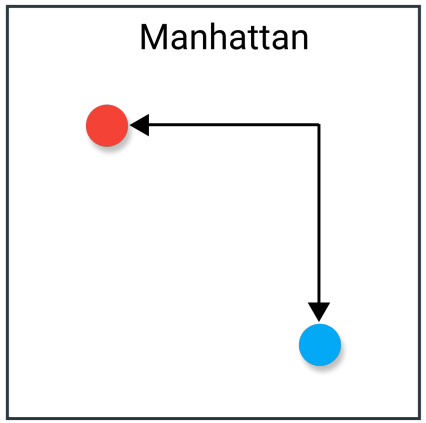
Other Distance Measures

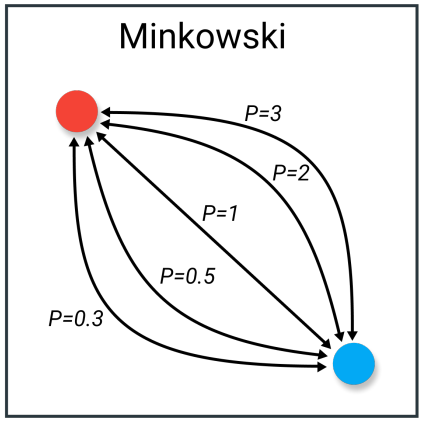
- There are other distance metrics like Euclidean distance that can also be used to calculate **similarity between features**.
- For Example: **Manhattan** and **Minkowski** distances.

$A = (x_1, y_1)$

$B = (x_2, y_2)$







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

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

$$\sqrt[p]{(x_2-x_1)^p + (y_2-y_1)^p}$$

INSAID

Manhattan Distance

- It estimate distance between two points as the **sum** of the **absolute differences** of their **Cartesian coordinates**.
- It is also known as **taxi-cab** distance.

$$d(a, b) = |a_1 - b_1| + |a_2 - b_2| + \cdots + |a_n - b_n|$$

Here,

- $d(a, b)$: Manhattan distance between two objects **a** and **b**
- $a_1, a_2, \ldots, a_n$: Features of object a.
- $b_1, b_2, \ldots, b_n$: Features of object b.

Manhattan Distance (Example)

- Let's say we have two points as shown below in the table and we want to estimate Manhattan distance.

	#	Feature 1	Feature 2	Feature 3	Feature 4
a	1	50	28	62	16
b	2	38	13	44	13

$$d(a, b) = |a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + |a_4 - b_4|$$

$$d(a, b) = |50 - 38| + |28 - 13| + |62 - 44| + |16 - 13|$$

$$d(a, b) = 12 + 15 + 18 + 3$$

$$d(a, b) = 48$$

INSAID

Manhattan Distance (Use Case)

- **Taxi cab drivers** and **food delivery systems** use Manhattan distance to estimate **distance** between **two blocks** in a city.

INSAID

Minkowski Distance

- It is the **generalized form** of Euclidean and Manhattan Distance.
- It is represented by the following formula:

$$d(a,b) = \sqrt[p]{(a_1 - b_1)^p + (a_2 - b_2)^p + \cdots + (a_n - b_n)^p}$$

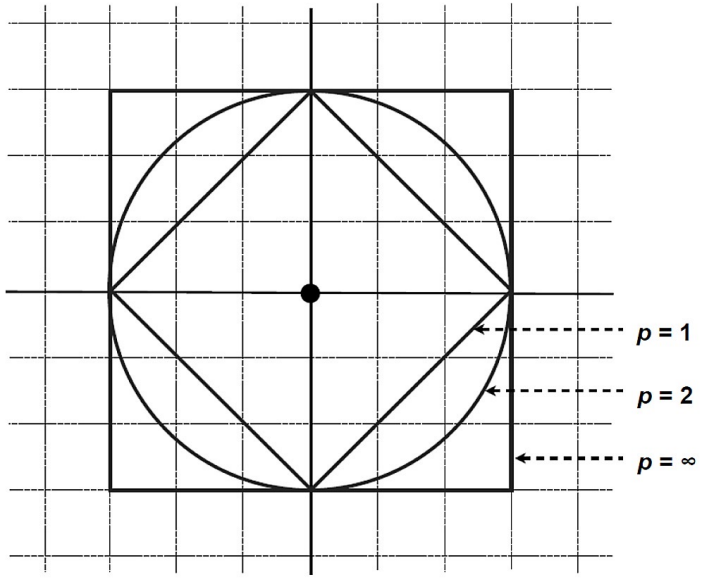
Here,

- $d(a,b)$: Manhattan distance between two objects **a** and **b**
- $a_1, a_2, \ldots, a_n$: Features of object a.
- $b_1, b_2, \ldots, b_n$: Features of object b.
- p: Order of the norm.

INSAID

Minkowski Distance

- When p=1, Minkowski distance is same as **Manhattan distance**.
- When p=2, it is same as **Euclidean distance**.
- In KNN, we can **iterate over p** through different values to find the **distance** that **works best** with our data.



INSAID

Minkowski Distance (Example)

- Let's say we have two points as shown below in the table and we want to estimate Minkowski distance for **p=1**.

#	Feature 1	Feature 2	Feature 3	Feature 4
1	50	28	62	16
2	38	13	44	13

Minkowski distance for $p = 1$

$$d(a,b) = \sqrt[p]{(a_1 - b_1)^p + (a_2 - b_2)^p + (a_3 - b_3)^p + (a_4 - b_4)^p}$$

$$d(a,b) = \sqrt[1]{(a_1 - b_1)^1 + (a_2 - b_2)^1 + (a_3 - b_3)^1 + (a_4 - b_4)^1}$$

$$d(a,b) = \sqrt[1]{(50 - 38)^1 + (28 - 13)^1 + (62 - 44)^1 + (16 - 13)^1}$$

$$d(a,b) = 12 + 15 + 18 + 3$$

$$d(a,b) = 48$$

INSAID

Minkowski Distance (Example)

Euclidean

- Let's say we have two points as shown below in the table and we want to estimate Minkowski distance for **p=2**.

#	Feature 1	Feature 2	Feature 3	Feature 4
1	50	28	62	16
2	38	13	44	13

Minkowski distance for $p = 2$

$$d(a,b) = \sqrt[p]{(a_1 - b_1)^p + (a_2 - b_2)^p + (a_3 - b_3)^p + (a_4 - b_4)^p}$$

$$d(a,b) = \sqrt[2]{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2}$$

$$d(a,b) = \sqrt[2]{(50 - 38)^2 + (28 - 13)^2 + (62 - 44)^2 + (16 - 13)^2}$$

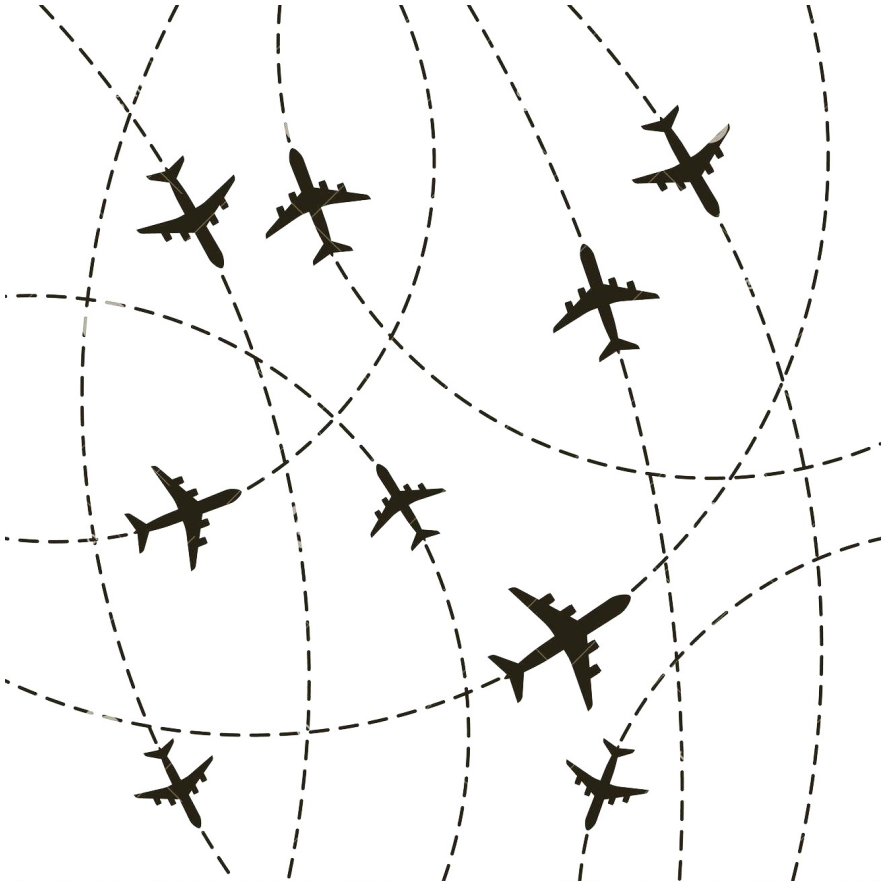
$$d(a,b) = \sqrt[2]{702}$$

$$d(a,b) = 26.49$$

INSAID

Minkowski Distance (Use Case)

- **Air Traffic Controls** use Minkowski distance to estimate **distance** between **planes** and to avoid overlaps in their routes.



INSAID

Agenda

1. What is K Nearest Neighbor algorithm?
2. When to use KNN?
3. Similarity Measures
4. Estimation of Similarity
5. Types of KNN
6. KNN for Regression
7. KNN for Classification
8. The K-Factor
9. Other Distance Measures
10. **Advantages**
11. Limitations
12. Applications

INSAID

Advantages

- KNN algorithm is widely used for **regression** and **classification.**
- It is **uncomplicated** and **easy** to implement.
- Only two metrics - **value of K** and the **distance metric**.
- Works with **any number of classes**.
- It is fairly **easy** to **add new data** to the algorithm.

INSAID

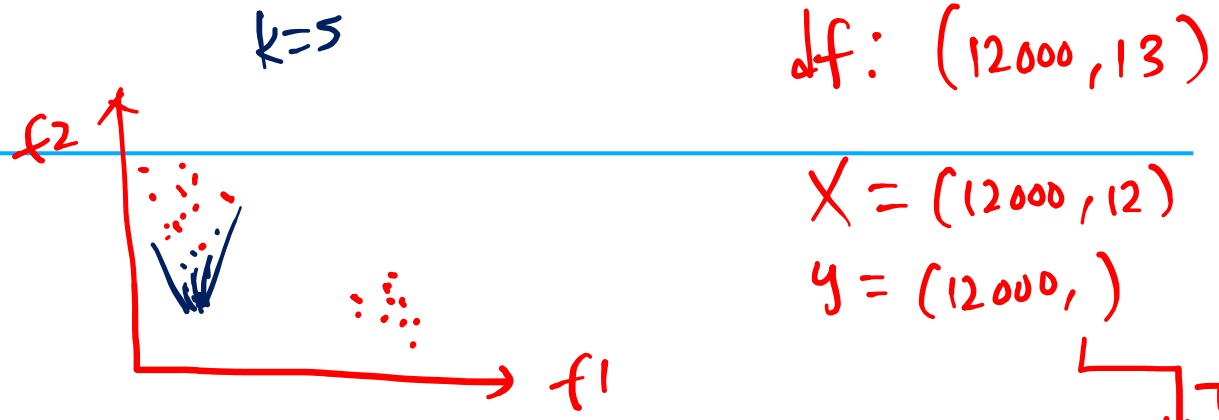
Agenda

1. What is K Nearest Neighbor algorithm?
2. When to use KNN?
3. Similarity Measures
4. Estimation of Similarity
5. Types of KNN
6. KNN for Regression
7. KNN for Classification
8. The K-Factor
9. Other Distance Measures
10. Advantages
11. **Limitations**
12. Applications

INSAID

Limitations

KNN: Lazy learning



df: (12000, 13)

X = (12000, 12)

y = (12000,)

TTS 80:20

- The **cost** of **predicting** the k nearest neighbors is **very high**.
- Does not work well with **high dimensions**.
- It is **hard** to work with **categorical** features.
- Sensitive to **noisy** data, **missing values** and **outliers**.
- As it stores an entire dataset into the memory, it can lead to **memory errors**.

X_train : (9600, 12)

y_train : (9600,)

X_test : (2400, 12)

y_test : (2400,)

INSAID

Agenda

1. What is K Nearest Neighbor algorithm?
2. When to use KNN?
3. Similarity Measures
4. Estimation of Similarity
5. Types of KNN
6. KNN for Regression
7. KNN for Classification
8. The K-Factor
9. Other Distance Measures
10. Advantages
11. Limitations
12. **Applications**

INSAID

Applications – Computer Vision

- Suppose we have an input image. **Reverse image search** systems can be used to find sets of **similar images** across the web or a local system.

INSAID

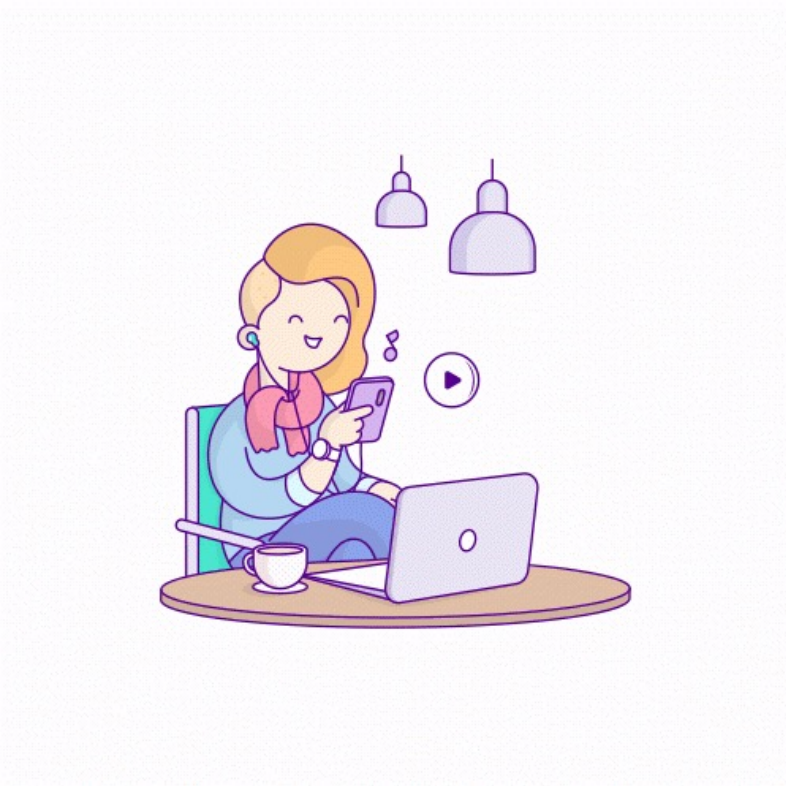
Applications – Product Recommendation

- Amazon can target you with **similar products**, or products that other customer that have the same **buying habits** as you.

INSAID

Applications – Content Recommendation

- Probably the most interesting example is **Spotify** and their great recommendation engine.
- Spotify will **recommend** you songs/podcasts based on what you have been **recently hearing**.



INSAID

Applications – Concept Search

- Softwares like **Relatively** use KNN to search for **semantically similar documents**.

INSAID

INSAID