Machine Learning
Nearest Neighbors Model

DR. BHARGAVI R
PROFESSOR
SCOPE
VIT CHENNAI

Classification - Types

Binary Classification: Involves categorizing data into one of the two classes

- Online transactions – Fraudulent / Not Fraudulent
- Email – Spam/ Not spam ?
- Tumor classification – Malignant/Benign

Multi-class Classification: Involves more than two classes. i.e A data instance can belong to one of may possible classes

- Optical Character Recognition
- Face classification

Classification Types (cont…)

Multi-Label Classification

A variant of the classification problem where multiple nonexclusive labels may be assigned to each instance.

- Photo tagging feature where each photo can have multiple tags such as 'Beach', 'Friends', 'Summer', and 'Vacation'. Each tag is a different label, and multiple tags can be correct for a single photo.

Nearest Neighbor Classification - Intuition

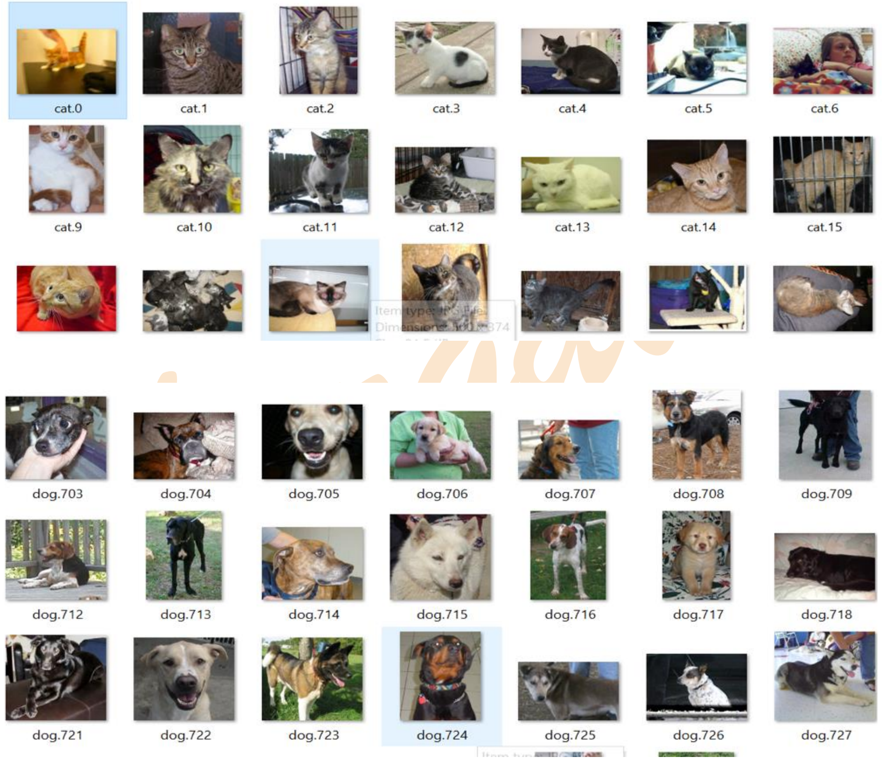
Can you recognize me?

?

?

How did we answer?

Similar inputs => Similar outputs



Cats

Dogs

Nearest Neighbors - Introduction

- Popular, intuitive and simple to understand.
- Supervised learning algorithm.
- Non Parametric Learning model – No functional form assumption.
- Lazy Learner – All the computations (similarity or distance computations) are postponed till the prediction time.
- Memorize the training instances (Memory based learning) and use at the time of prediction.
- Used for classification and regression.

Use case: House Number Identification

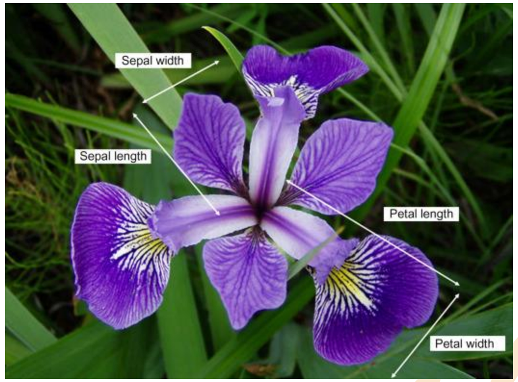
- How do you identify the house number from the image captured?

What are the *Challenges* we see here

- How to represent the data?
- Which instances are nearest neighbors?
- How to find the similarity?
- How many nearest neighbors are to be considered for decision making?

How to Represent the Data? (cont…)

Structured Data



Feature	Value (cms)
sepal length	6.3
sepal width	3.3
petal length	6
petal width	2.5

Id	Sepal Length	Sepal Width (cm)	Petal Length	Petal Width (cm)	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa

How to Represent the Data? (cont…)

Unstructured Data

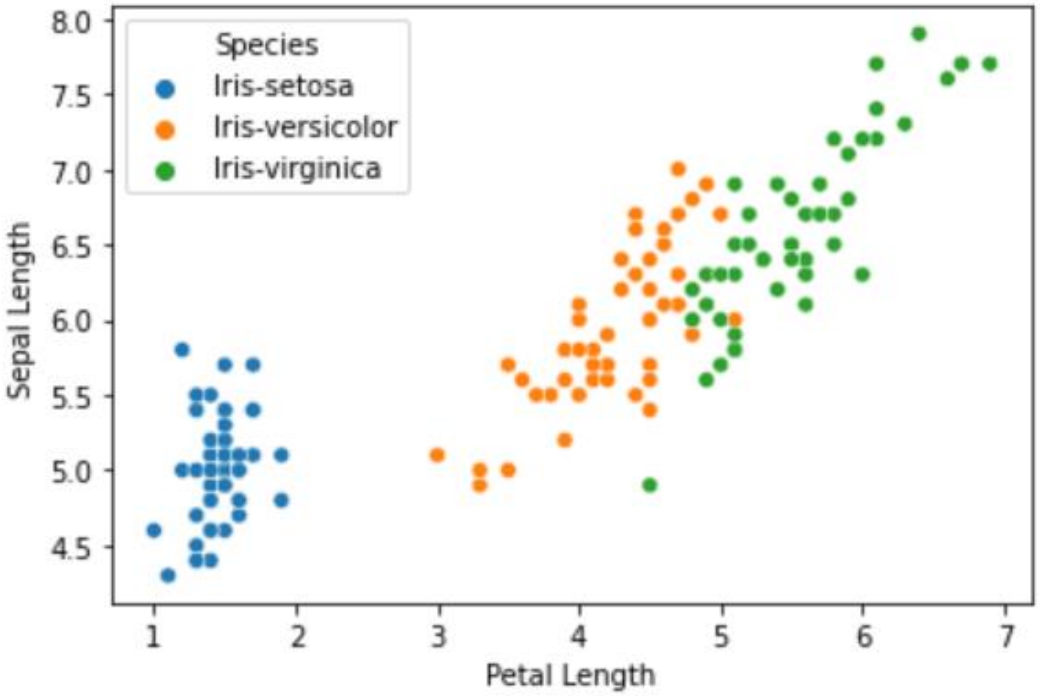
0	0	7	8	13	16	15	1
0	0	7	7	4	11	12	0
0	0	0	0	8	13	1	0
0	4	8	8	15	15	6	0
0	2	11	15	15	4	0	0
0	0	0	16	5	0	0	0
0	0	9	15	1	0	0	0
0	0	13	5	0	0	0	0

0	0	7	8	13	16	15	1	---------	0	0	13	5	0	0	0	0

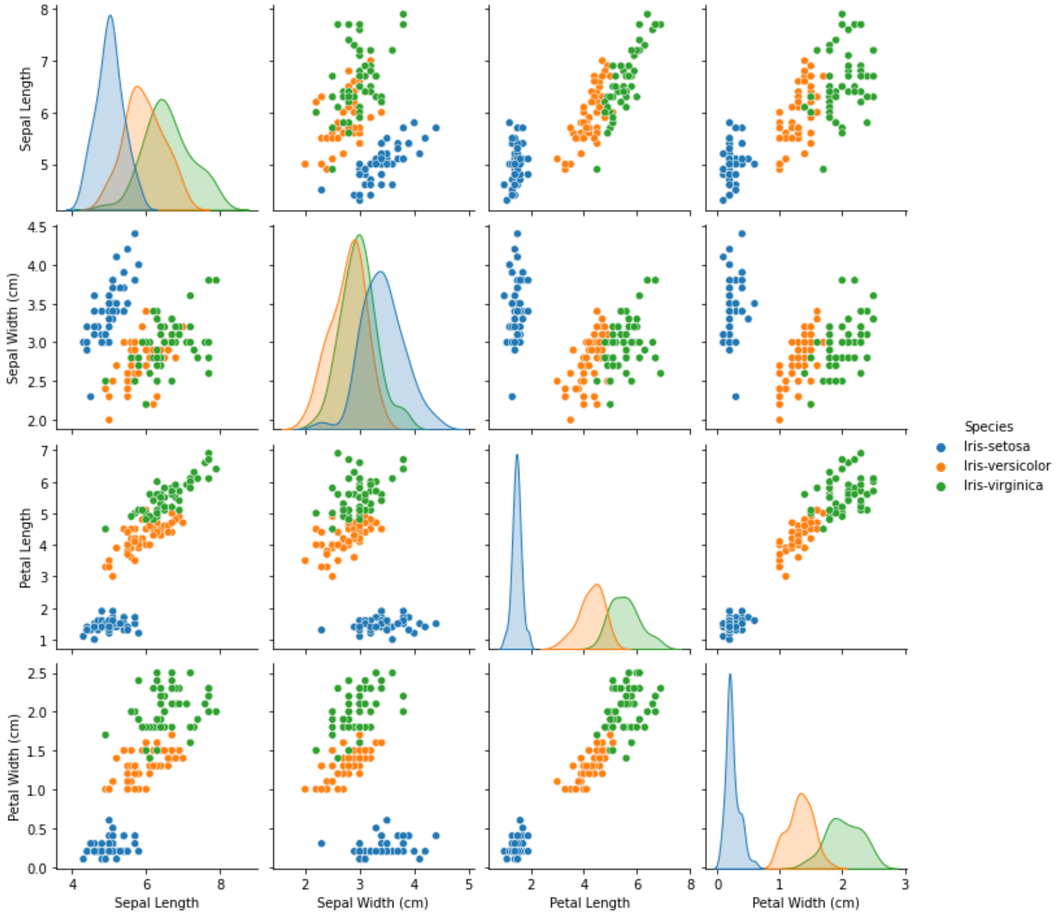
What are Nearest Neighbors

Which instances are nearest neighbors?

Different classes seem to be well separated from the other



Neighbors (cont…)



One NN Classification

Input data

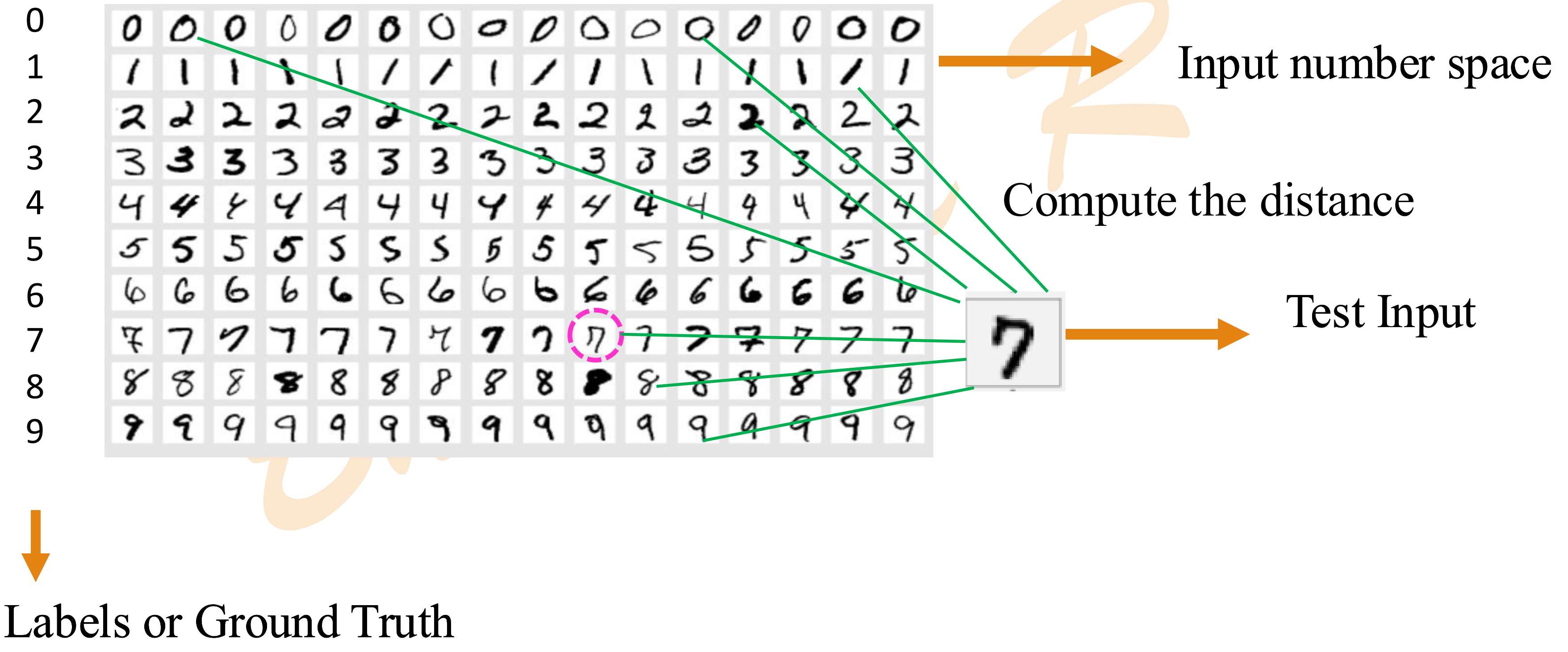
Input number space

Labels or Ground Truth

Test Input

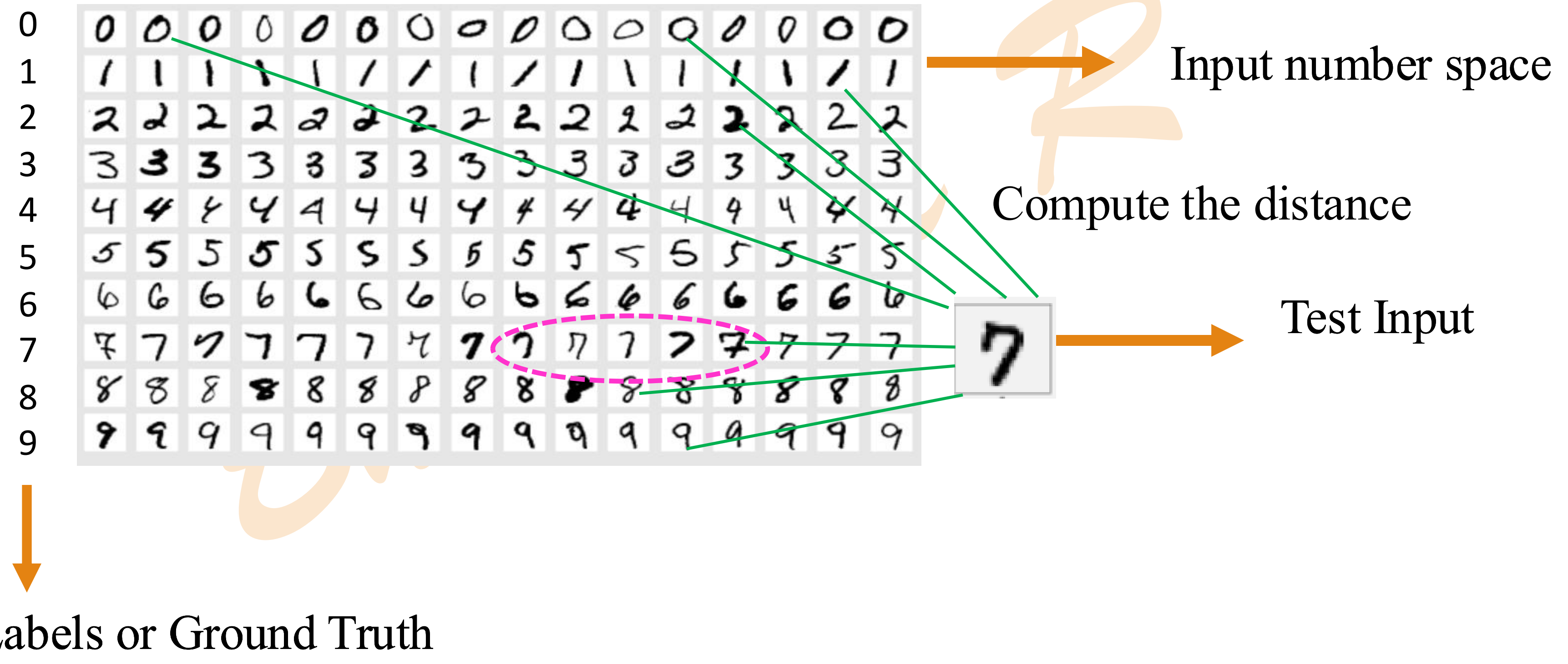
One NN Classification (cont…)

Choose the minimum distance or ***most similar input instance*** and label the test input with the corresponding input instance's label



K NN Classification

Choose the set of minimum distances or ***most similar input instances*** and label the test input with the label of the majority class from the set



How to find the most similar/nearest instances?

Distance Metrics

- Similarity – Domain specific
- Euclidian distance: Euclidean distance between two vectors x_i and y_i is

$$D(x_i , y_i) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

  - In 1 Dimension Distance between x_i and y_i is given by $D(x, y) = |x - y|$
- The more the distance is, less similar the data points are.

Distance Metrics (cont …)

- Manhattan distance: Manhattan distance between two vectors x_i and y_i is

$$D(x_i\,, y_i) = \sum_{i=1}^{n} |x_i - y_i|$$

- The more the distance is, less similar the data points are.

Distance Metrics (cont …)

- Scaled Euclidean
- Mahalanobis
- Correlation
- Cosine Similarity
- Jaccard etc

One NN Example

Consider the dataset of humans indicating the age of the person (input) and the stage (label/output)to which the person belongs to.

Sl.No	Age	Category
1	1	Child
2	12	Child
3	13	Adolescence
4	18	Adolescence
5	19	Adult
6	59	Adult
7	60	Senior Adult
8	100	Senior Adult

What is the category of a person who is 35 years old?

One NN Example

Sl.No	Age	Category	Distance from test data
1	1	Child	34
2	12	Child	23
3	13	Adolescence	22
4	18	Adolescence	17
5	19	Adult	16
6	59	Adult	24
7	60	Senior Adult	25
8	100	Senior Adult	65

Nearest neighbor to x_test (35). So, prediction = Adult

Therefore, the person with Age = 35 belongs to Adult category.

K-NN Example

Sl.No	Age	Category	Distance from test data
1	1	Child	34
2	12	Child	23
3	13	Adolescence	22
4	18	Adolescence	17
5	19	Adult	16
6	59	Adult	24
7	60	Senior Adult	25
8	100	Senior Adult	65

3 Nearest neighbor to x_test (35). So, prediction = Adolescence

With k = 3, the person with Age = 35 belongs to Adolescence category.

Another Example

Consider the dataset given below, indicating the age, salary of the person (input) and the loan approval status (label/output).

Sl.No	**Age**	**Salary**	**Loan Approval Status**
1	25	40,000	N
2	35	60,000	N
3	45	80,000	N
4	23	95,000	Y
5	40	62,000	Y
6	60	1,00,000	Y

What's the Loan approval status of a person who is 48 years old & salary of 90000?

Example (cont…)

Normalize the data with min-max normalization

$(x_i^{test} - \min(x_i))/(\max(x_i) - \min(x_i))$

Sl.No	Age	Salary	Status
1	0.054	0	N
2	0.324	0.333	N
3	0.595	0.667	N
4	0	0.917	Y
5	0.459	0.367	Y
6	1	1	Y

Normalizing the test input 48 = 0.675, 90000 = 0.833

Example (cont…)

Compute the distance

Sl.No	Age	Salary	Status	Distance
1	0.054	0	N	1.04
2	0.324	0.333	N	0.611
3	0.595	0.667	N	0.185
4	0	0.917	Y	0.681
5	0.459	0.367	Y	0.514
6	1	1	Y	0.365

3 Nearest Neighbors

With K = 3 Loan approval status is predicted as “Yes” with majority class prediction

K-NN Classification Algorithm

Input : input, label pairs($\mathbf{x_1}$, y_1), ($\mathbf{x_2}$, y_2),…. ($\mathbf{x_i}$, y_i), $\mathbf{x_q}$ (test sample for which label needs to predicted).

Output: y_q (Predicted label for $\mathbf{x_q}$)

Procedure:

(a) Find K most similar (nearest) examples to $\mathbf{x_q}$ from the training dataset based on some distance measure.

(b) Get the labels of these K nearest examples.

(c) Predict the label of $\mathbf{x_q}$ by applying some aggregation (eg. Majority voting) on these labels of the K nearest neighbours.

K-NN Regression

Consider the employee dataset indicating the age, years of experience of the employee (input) and the Salary (output).

Age	Experience(yrs)	Salary (Lakhs)
28	3	22
26	4	25
30	5	30
34	8	35
38	15	40
46	20	42
48	25	47
55	30	70
52	23	60

What is the Expected salary of a person who is 32 years old with 7 years Experience?

K-NN Regression (cont…)

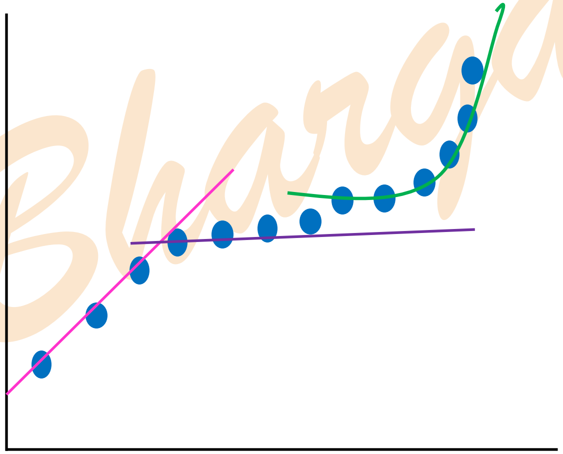
Age	Experience (Yrs)	Salary (Lakhs)	Distance (Euclidean)
28	3	22	5.656854
26	4	25	6.708204
30	5	30	2.828427
34	8	35	2.236068
38	15	40	10
46	20	42	19.10497
48	25	47	24.08319
55	30	70	32.52691
52	23	60	25.6125

3 Nearest Neighbors

With K = 3 Expected salary of the person is 29 Lakhs which is average of Nearest neighbors

Nearest Neighbors - Advantages

- Simple computations, easy to implement.
- Can learn complex decision boundaries.
- Performs well when a single function can not fit the entire input space.



Limitations/Disadvantages

- Intensive computations in the case of large data sets.
- Do not work well with high dimensional input (curse of Dimensionality)
  - Curse of Dimensionality: Distance metrics become less informative
  - Volume of data increases exponentially.
  - Chances more sparsity of input feature space due to data unavailability for all the features.
  - Computational complexity increases.
  - Models tend to overfit.
  - Difficulty in visualization
- Since K-NN is memory based, entire data needs to be preserved and carried around for predictions.
- Choosing the right distance metric and the value of ‘K’ can be difficult.