Big Data Analytics with Deep Learning

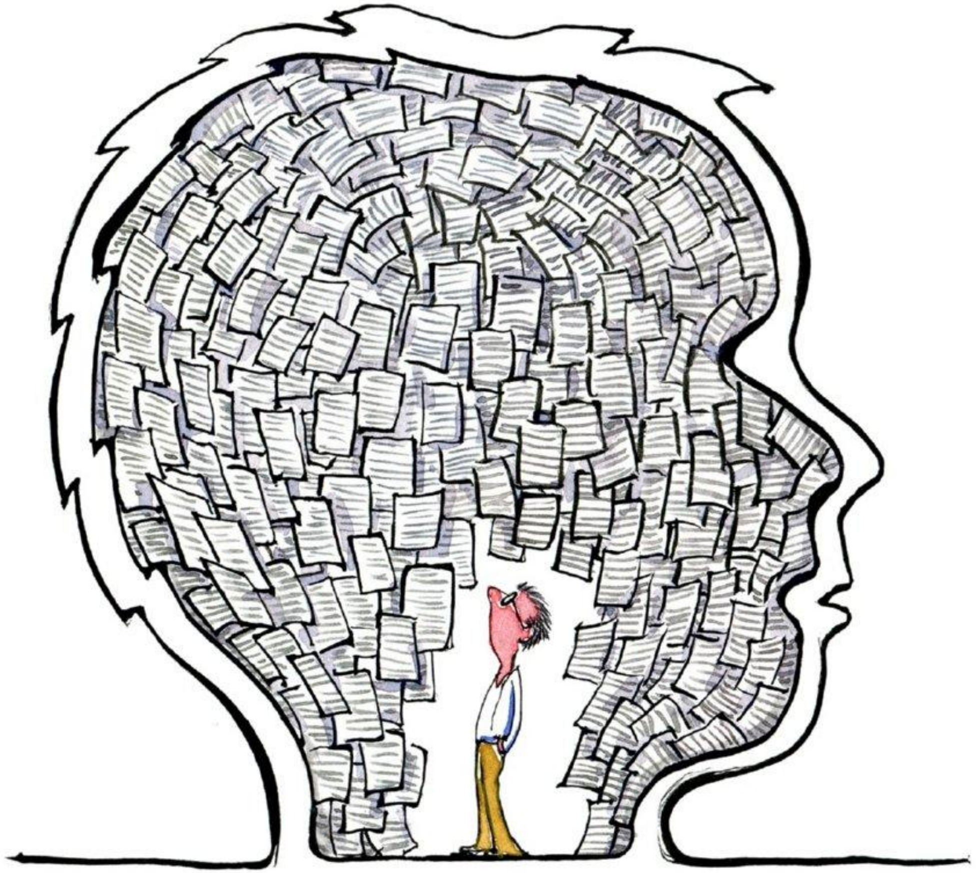
Dr. Tapan Kumar Jain,

Assistant Professor (ECE), IIIT Nagpur

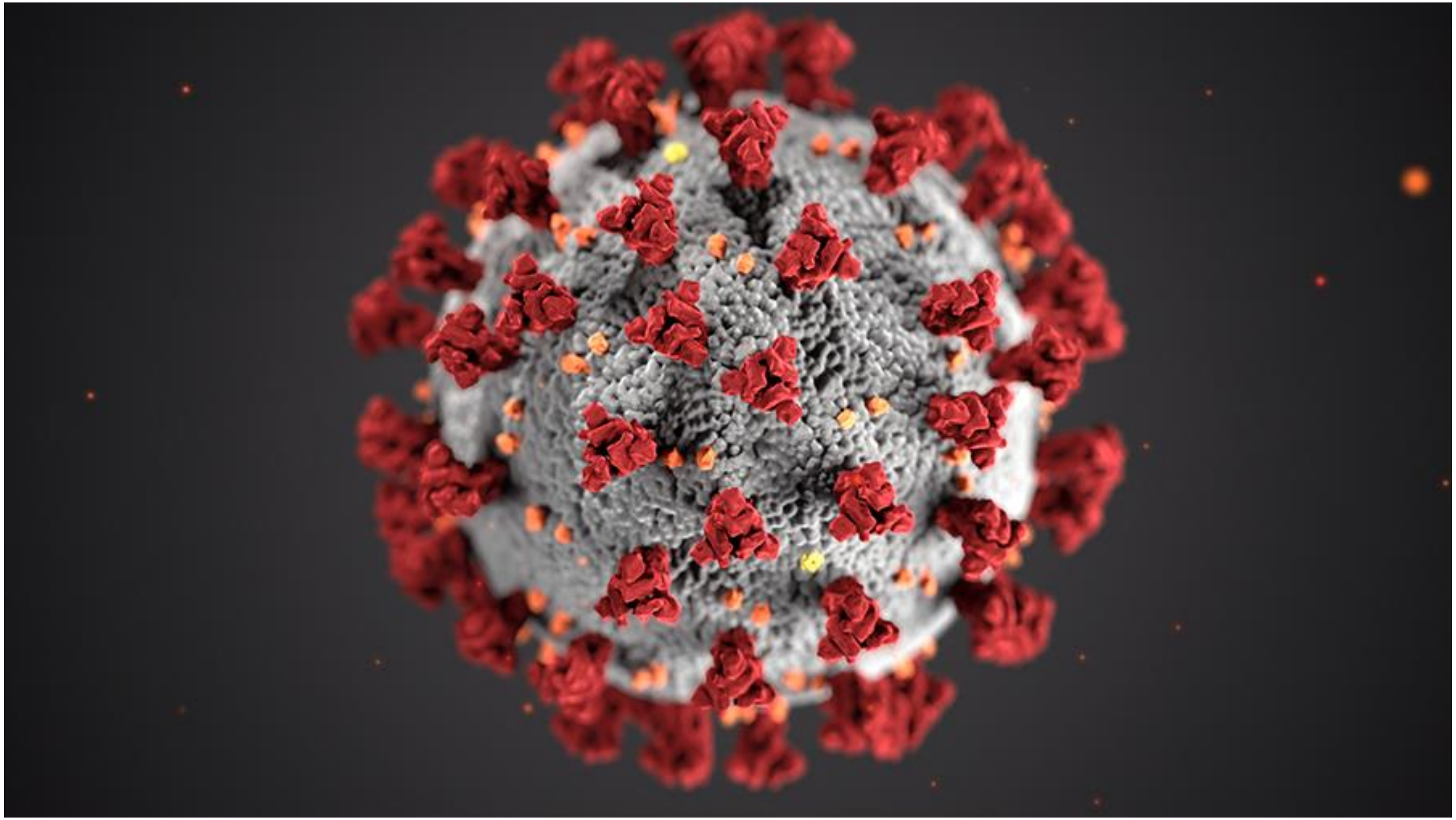
Senior Member, IEEE

IIIT Nagpur

1. IIIT NagpurEstablish: 2016
2. Director: Prof. O. G. Kakde
3. Dean: Dr. Ashwin Kothari
4. Registrar: Er. K. N. Dakhale
5. Departments: BSE, CSE, & ECE
6. BTech, PhD [Certification Course]



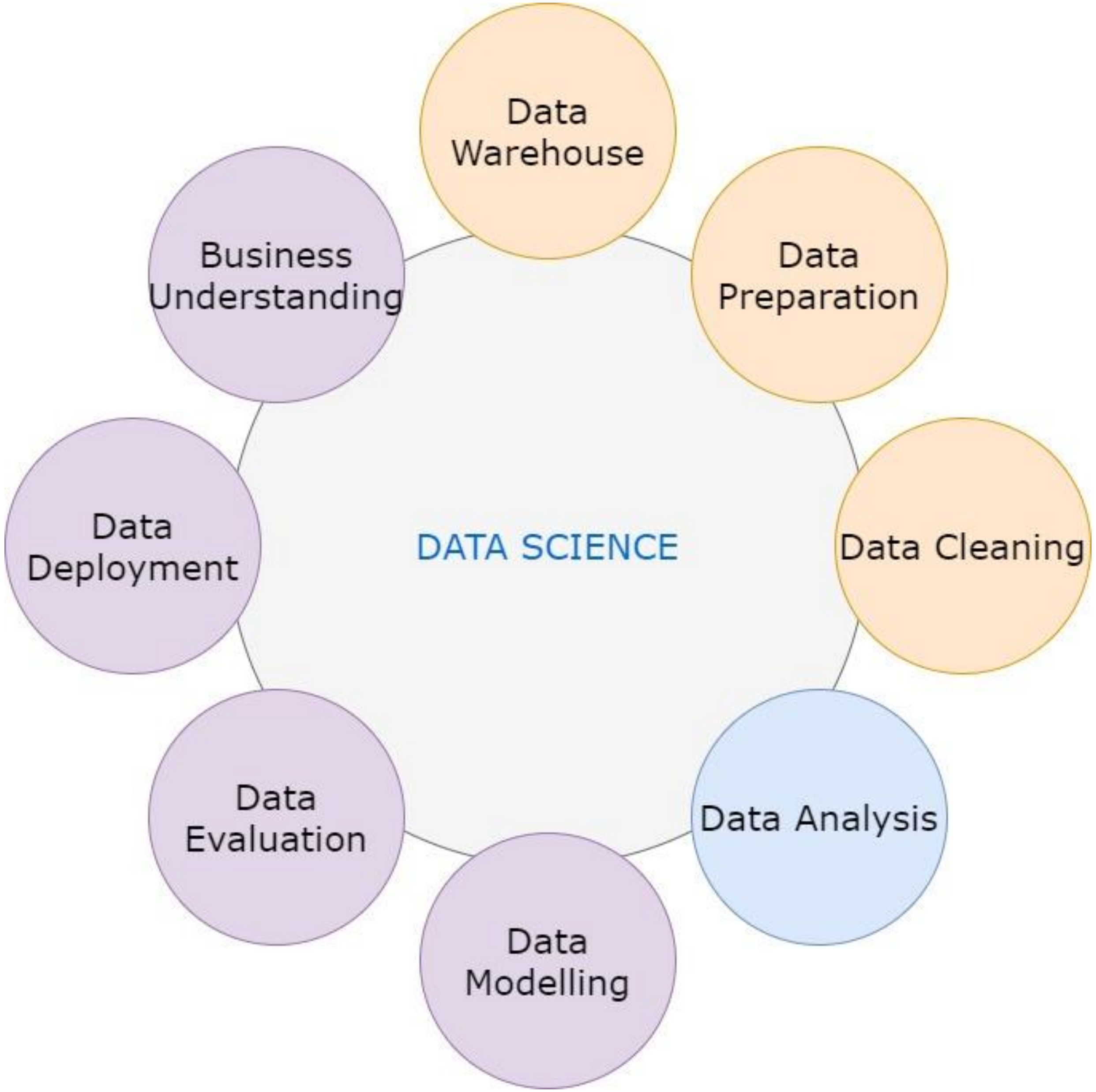
A picture is worth a thousand words,
Now Imagine the volume of Data

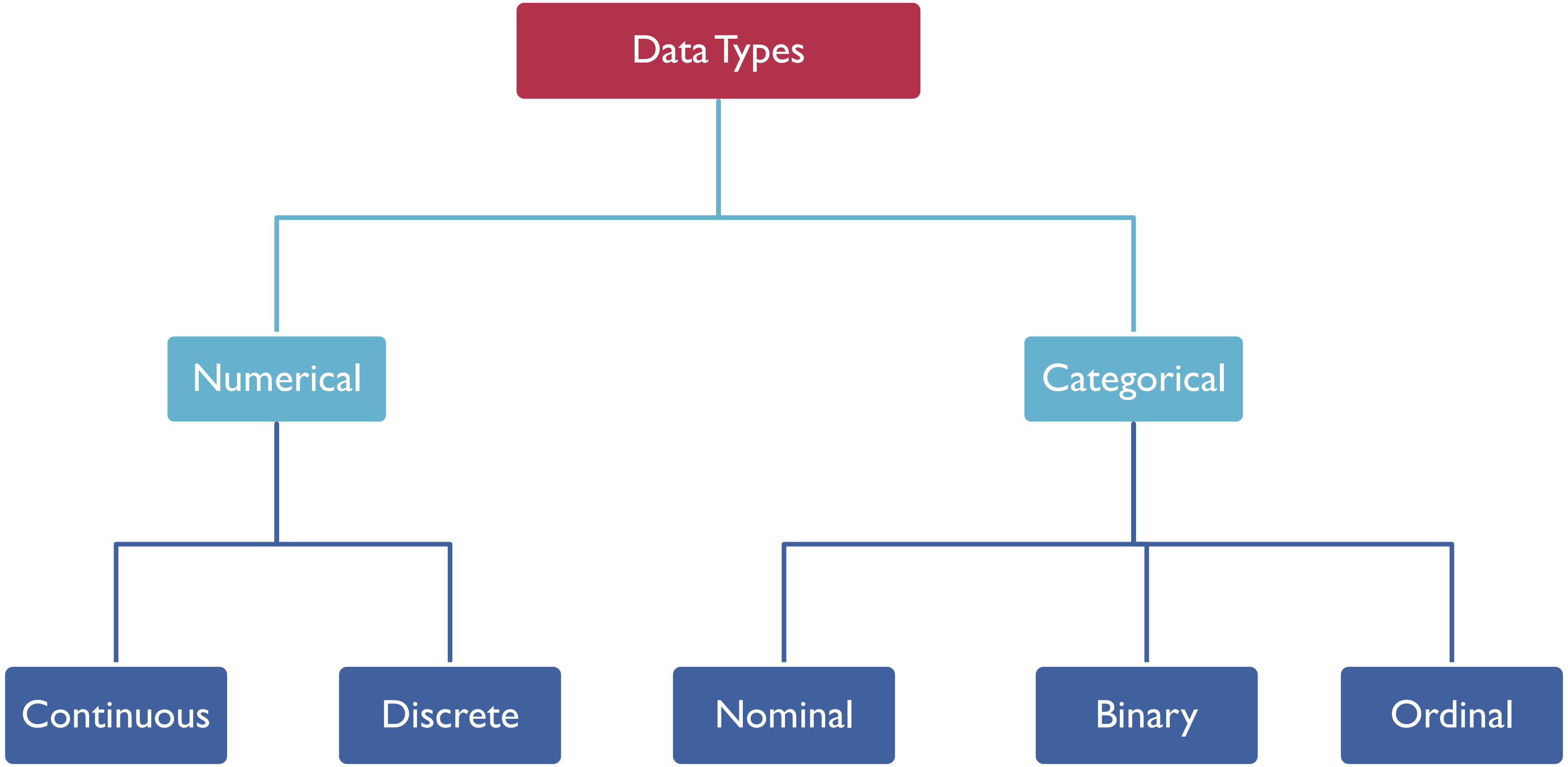


CORONA / CORONAVIRUS / COVID 19

Role of

- Data Engineer
- Data Scientist
- Data Analyst




Data Types
Numerical
Categorical
Continuous
Discrete
Nominal
Binary
Ordinal

Estimations (Location)

1. **Mean :** The sum of all values divided by the number of values. (Average)
2. **Weighted Mean :** The sum of all values times a weight divided by the sum of the weights. (weighted average)
3. **Median :** The value such that one-half of the data lies above and below. (50th percentile)
4. **Weighted Median :** The value such that one-half of the sum of the weights lies above and below the sorted data.
5. **Trimmed Mean :** The average of all values after dropping a fixed number of extreme values. (truncated mean)
6. **Robust :** Not sensitive to extreme values. (Resistant)
7. **Outlier :** A data value that is very different from most of the data. (extreme value)

Estimations (Variability)

1. **Deviations :** The difference between the observed values and the estimate of location. (errors, residuals)
2. **Variance :** The sum of squared deviations from the mean divided by *n* – *1 where n* is the number of data values. (mean-squared-error)
3. **Standard deviation:** The square root of the variance. (l2-norm, Euclidean norm)
4. **Mean absolute deviation:** The mean of the absolute value of the deviations from the mean. (l1-norm, Manhattan norm)
5. **Median absolute deviation from the median :** The median of the absolute value of the deviations from the median.
6. **Range:** The difference between the largest and the smallest value in a data set.

7. **Order statistics:** Metrics based on the data values sorted from smallest to biggest. (ranks)
8. **Percentile:** The value such that *P* percent of the values take on this value or less and (100–P) percent take on this value or more. (quantile)
9. **Interquartile range:** The difference between the 75th percentile and the 25th percentile.(IQR)

Supervised Learning Algorithm: kNN

k – Nearest Neighbours

1. kNN
2. Factor k
3. Examples

KNN – Different Nomenclature

1. K-Nearest Neighbors
2. Memory-Based Reasoning
3. Example-Based Reasoning
4. Instance-Based Learning
5. Lazy Learning

What is kNN

•kNN – k Nearest Neighbors, is one of the simplest Supervised Machine Learning algorithm mostly used for Classification It classifies a data point based on how its neighbors are classified.

•A powerful classification algorithm used in pattern recognition.

•k nearest neighbors stores all available cases and classifies new cases based on a similarity measure(e.g distance function)

•One of the top data mining algorithms used today.

•A non-parametric lazy learning algorithm (An Instance- based Learning method).

Problem: Find the green data point class?

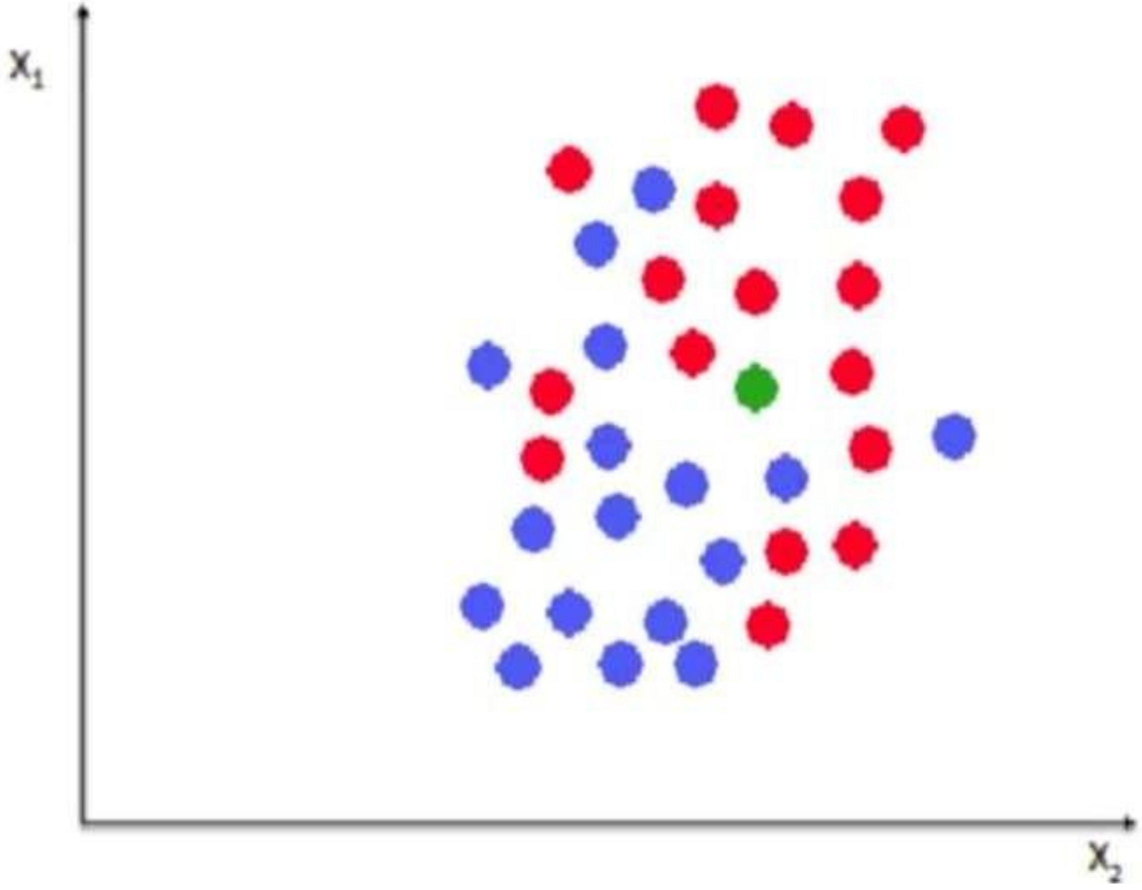


Figure A

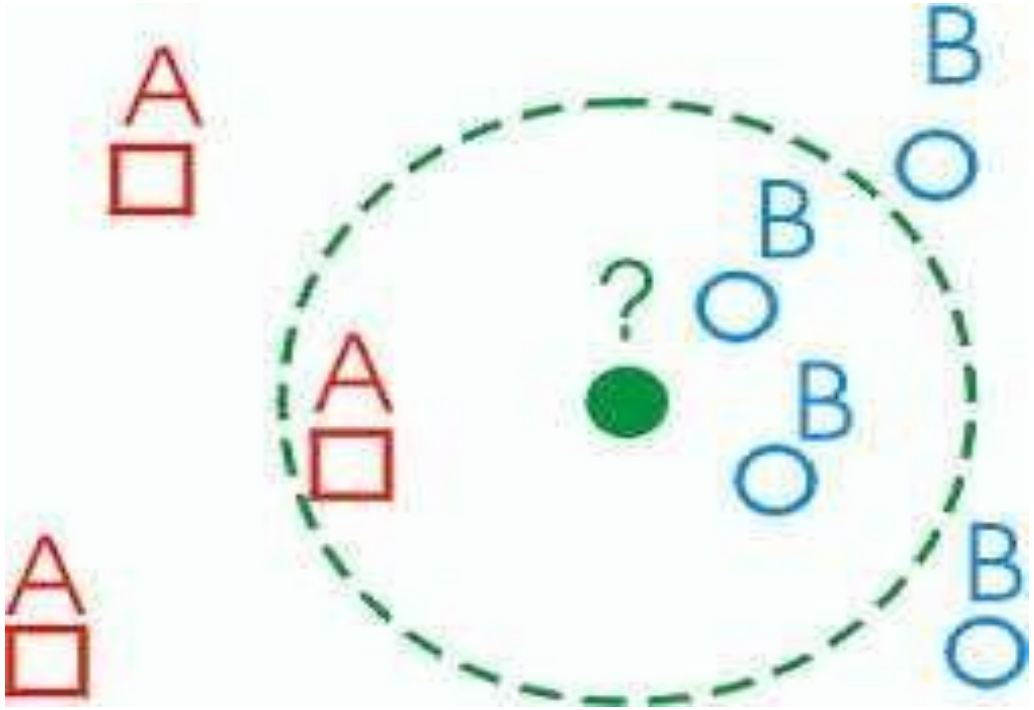


Figure B

kNN: k Factor

1. k in kNN is a parameter that refers to the number of nearest neighbors to include in the majority voting process.
2. kNN Algorithm is based on feature similarity: Choosing the right value of k is a process called parameter tuning, and is important for better accuracy.
3. If K is too small it is sensitive to noise points. (k=3)
4. Larger K works well. But too large K may include majority points from other classes.
5. Odd value of k is selected to avoid confusion.
6. Maximum value of k = sqrt(n), where n is the data points

When do we use kNN

1. We can use when data is labeled.
2. Data is noise free
3. Data is small (computational time)

kNN : Example - I

Customer	Age	Income	No. credit cards	Class	Distance from Tapan
Anne	35	35K	3	No	sqrt [$(35-37)^2+(35-50)^2+(3-2)^2$]=15.16
John	22	50K	2	Yes	sqrt [$(22-37)^2+(50-50)^2+(2-2)^2$]=15
George	63	200K	1	No	sqrt [$(63-37)^2+(200-50)^2+(1-2)^2$]=152.23
Kevin	59	170K	1	No	sqrt [$(59-37)^2+(170-50)^2+(1-2)^2$]=122
Tom	25	40K	4	Yes	sqrt [$(25-37)^2+(40-50)^2+(4-2)^2$]=15.74
Tapan	37	50K	2	?	

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

kNN : Example - II

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
48	**$142,000**	**Y**	

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

sqrt [(48-25)^2+(142000-40000)^2]=102000.0025

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

kNN : Example - II

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**N**	

$$a_i = \frac{v_i - \min v_i}{\max v_i - \min v_i}$$

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Strengths of kNN

1. We select the k entries in our database which are closest to the Very simple and intuitive.
2. Can be applied to the data from any distribution.
3. Good classification if the number of samples is large enough.

Weaknesses of kNN

1. Takes more time to classify a new example.
2. Need to calculate and compare distance from new example to all other examples.
3. Choosing k may be tricky.
4. Need large number of samples for accuracy.

Summary

1. A positive integer k is specified, along with a new sample
2. We select the k entries in our database which are closest to the new sample
3. We find the most common classification of these entries
4. This is the classification we give to the new sample

Supervised Learning Algorithm: Decision Tree

Decision Tree

1. What is Decision Tree
2. Terminologies
3. Different Criterion
4. Pros / Cons of Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Decision Tree

Thank you

tapankumarjain@gmail.com