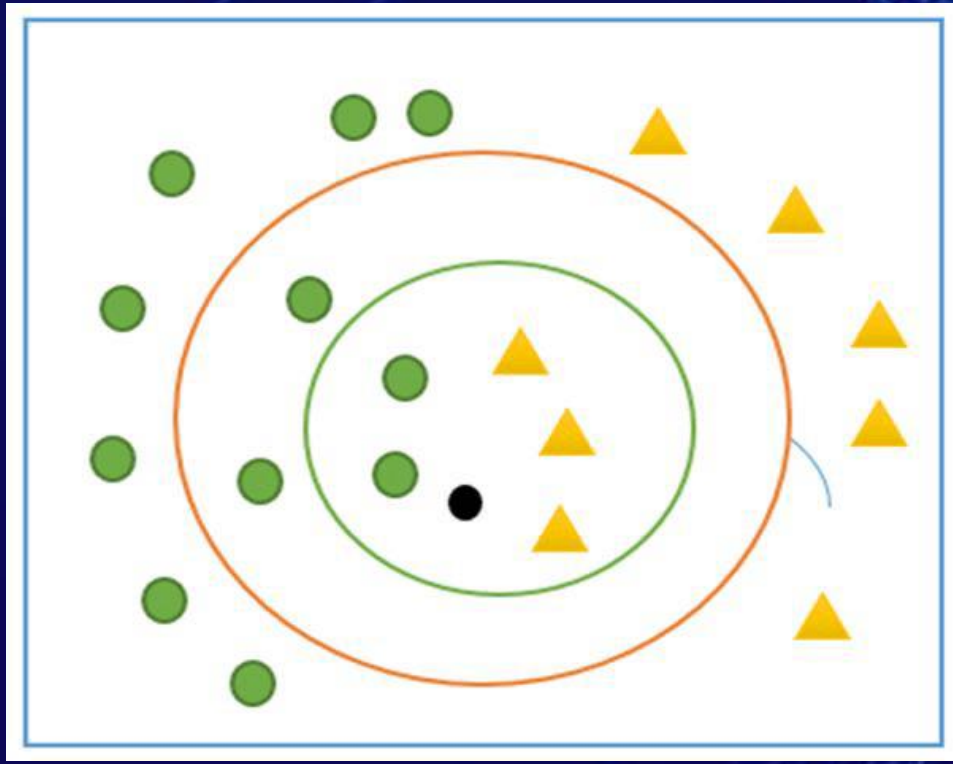


Unit 2.3
K-Nearest Neighbours

Reference

Disclaimer

The content is curated from online/offline resources and used for educational purpose only

K-Nearest Neighbours

Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.

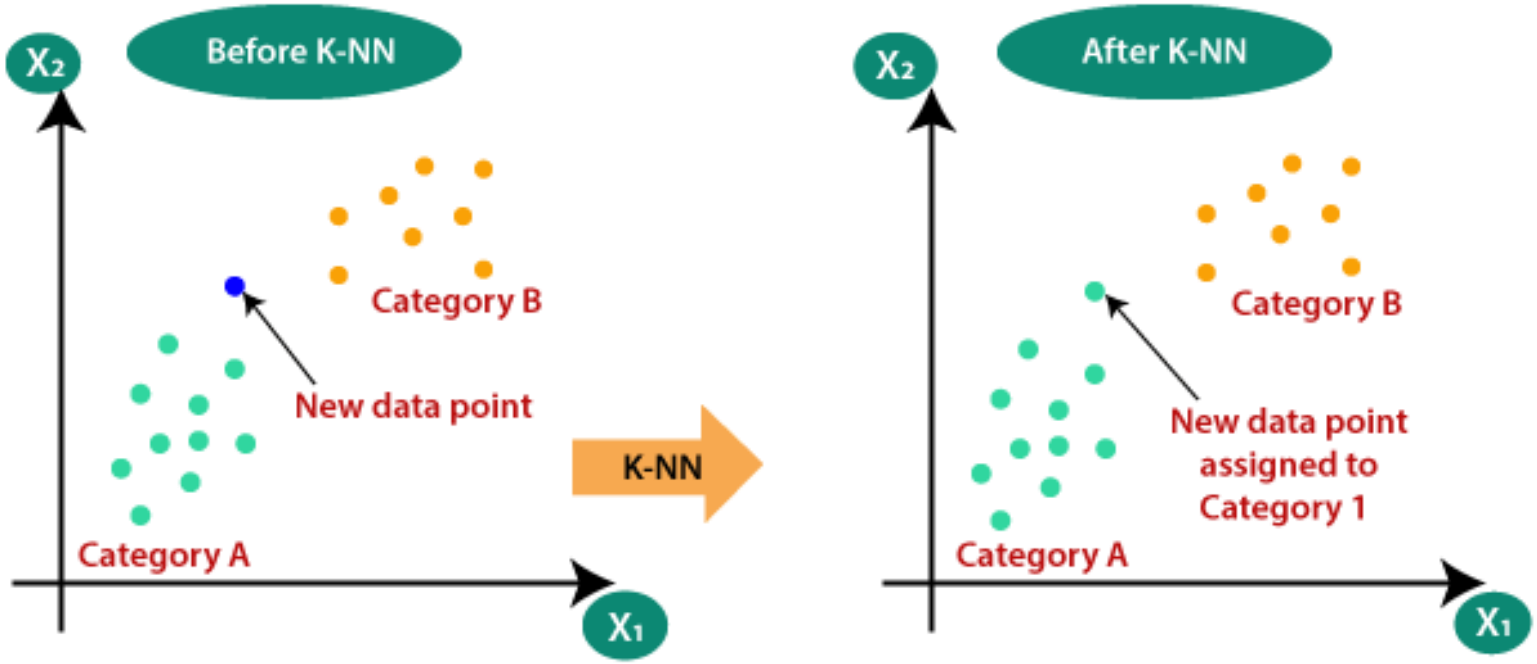
KNN

Learning Objectives

You will learn in this lesson:

- Concept of Neighbour
- Data Similarity Measure
- Modelling
- Estimation of K Neighbours
- Model Inferencing

Why do we need a K-NN Algorithm?



Reference

Introduction

- It is a supervised learning algorithm and easy to implement.
- It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category.
- K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- used for Regression as well as for Classification but mostly it is used for the Classification problems.
- Major challenge is to identify value K

Introduction

- Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs' images and based on the most similar features it will put it in either cat or dog category.

Reference

Telecom Customer Dataset

- In Telecommunication dataset, with predefined labels, need to build a model which is used to predict the class of a new or unknown case.
- The example focuses on using demographic data to predict usage patterns.

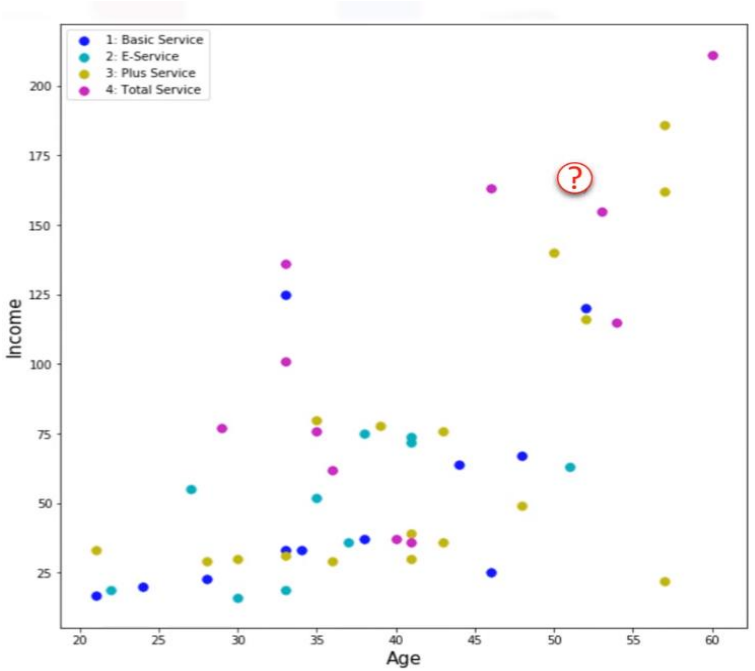
Features | Labels

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

KNN Classifier

Telecom Customer Dataset

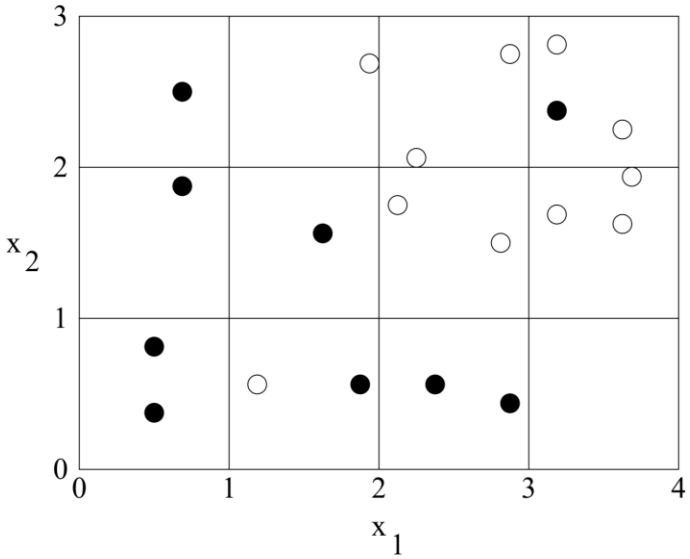
- Objective is to build a classifier, using the rows 0 to 7, to predict the class of row 8.
- We will use a specific type of classification called K-nearest neighbor.
- Just for sake of demonstration, let's use only two fields as predictors - specifically, **Age** and **Income**, and then plot the customers based on their group membership.



Telecom Dataset

Intuition of Nearest Neighbour

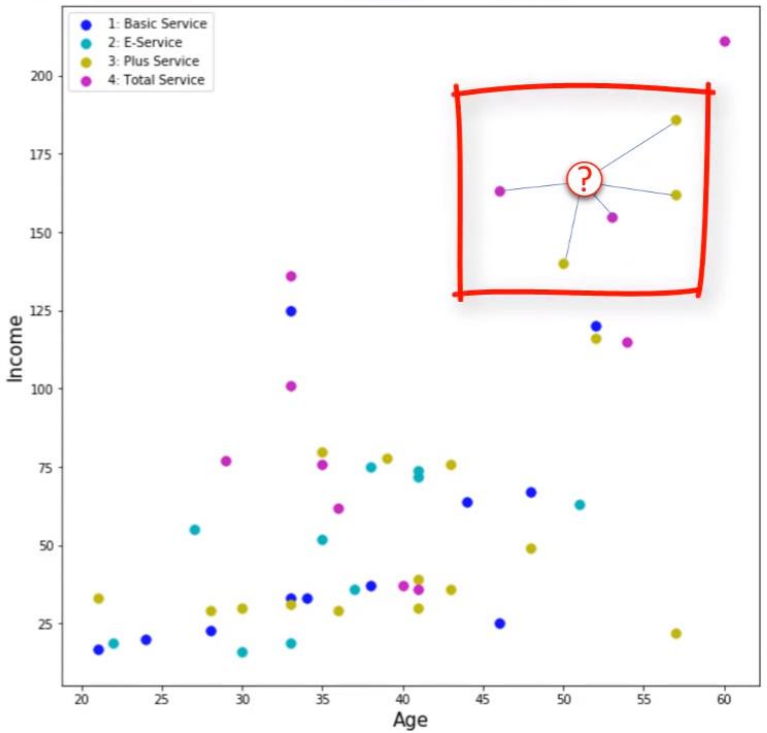
- How can we find the class of new customer, available at record number 8 with a known age and income?
- Can we say that the class of our new customer is most probably group 4 because its nearest neighbour is also of class 4?
- Yes, we can say so!



Notion of Neighbourhood

Inference

- Now, the question is, “To what extent can we trust our judgment, which is based on the first nearest neighbor?”
- It might be a poor judgment, especially if the first nearest neighbor is a very specific case, or an outlier !
- What if we chose the five nearest neighbors, and did a majority vote among them ?



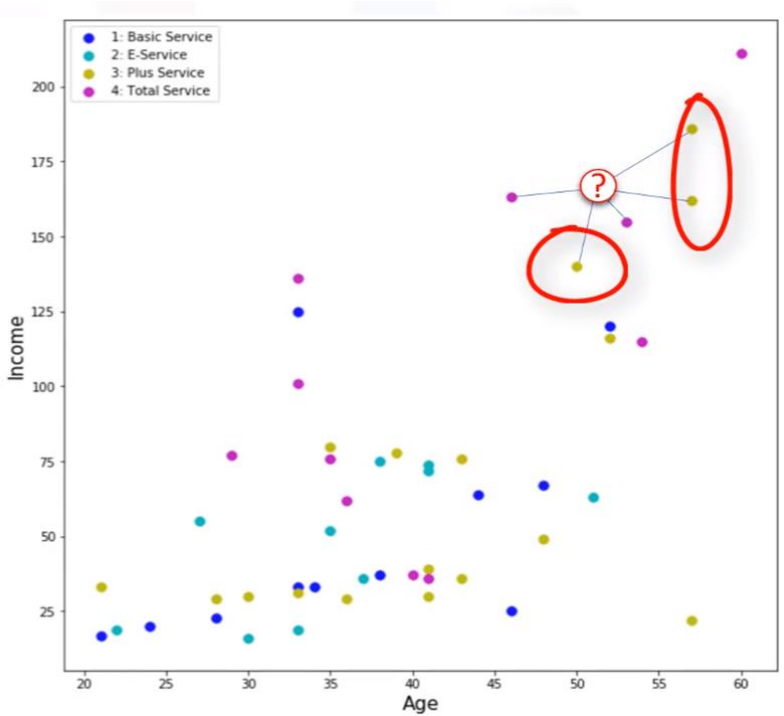
Telecom Dataset

Decision Resolving

- Does this make more sense?

  Yes !

- In this case, the value of K in the k-nearest neighbours' algorithm is 5.
- This example highlights the intuition behind the k-nearest neighbours' algorithm.



Telecom Dataset

Implementation Steps

In a classification problem, the k-nearest neighbors algorithm is implemented using following steps:

1. Pick a value for K.
2. Calculate the distance(of Similarities) of unknown case from all cases.
3. Search for the K observations in the training data that are ‘nearest’ to the measurements of the unknown data point.
4. Predict the response of the unknown data point using the most popular response value from the K nearest neighbors.

Similarity between Data Points

How can we calculate the similarity between two data points?

- Assume that we have two customers, customer 1 and customer 2 who have only one feature, Age.
- We can easily use a specific type of Euclidean distance to calculate the distance of these 2 customers.
- Lower distance resembles higher similarity.

$$\boldsymbol{Dis}(\boldsymbol{x},\boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{i}=\boldsymbol{0}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{y_i})^{\boldsymbol{2}}}$$

Similarity between Data Points

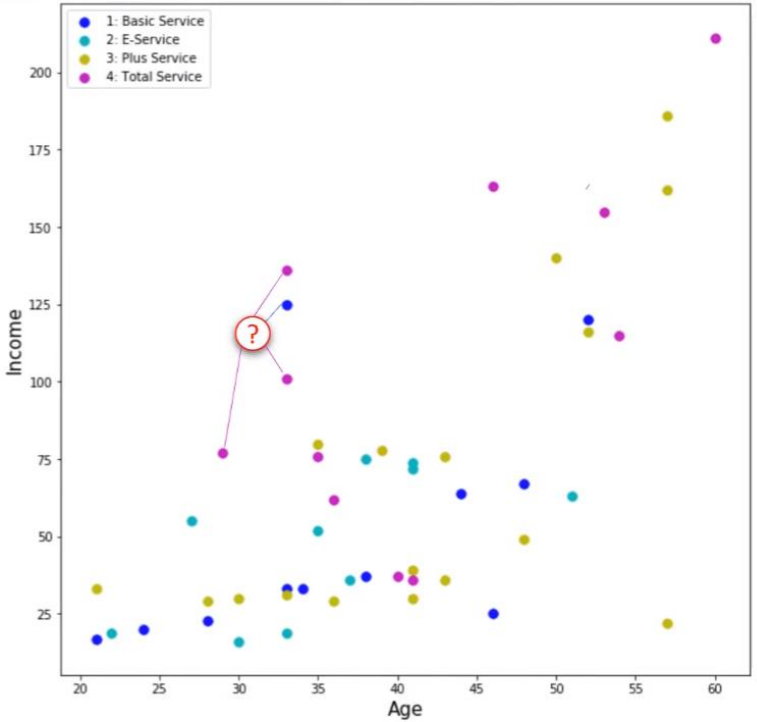
- Age of customer 1 = 54 and
- Age of customer 2 = 50,
- Distance between both customer 1 & customer 2 “age” feature are :

  Dis(x,y)=√((54-50)^2)=4
- If we have both income and age features of both customers.
- Age of customer 1 = 54 and income = 250
- Age of customer 2 = 50 and income = 240
- Distance between Customer 1 & Customer 2 “age” and “income”

  Dis(x,y)=10.77

Value of K ?

- A low value of K causes a highly complex model, which might result in over-fitting of the model.
- It means the prediction process is not generalized enough to be used for out-of-sample cases.



Telecom Dataset

Optimizing K?

- So, how we can find the best value for **K**?
- Calculate the accuracy of the model by choosing **K=1** using all samples in your test set.
- Repeat this process, increasing the **K**, and see which **K** is best for your model.
- In this example, **K=4** gives the **best accuracy.**

Lab 1 – Implement K-NN Machine Learning Algorithm

Summary

- KNN is Supervised Machine Learning Algorithm
- We have seen Euclidean distance to find the similarities between data points
- Working of KNN with telecom customer dataset
- Choosing correct or Optimum value of K is important

Quiz

1) Which is the number of nearby neighbours to be used to classify the new record ?

a. KNN
b. Validation data
c. Euclidean Distance
d. All the above

a. KNN

Quiz

2) Which of the following option is true about k-NN algorithm?

a) It can be used for classification
b) It can be used for regression
c) It can be used in both classification and regression
d) None of the above

c). It can be used in both classification and regression

Quiz

3) Which of the following distance metric we have used in k-NN?

a) Manhattan
b) Minkowski
c) Tanimoto
d) Euclidean

d). Euclidean

Quiz

4) K-NN algorithm does more computation on test time rather than train time.

a TRUE
b) FALSE

a) TRUE

Reference

- https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
- https://towardsdatascience.com/basic-probability-theory-and-statistics-3105ab637213
- https://www.analyticsvidhya.com
- https://www.researchgate.com
- https://www.towardsdatascience.com
- https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/

Thank you...!