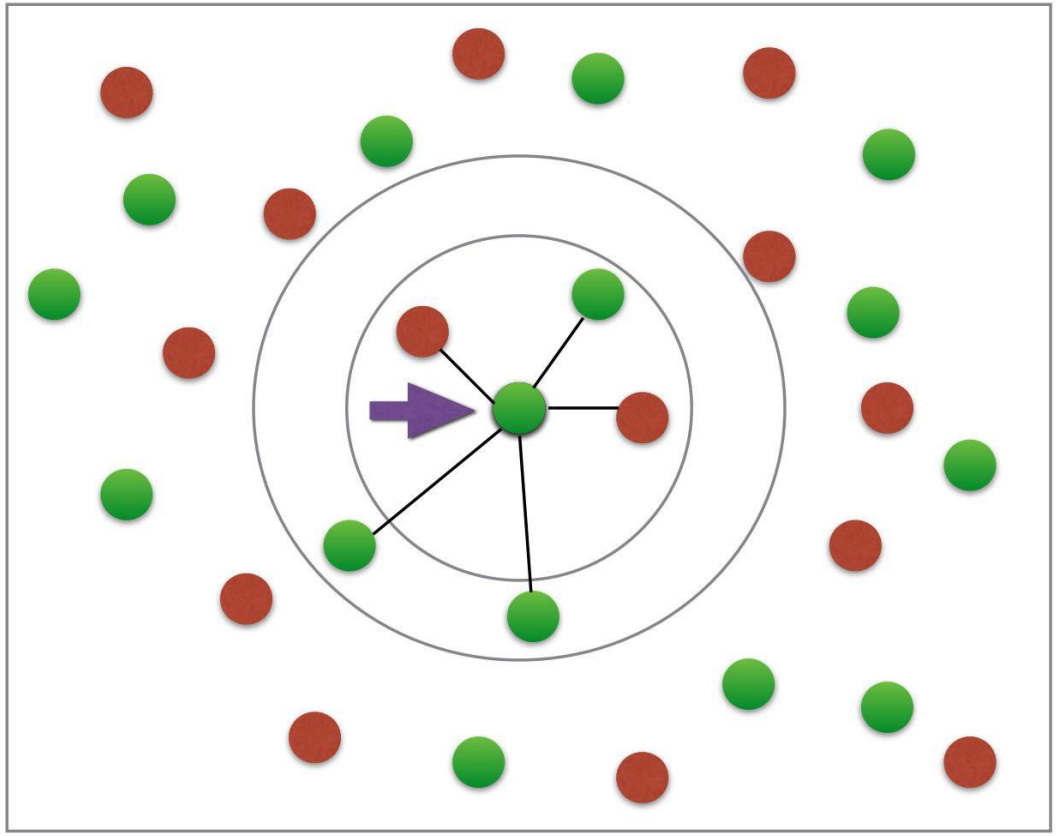
K-Nearest Neighbours

Objective

- Nearest Neighbours
- Telecome customer dataset
- Inference
- Implementation of KNN
- Feature Normalization
- Identify value of K

Introduction

- It is a supervised learning algorithm
- Simple to implement and most widely used machine learning Inference
- K-NN can outperform more powerful classifiers
- Non-parametric method for pattern classification
- Major challenge is to identify value K



Reference : KNN Classifier

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

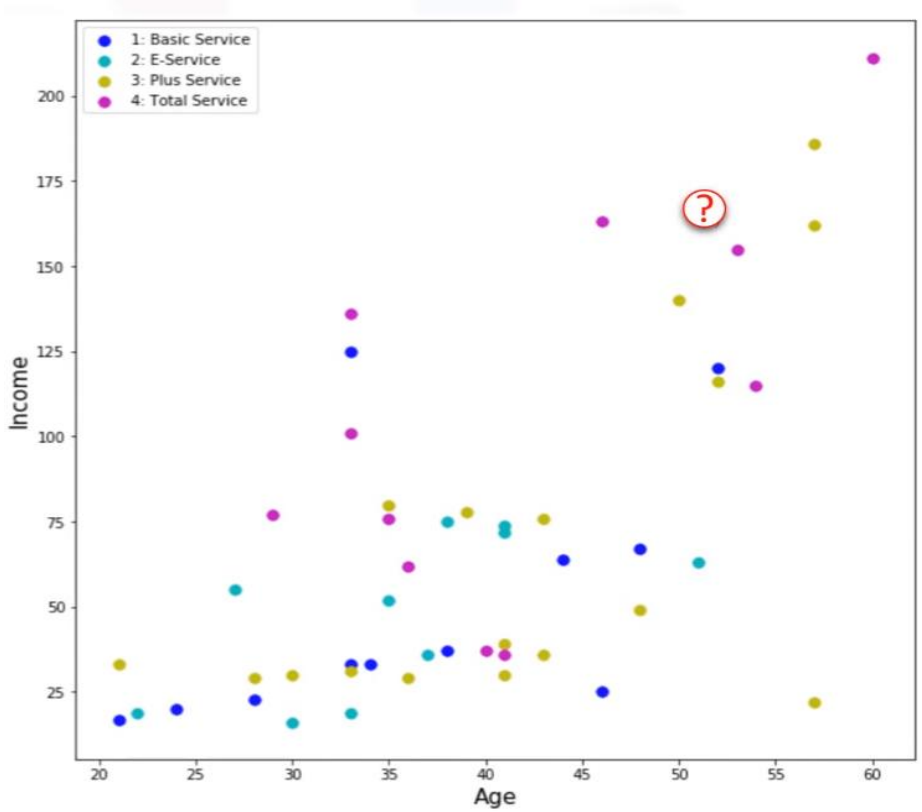
Labels

	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	?

Reference : 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.



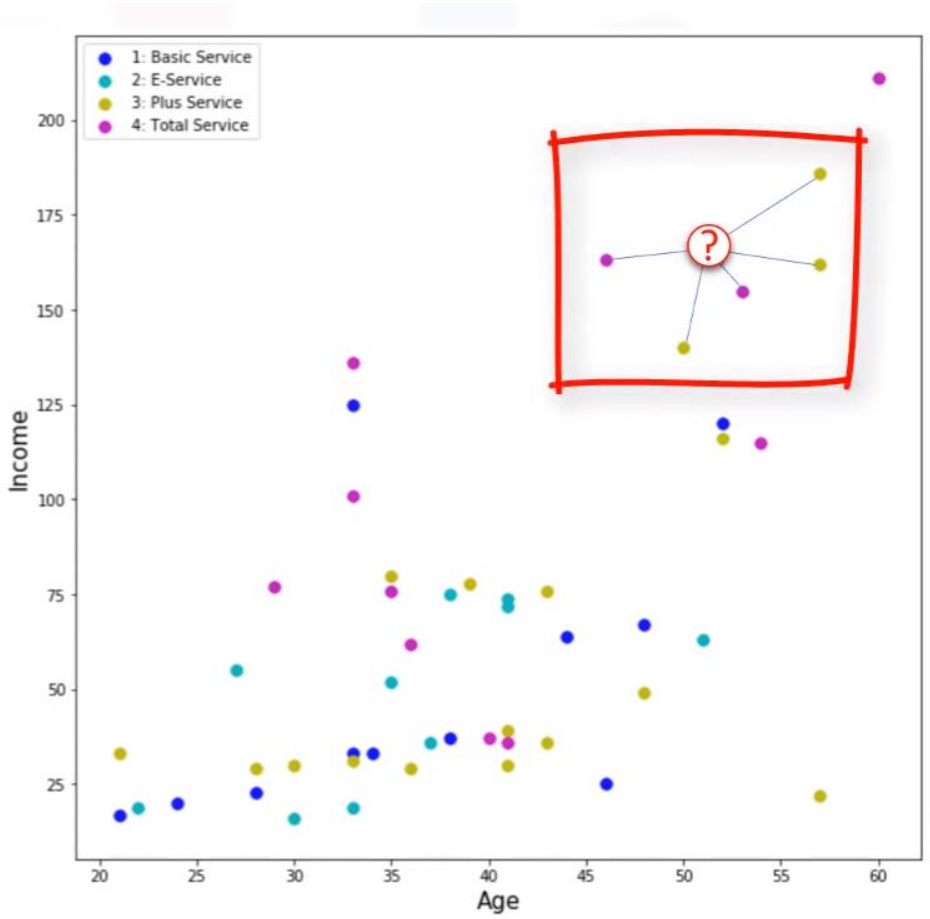
Reference : Telecom Dataset

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!

Inference

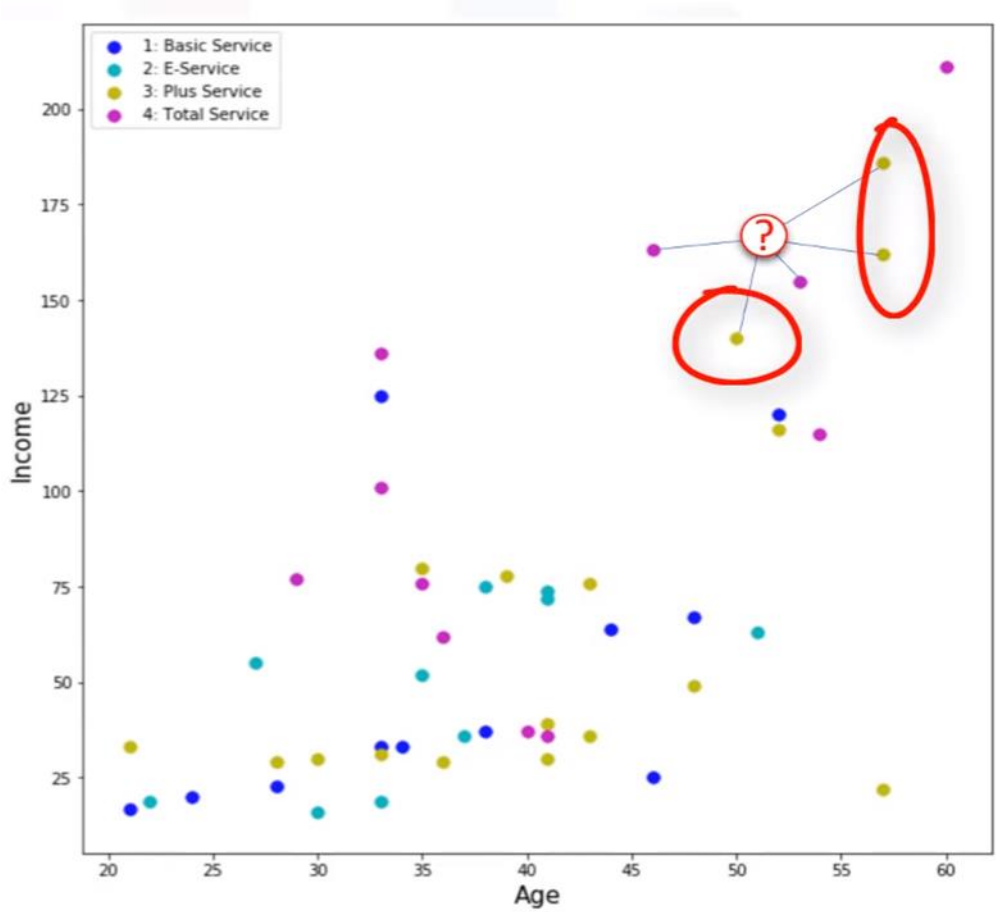
- Now, the question is, “To what extent can we trust our judgment, which is based on the first nearest neighbor?”
- It might be a poor judgment, especially if the first nearest neighbor is a very specific case, or an outlier !
- What if we chose the five nearest neighbors, and did a majority vote among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



Reference : 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.



Reference : Telecom Dataset

Implementation steps

- In a classification problem, the k-nearest neighbors algorithm is implemented using following steps:
- Pick a value for K.
- Calculate the distance of unknown case from all cases.
- Search for the K observations in the training data that are ‘nearest’ to the measurements of the unknown data point.
- 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.

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

Similarity between data points

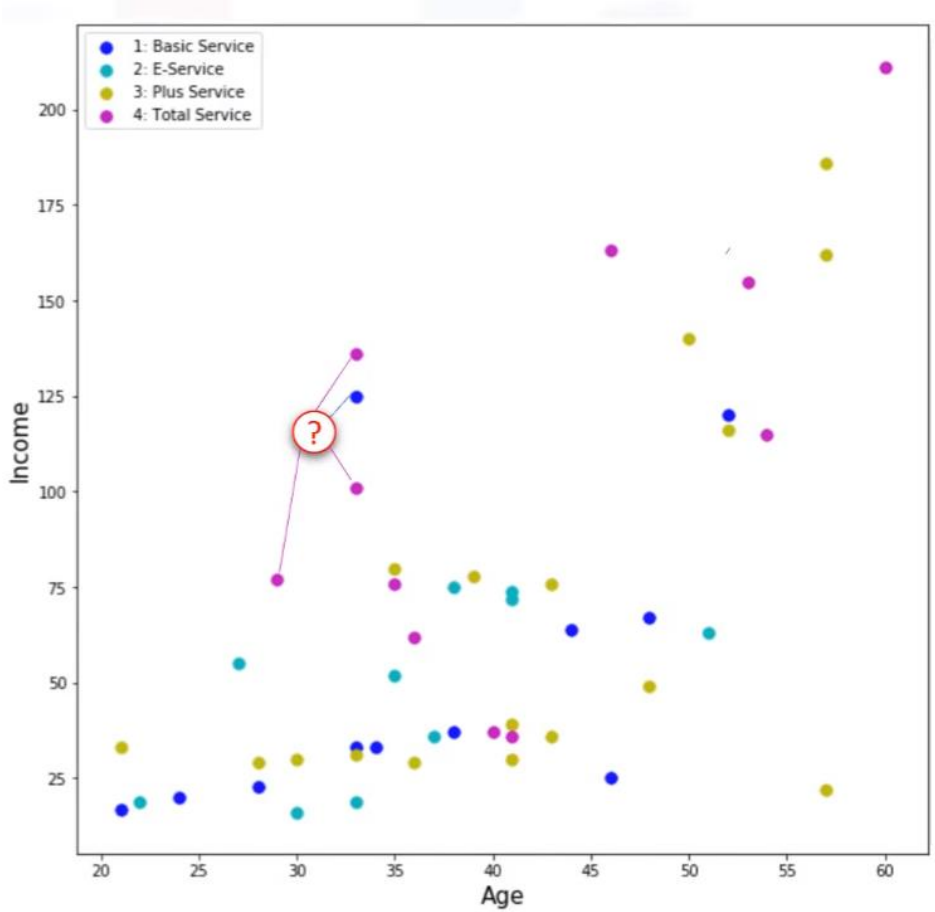
- Age of customer 1 = 54 and
- Age of customer 2 = 50,
- Distance between both customer 1 & customer 2 “age” feature are :

  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.



Reference : 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.**

Hands On