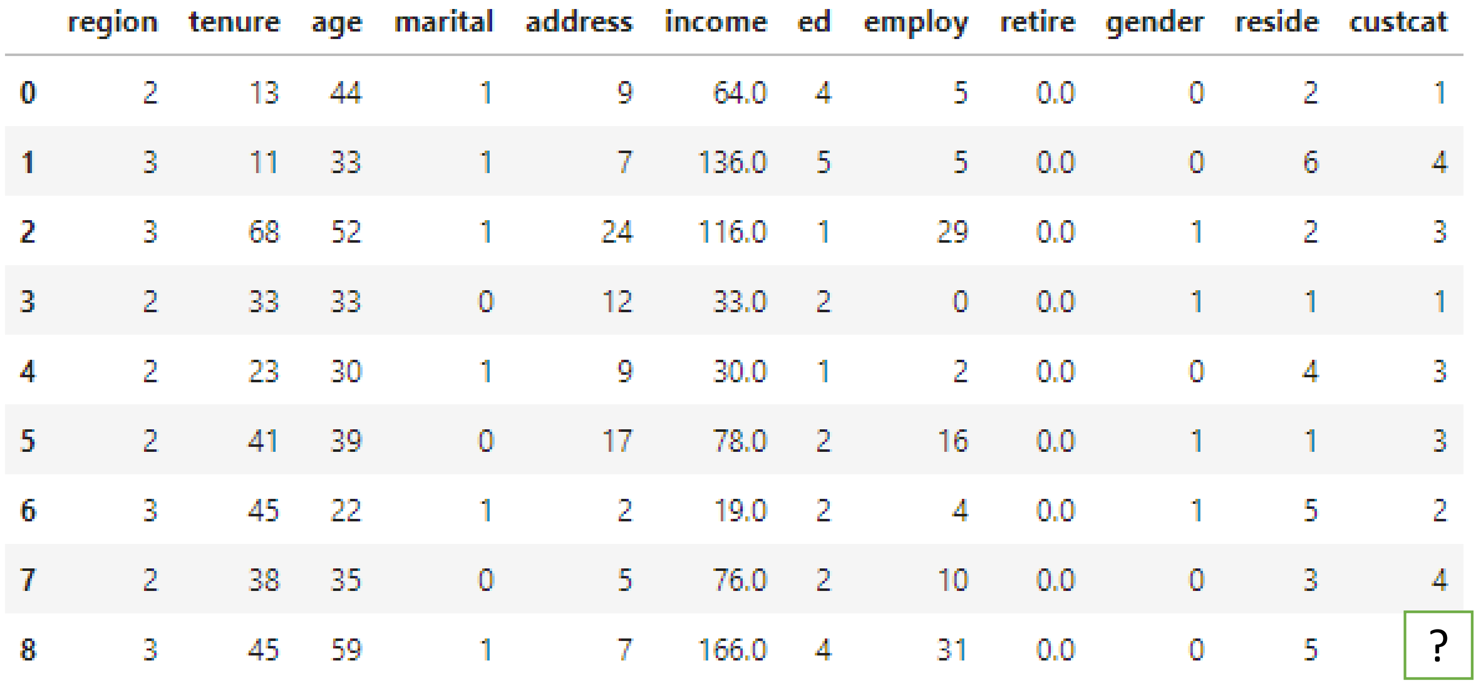
K-Nearest Neighbors

Intro to KNN

X: Independent variable

Y: Dependent variable

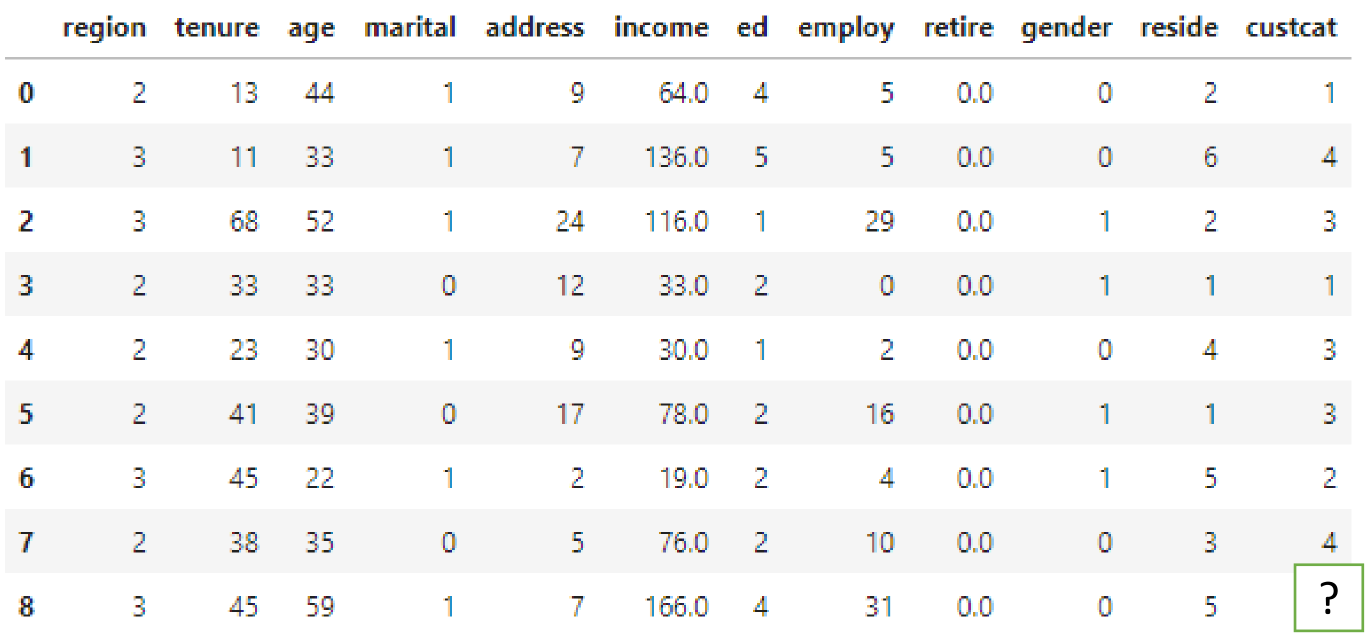


	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	13	44	1	9	64.0	4	5	0.0	0	2	1
1	3	11	33	1	7	136.0	5	5	0.0	0	6	4
2	3	68	52	1	24	116.0	1	29	0.0	1	2	3
3	2	33	33	0	12	33.0	2	0	0.0	1	1	1
4	2	23	30	1	9	30.0	1	2	0.0	0	4	3
5	2	41	39	0	17	78.0	2	16	0.0	1	1	3
6	3	45	22	1	2	19.0	2	4	0.0	1	5	2
7	2	38	35	0	5	76.0	2	10	0.0	0	3	4
8	3	45	59	1	7	166.0	4	31	0.0	0	5	?

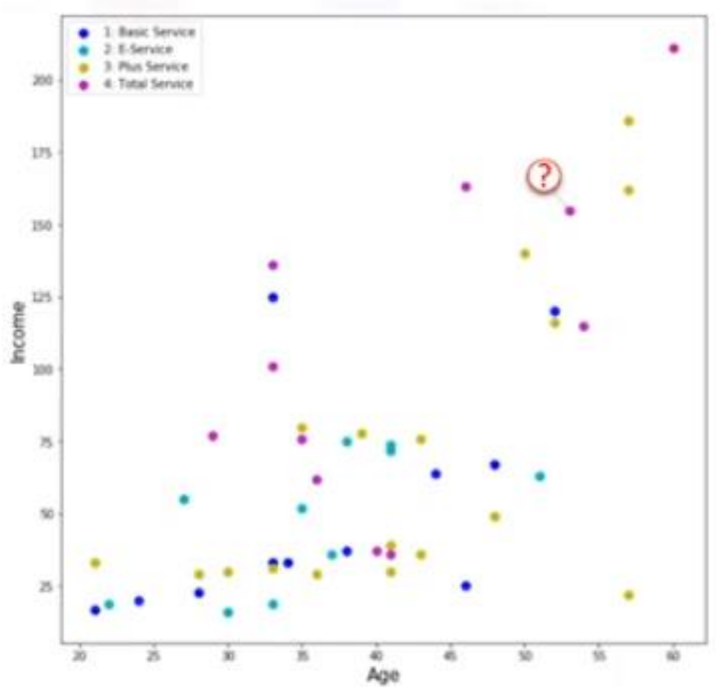
Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Our objective is to build the classifier

Determining the class using 1st KNN

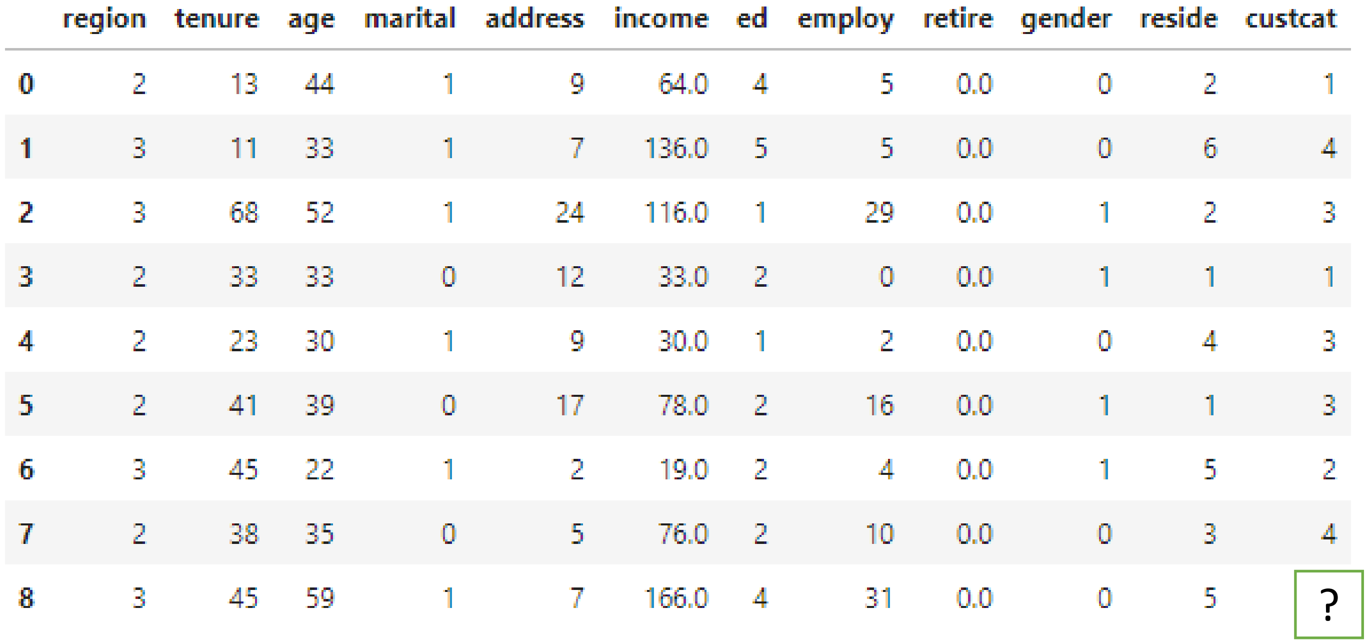


	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	13	44	1	9	64.0	4	5	0.0	0	2	1
1	3	11	33	1	7	136.0	5	5	0.0	0	6	4
2	3	68	52	1	24	116.0	1	29	0.0	1	2	3
3	2	33	33	0	12	33.0	2	0	0.0	1	1	1
4	2	23	30	1	9	30.0	1	2	0.0	0	4	3
5	2	41	39	0	17	78.0	2	16	0.0	1	1	3
6	3	45	22	1	2	19.0	2	4	0.0	1	5	2
7	2	38	35	0	5	76.0	2	10	0.0	0	3	4
8	3	45	59	1	7	166.0	4	31	0.0	0	5	?

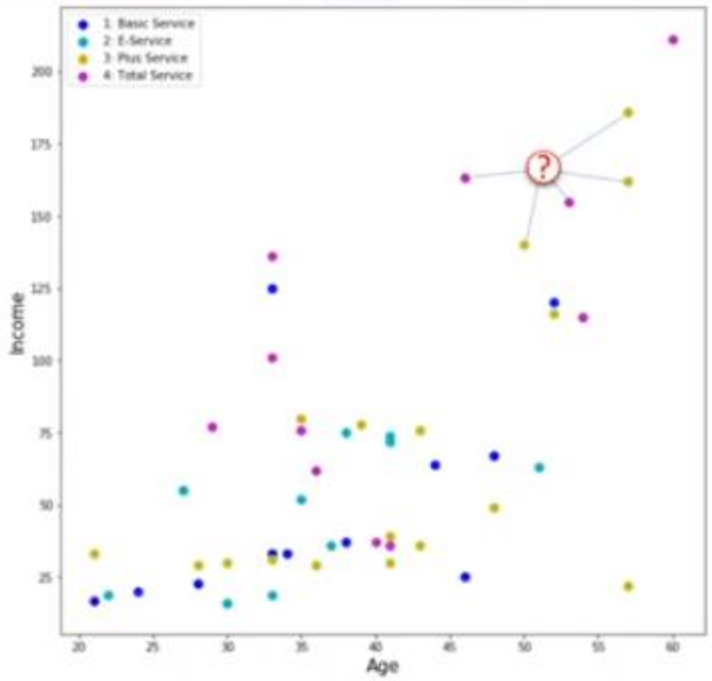


- With the known age and income, how to we predict the class of the costumer?
- To what extent can we trust our judgement which is based on the first nearest neighbor? The first neighbor can be the outlier.

Determining the class using 5 KNNs



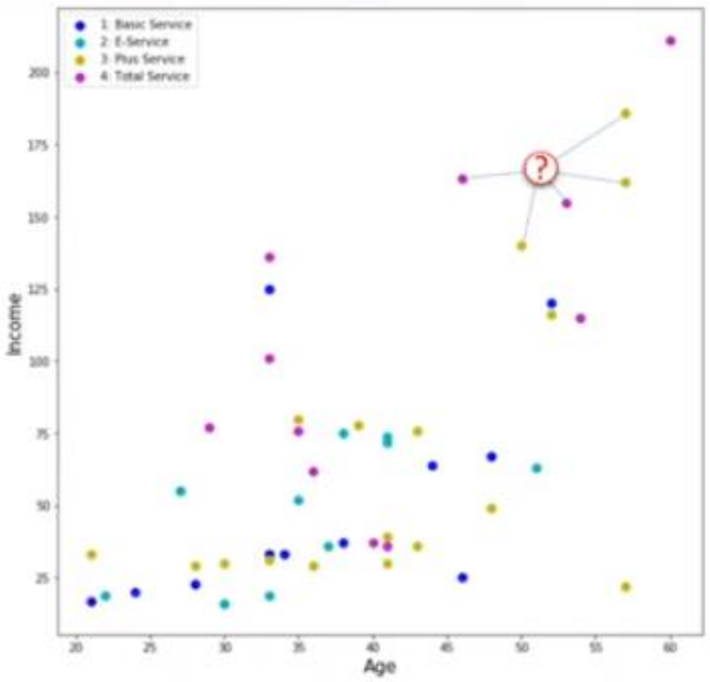
	region	tenure	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	13	44	1	9	64.0	4	5	0.0	0	2	1
1	3	11	33	1	7	136.0	5	5	0.0	0	6	4
2	3	68	52	1	24	116.0	1	29	0.0	1	2	3
3	2	33	33	0	12	33.0	2	0	0.0	1	1	1
4	2	23	30	1	9	30.0	1	2	0.0	0	4	3
5	2	41	39	0	17	78.0	2	16	0.0	1	1	3
6	3	45	22	1	2	19.0	2	4	0.0	1	5	2
7	2	38	35	0	5	76.0	2	10	0.0	0	3	4
8	3	45	59	1	7	166.0	4	31	0.0	0	5	?



In this case, K = 5

What is K-Nearest Neighbor (or KNN)?

- A method for classifying cases based on their similarity to other cases
- Cases that are near to each other are said to be neighbors
- Based on similar cases with same class labels are near each other



The K-Nearest Neighbors Algorithm

1. Pick a value for K.
2. Calculate the distance of unknown case from all the cases.
3. Select the K-observations in the training data that are “nearest” to the unknown data point.
4. Predict the response of the unknown data point using the most popular response value form the K-nearest neighbors.

Questions

- How to select K?
- How to compute the similarity between cases?

Calculate the similarity/distance in 1-dimensional space

Customer
Age
54

Customer
Age
50

$$Dis(x_1, x_2) = \sqrt{\Sigma_{i=0}^{n}(x_{1i} - x_{2i})^2} = \sqrt{(54-50)^2}\text{=4}$$

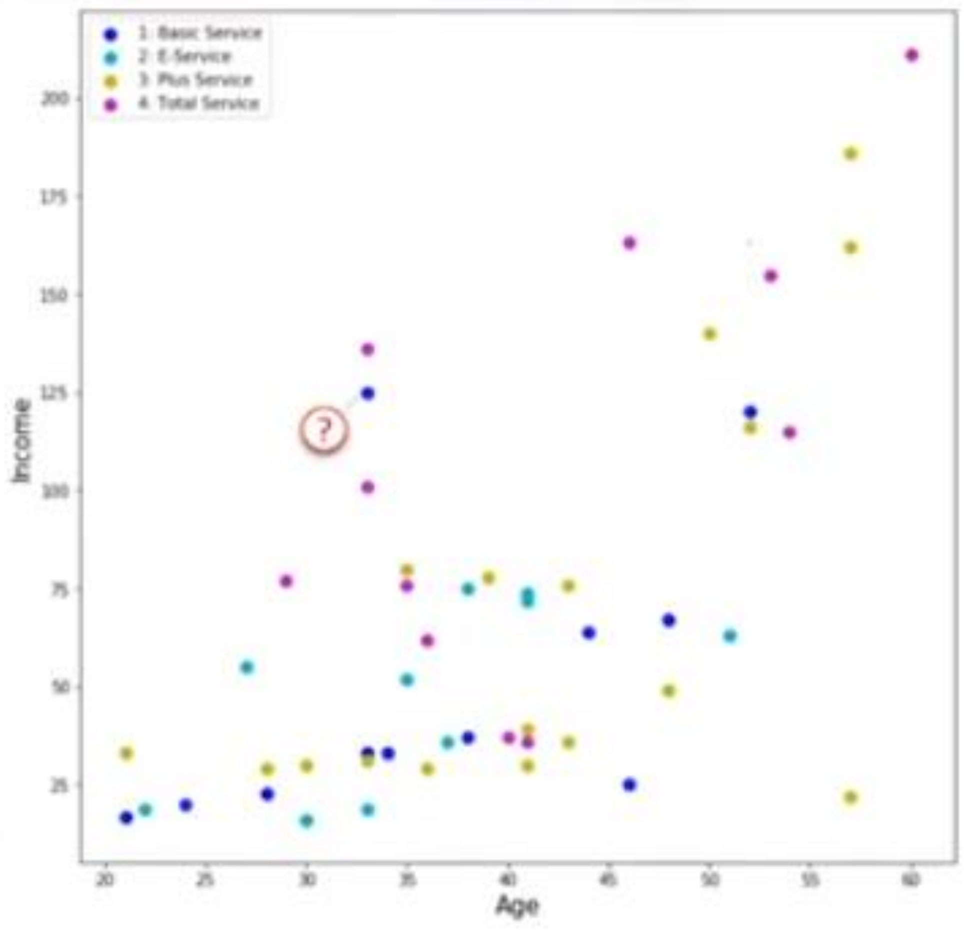
Customer 1	
Age	Income
54	190

Customer 2	
Age	Income
50	200

$$Dis(x_1, x_2) = \sqrt{\Sigma_{i=0}^{n}(x_{1i} - x_{2i})^2} = \sqrt{(54-50)^2 + (190-200)^2}\text{=10.77}$$

What is the best value of K of KNN?

- Small K (Ex. K = 1):
  - It capture noise (in this case).
  - It doesn’t work with out-of-sample data.
- Large K(can be over generalized)



General solution for testing the accuracy

- Reserve data for testing the accuracy of the model.
- Choose K = 1, and then use the training part for modeling. Then calculate the accuracy of the prediction using all sample in the test set.
- Repeat the process by increasing K.
- Then you compare which K is the best.