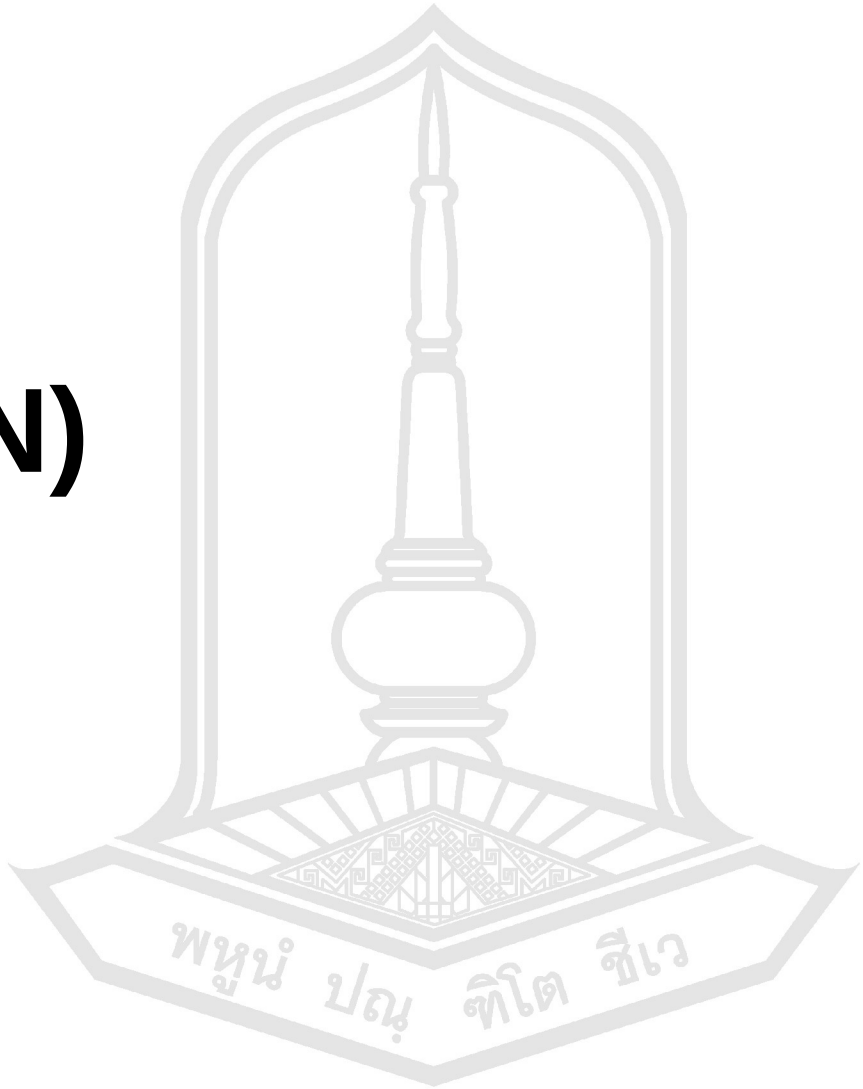
MACHINE LEARNING

1211635

MAHASARAKHAM
UNIVERSITY

K-NEAREST NEIGHBOR (KNN)

Supervised learning algorithm



Olarik Surinta, PhD.
Lecturer

What is KNN

- A very simple classification and regression algorithm
  - In case of classification, new data points get classified in a particular class
  - In case of regression, new data gets labeled based on the average value of k nearest neighbor
- It is a lazy learner because it doesn’t learn much from the training data
- It is a supervised learning algorithm
- Default method is Euclidean distance

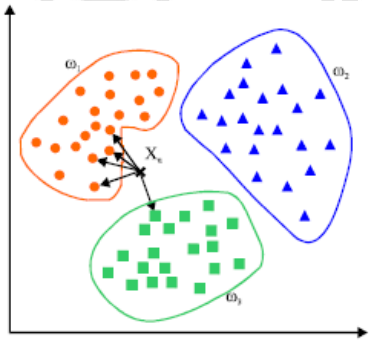
KNN

- KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).
- KNN has been used in statistical estimation and pattern recognition already in the beginning of 1971's as a non-parametric technique.

Cr. http://www.saedsayad.com/k_nearest_neighbors.htm

algorithm

- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a **distance function**.
- If K=1, then the case is simply assigned to the class of its nearest neighbor.



Distance functions

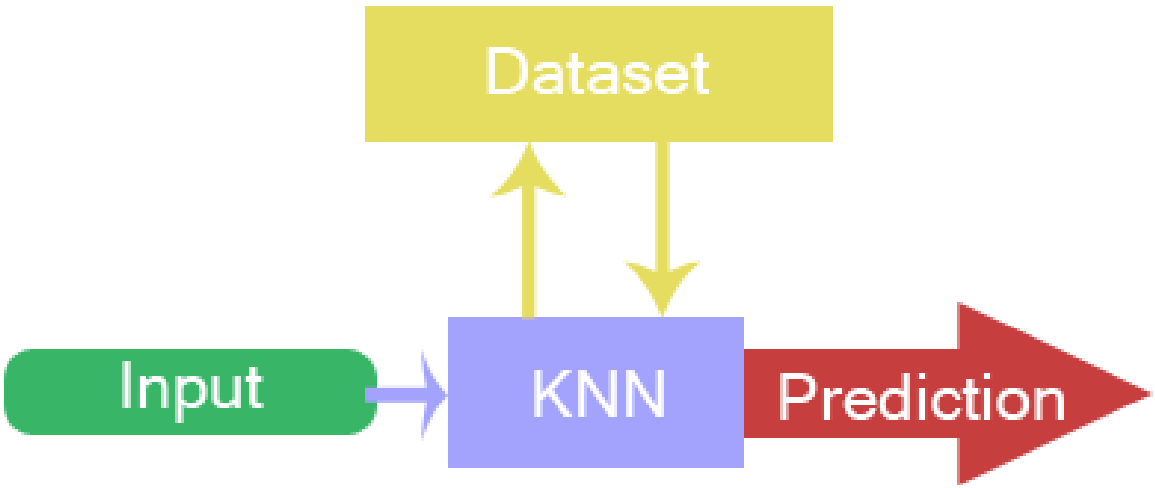
Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

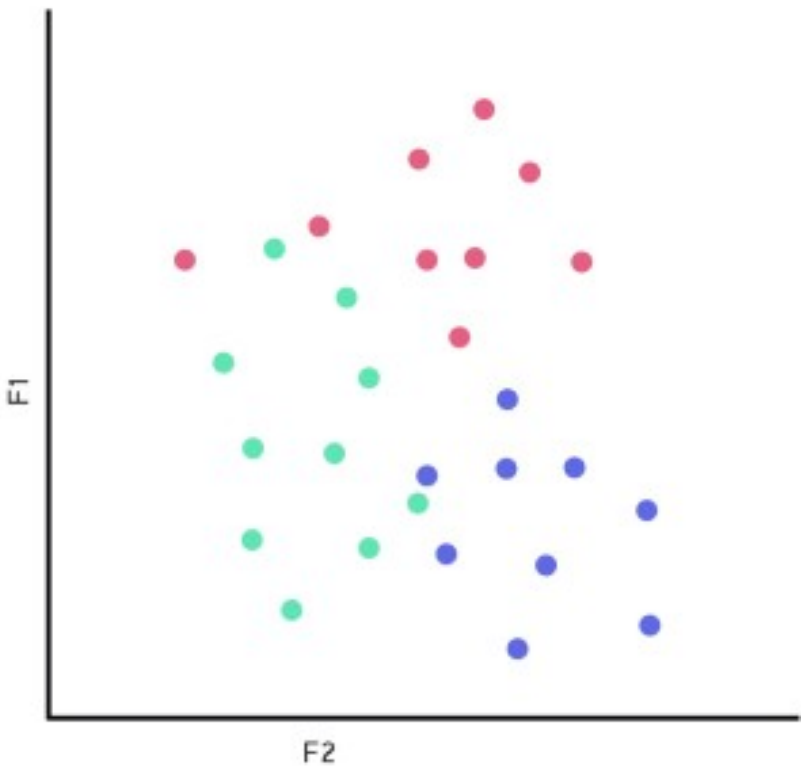
Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

KNN

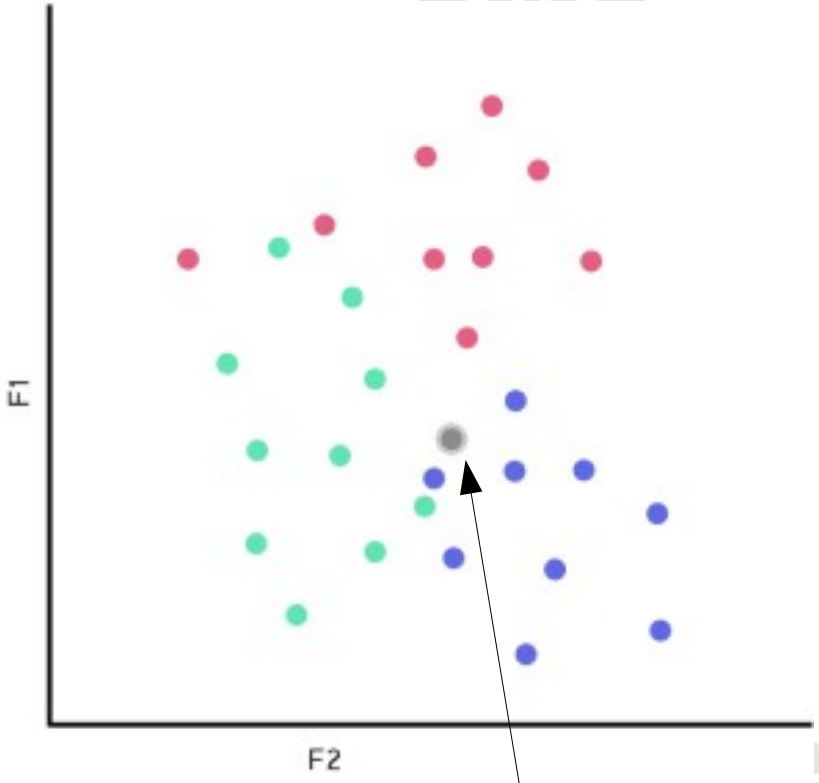


Cr. https://depiesml.wordpress.com/2015/09/03/learn-by-implementation-k-nearest-neighbor/

KNN

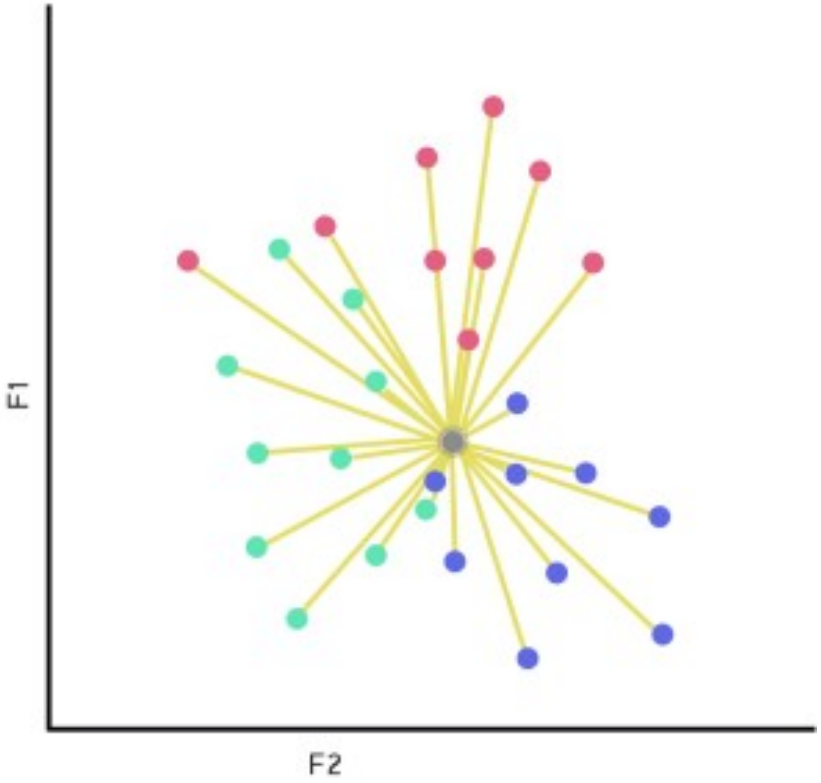


Dataset, 3 classes

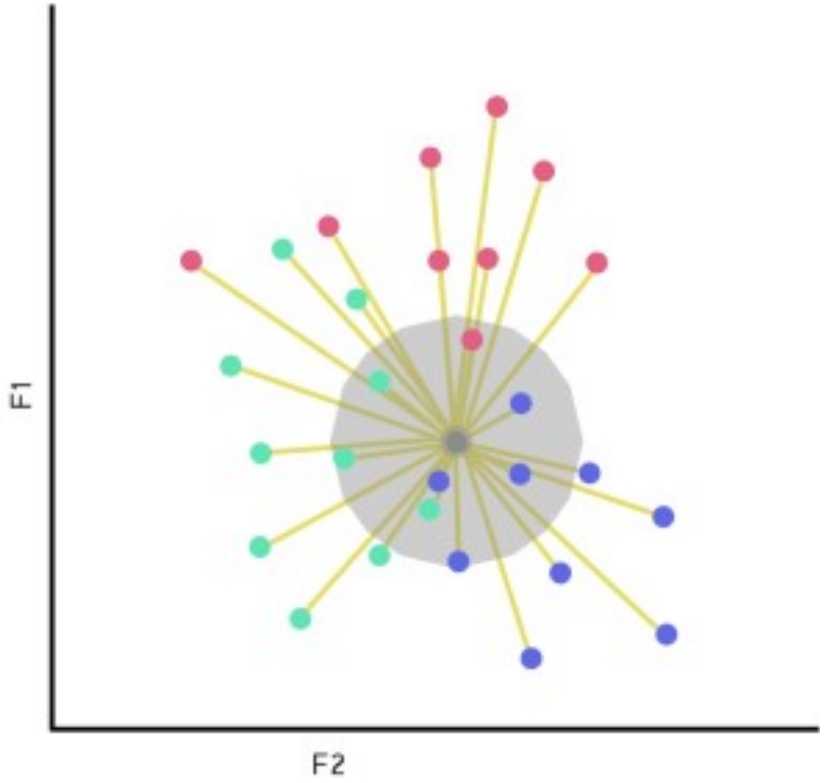


New input data

KNN

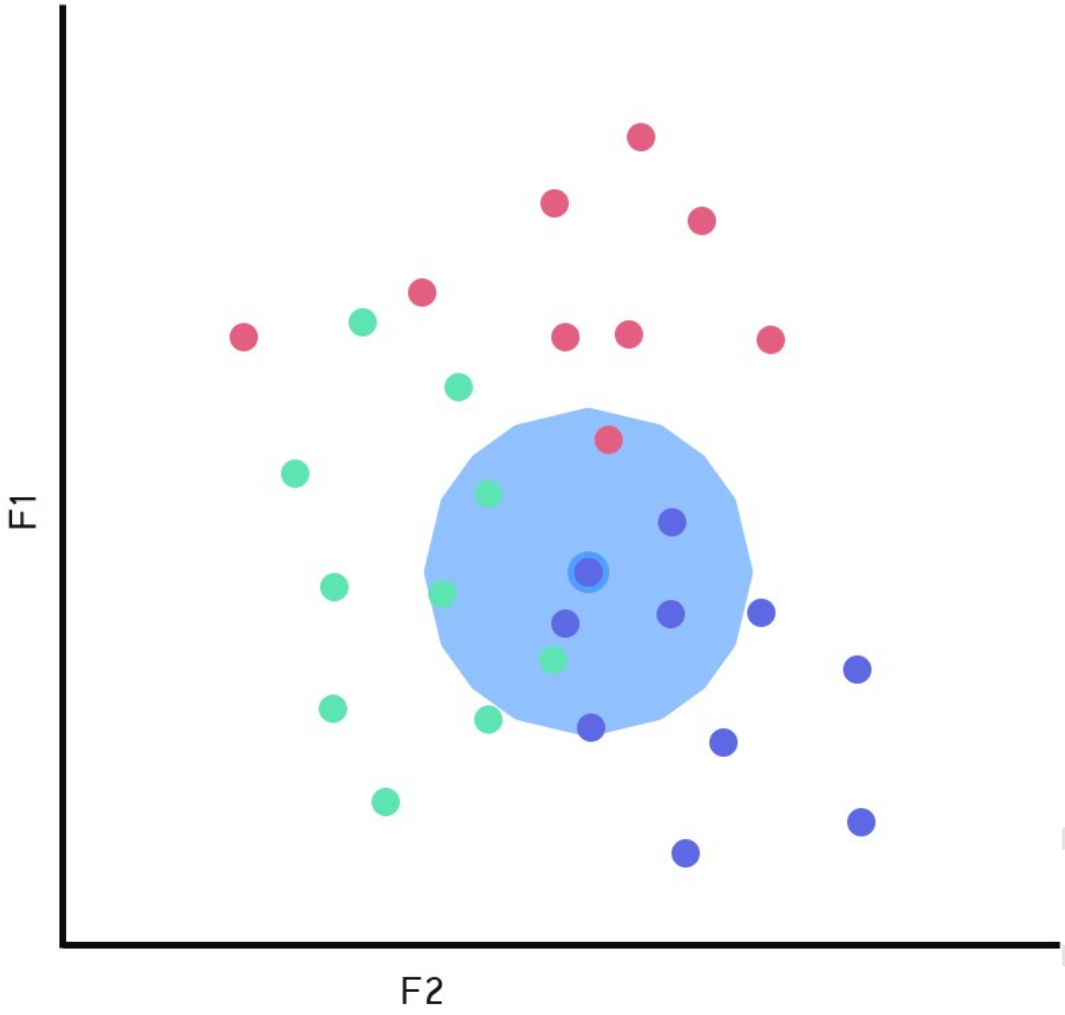


Finding a distance between new input data and all training data



Select number of neighborhood, such as K=1, 3, 5, etc.

KNN, K=7



Using a majority vote and then assigned a class to the new input data

KNN

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	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

What is the output If K=1, K=3, and K=5?

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

References

- http://www.saedsayad.com/k_nearest_neighbors.htm
- https://depiesml.wordpress.com/2015/09/03/learn-by-implementation-k-nearest-neighbor/