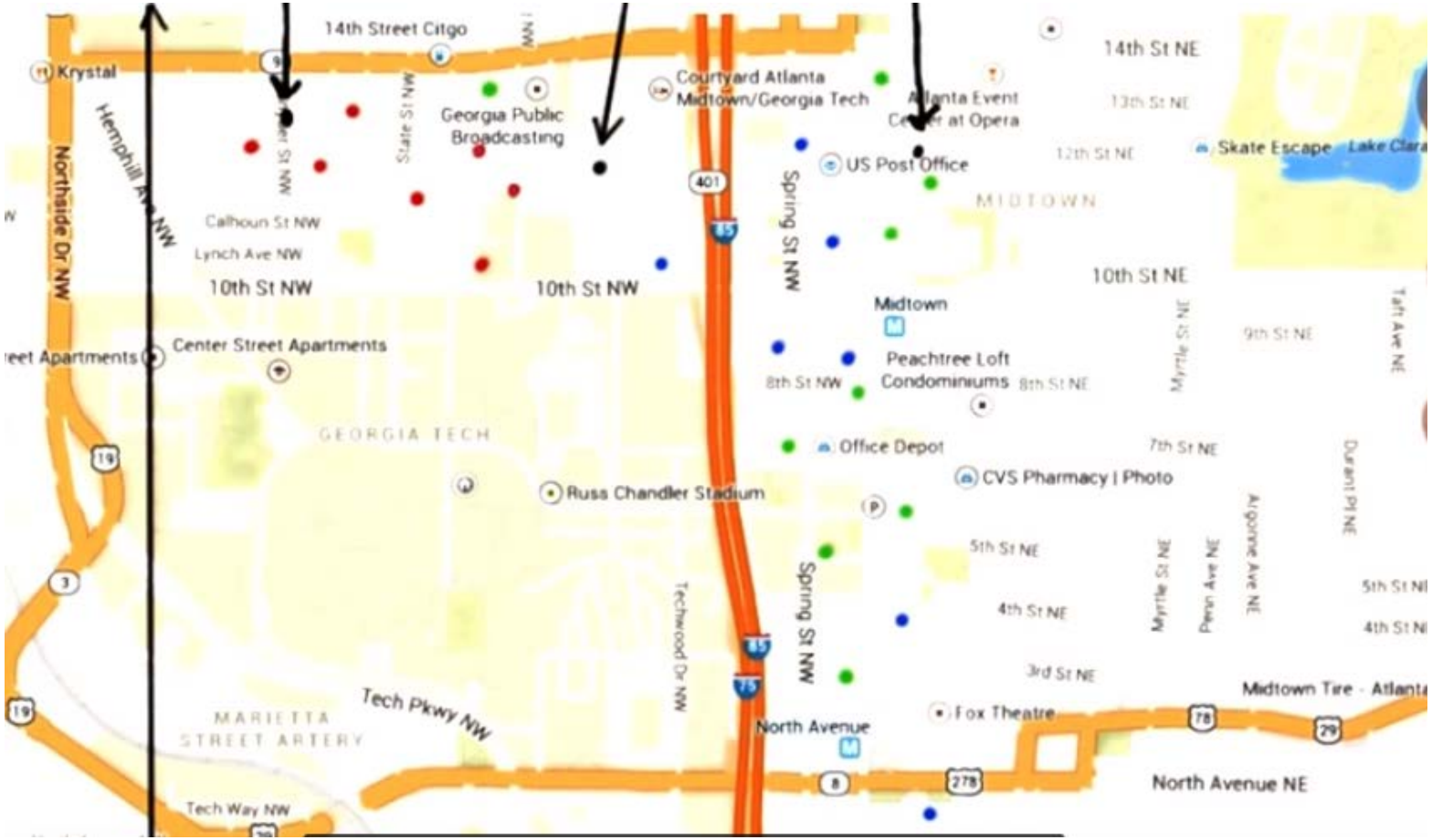
K- Nearest Neighbors

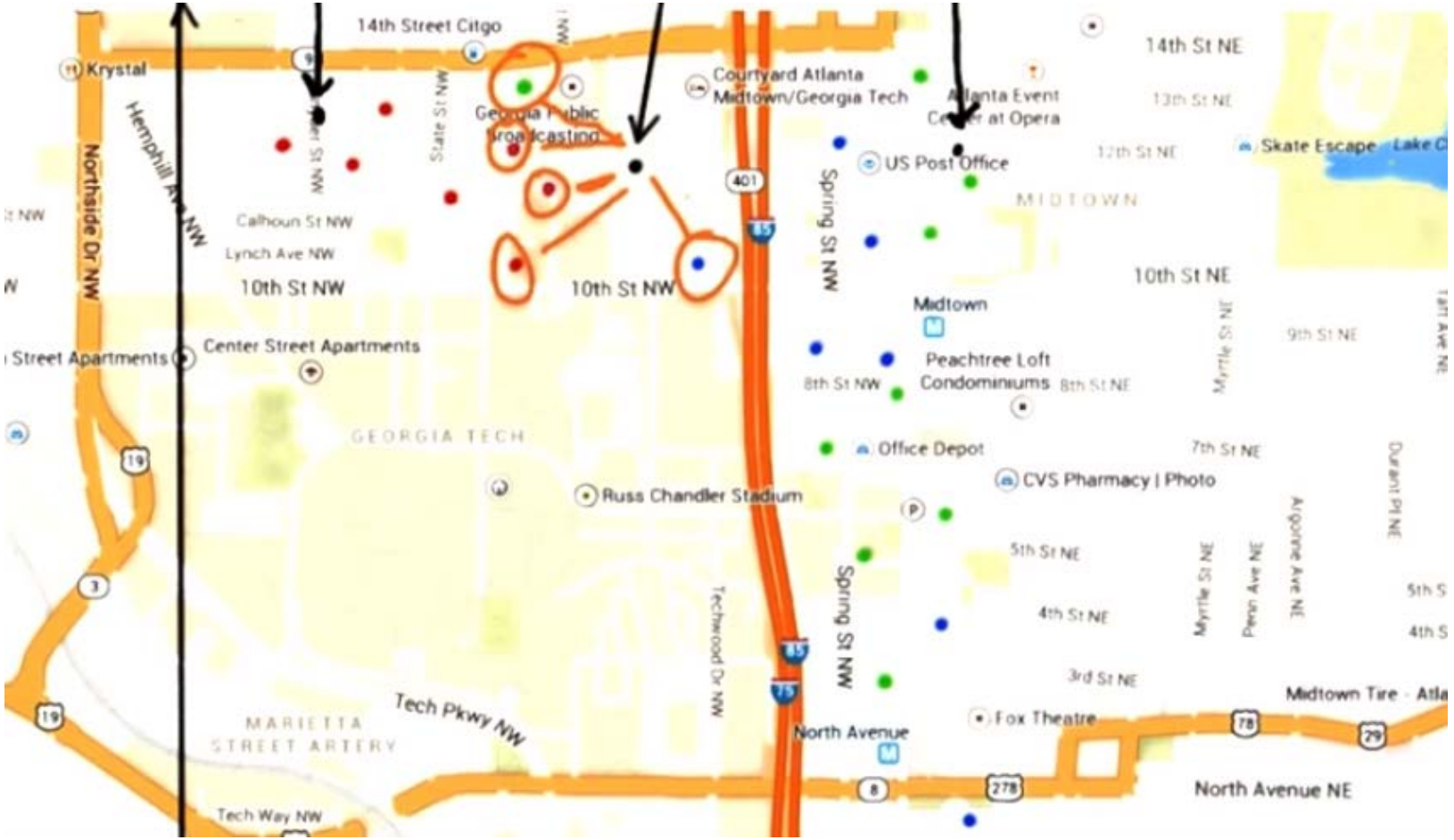
KNN

- K Nearest Neighbors is a simple **classification** algorithm
- 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 1970's as a non-parametric technique.
- It is best shown through example!

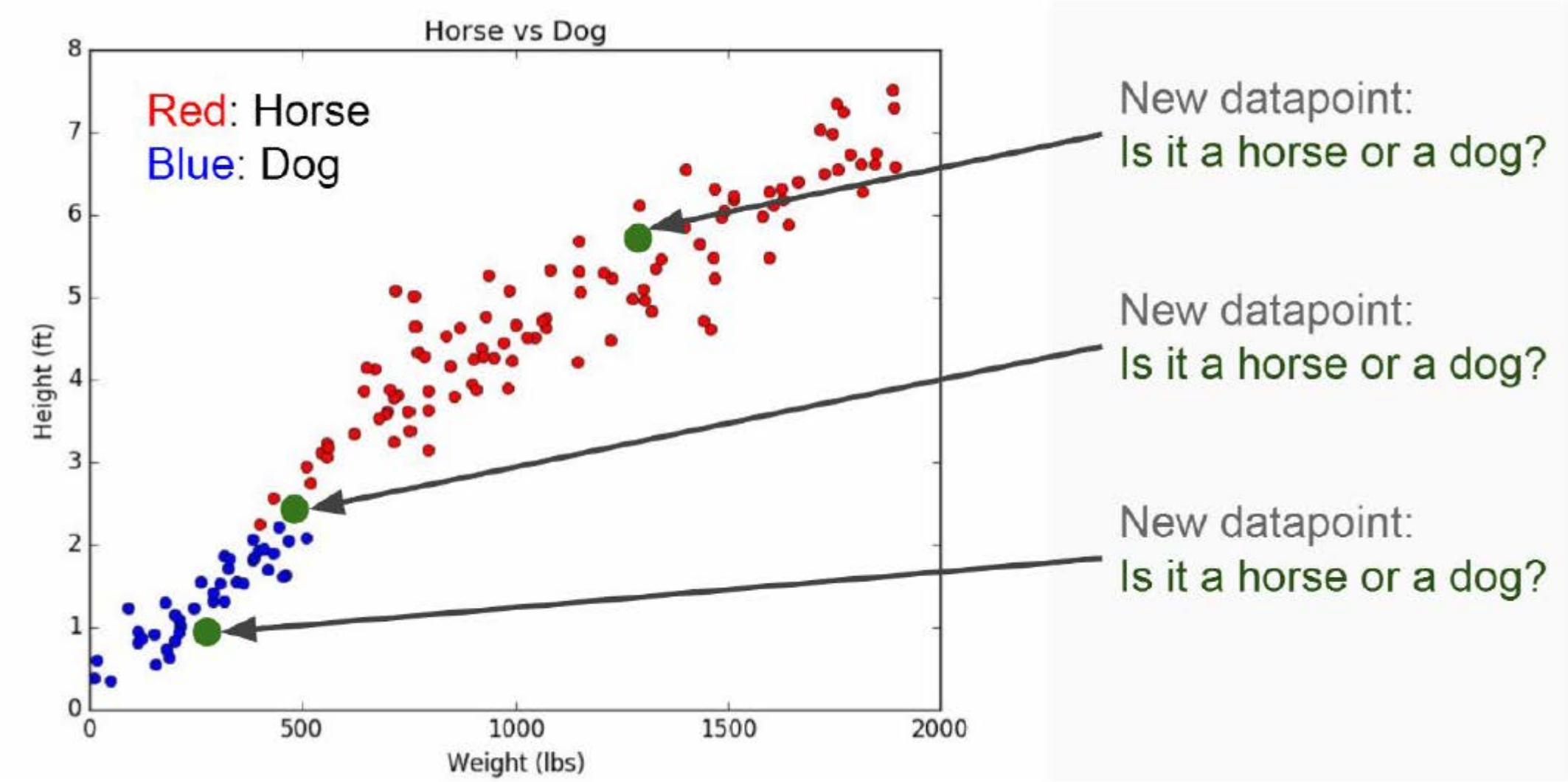
Let’s check the price of our neighbour…



Let’s check the price of our neighbourS…

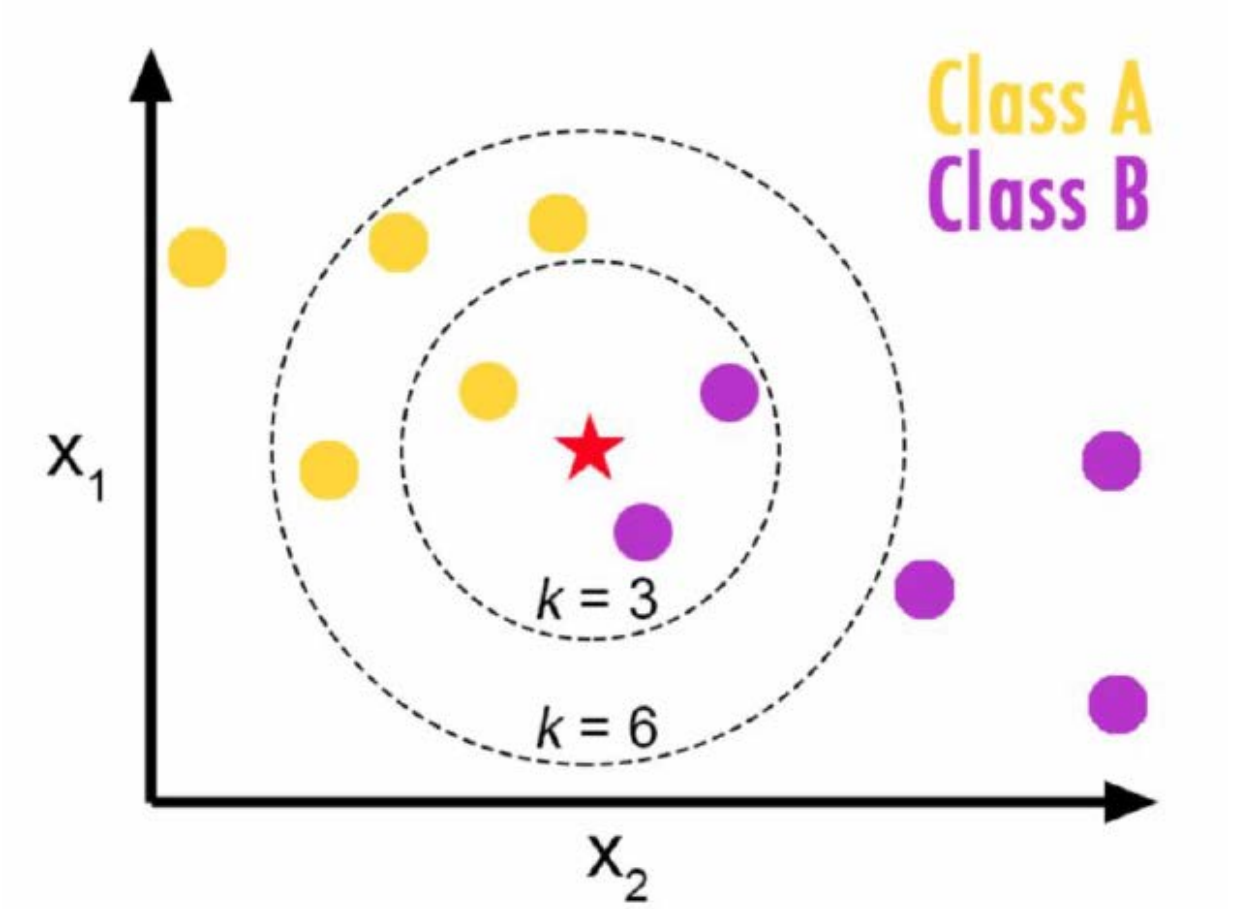


KNN – Another Example



Choosing K ?

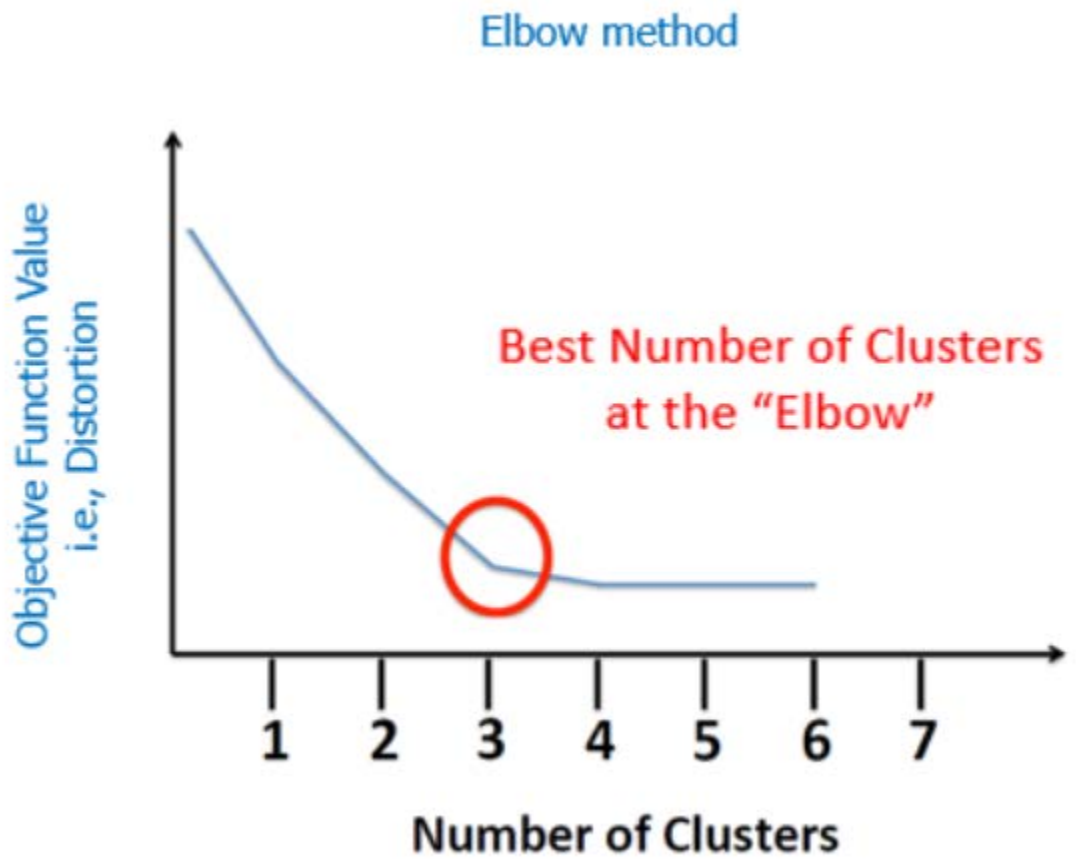
Choosing the K will effect what class the new point will belong to.



Choosing the K vlaue

- Choice of k is very critical – A small value of k means that noise will have a higher influence on the result. A large value make it computationally expensive and kinda defeats the basic philosophy behind KNN (that points that are near might have similar densities or classes) .A simple approach to select k is set k = n^(1/2).
- N – no. of features

Elbow Method



Distance Metrics

Distance functions

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

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

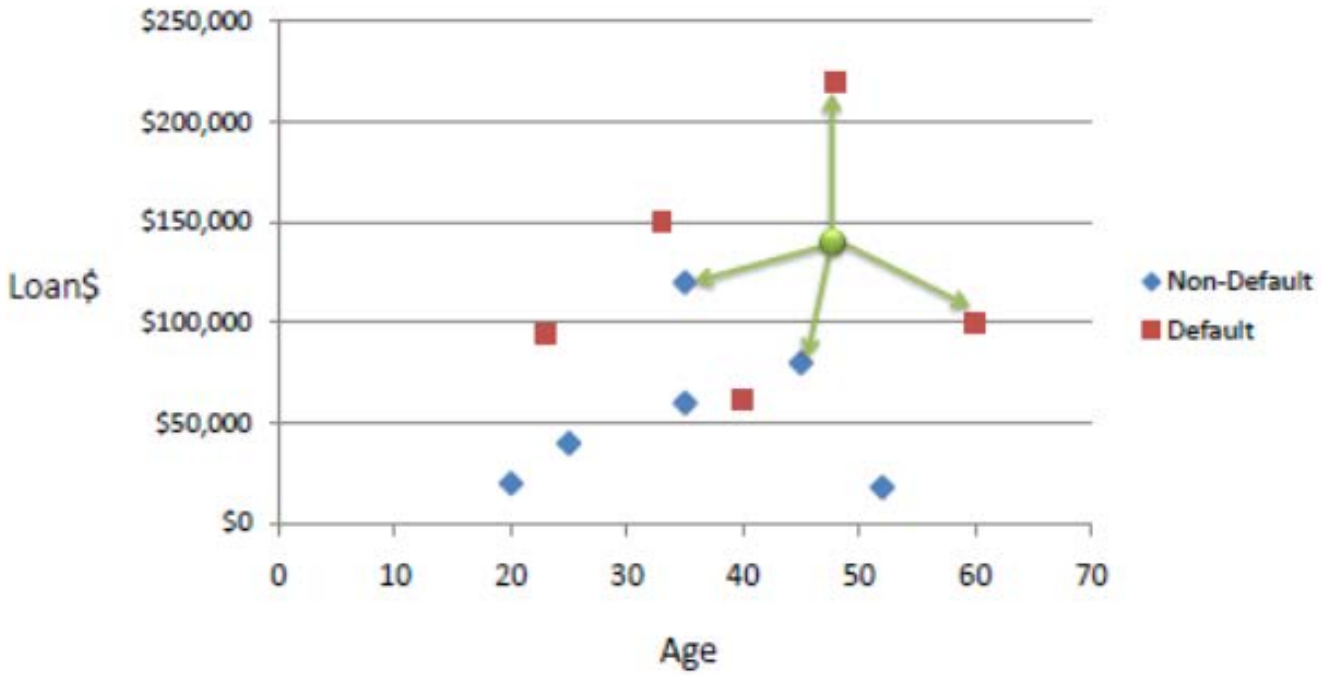
Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

Refer - [Diff between Euclidean and Minkowski]

Example

Consider the following data concerning credit default.

Age and Loan are two numerical variables (predictors) and Default is the target.



Example

- We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance. If K=1 then the nearest neighbor is the last case in the training set with Default=Y.

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

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	?		

Euclidean Distance

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

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

Pros

- Very simple
- Training is trivial
- Works with any number of classes
- Easy to add more data
- Few parameters
    - K
    - Distance Metric

Cons

- High Prediction Cost (worse for large data sets)
- Not good with high dimensional data
- Categorical Features don’t work well

Is KNN Supervised?

- It warrants noting that **kNN** is a "**supervised**" classification method in that it uses the class labels of the training data.
- **Unsupervised** classification methods, or "clustering" methods, on the other hand, do not employ the class labels of the training data.
- It is the most fundamental classification method

Curse of Dimensionality - Terminology

- As the number of **Features** or **Dimensions grows**, the amount of data we need to **generalize** accurately grows exponentially.