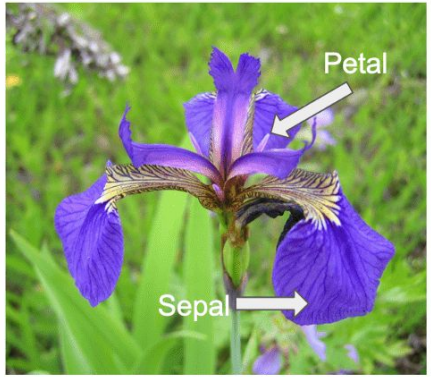
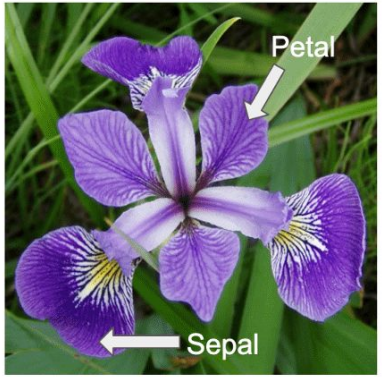
Supervised Learning: PART 1: KNN

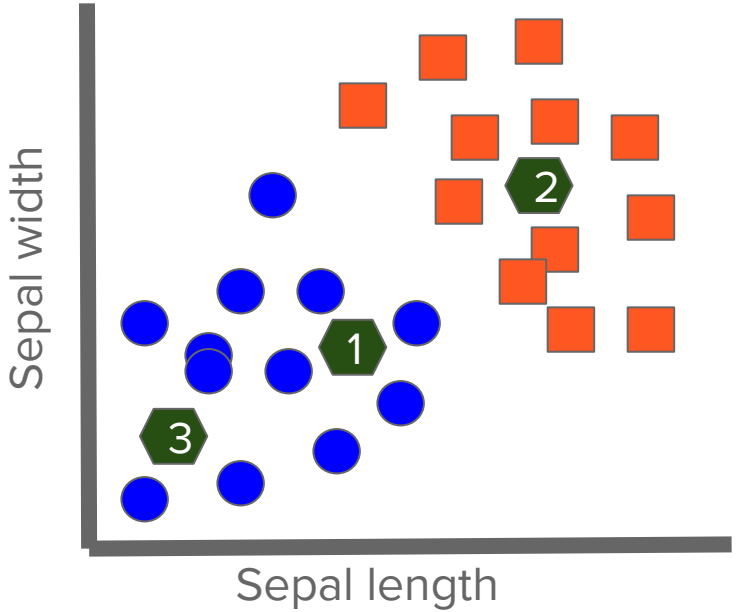
Mahesh Mohan M R
Centre of Excellence in AI
Indian Institute of Technology Kharagpur

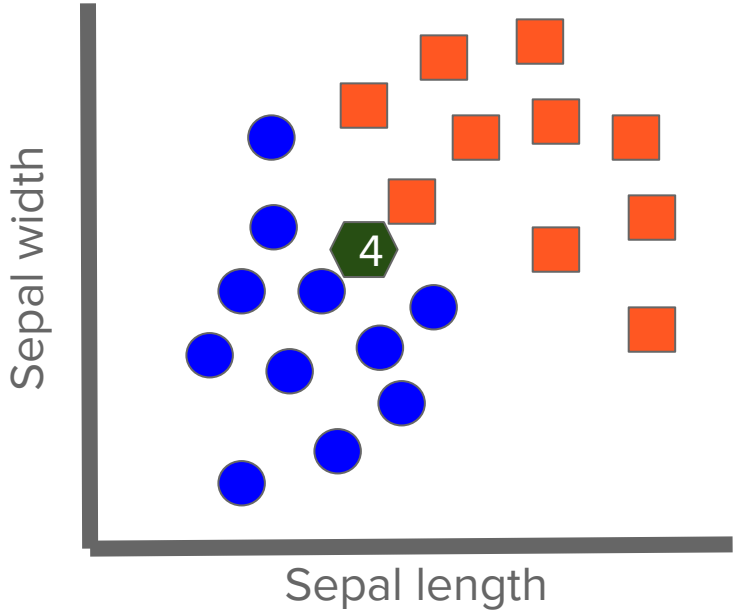
Iris setosa



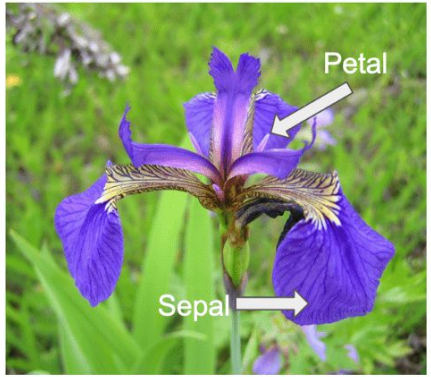
Iris versicolor



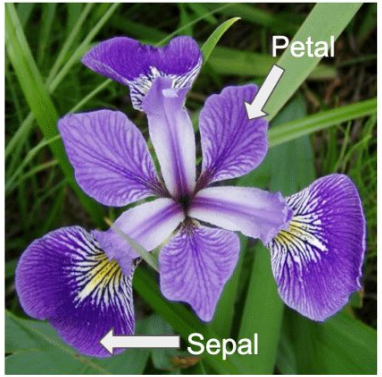


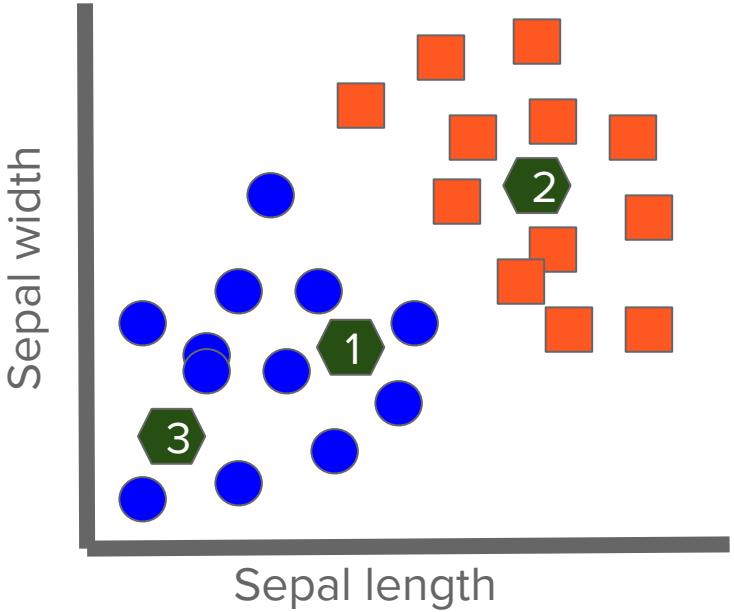


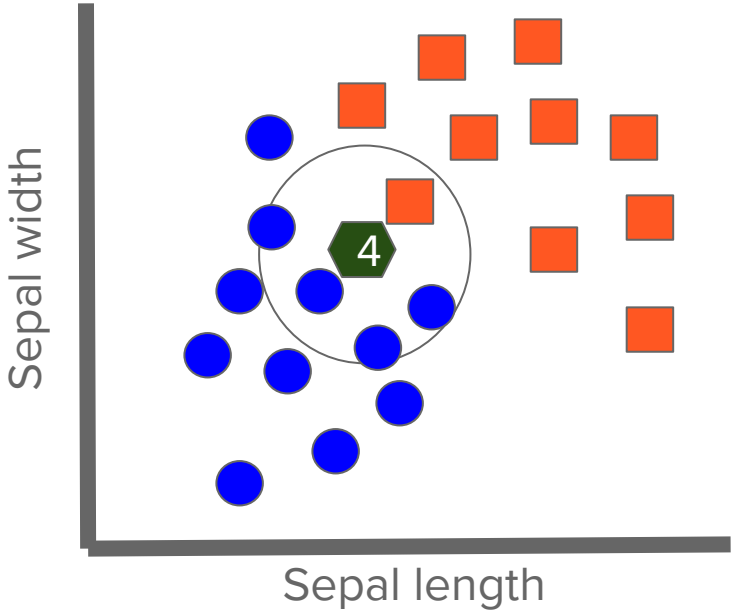
Iris setosa



Iris versicolor

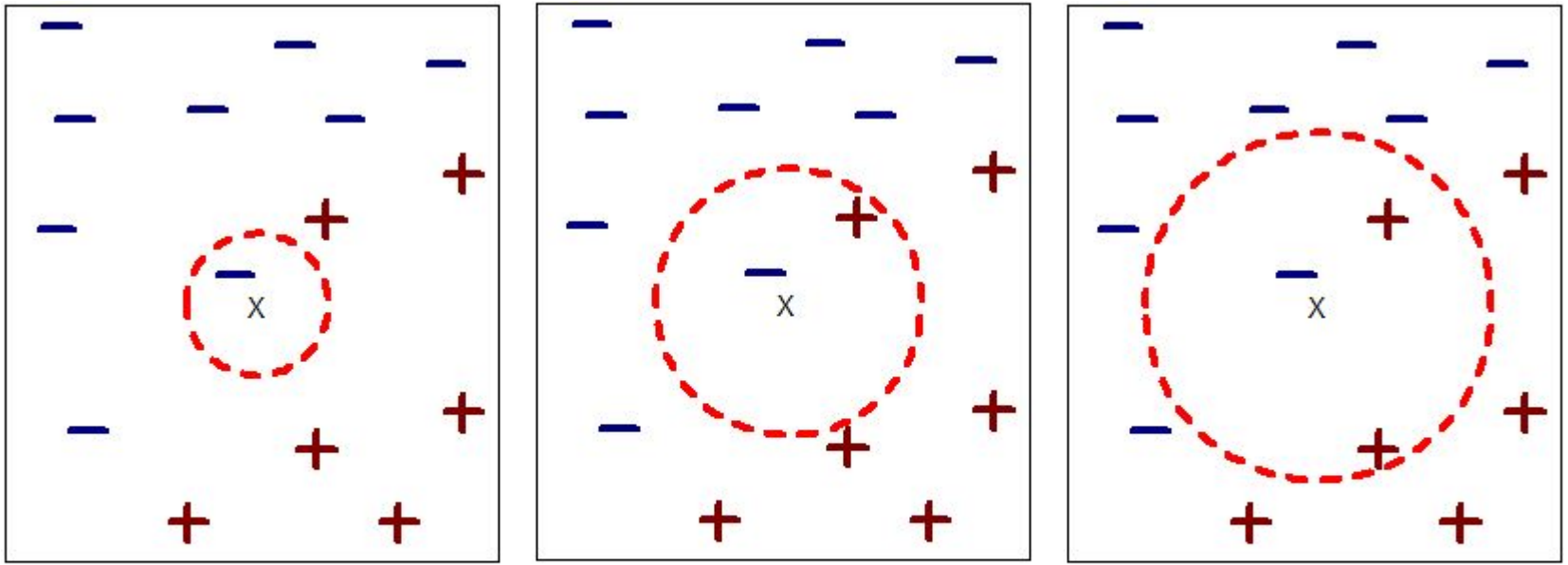






K Nearest Neighbour (KNN) Classifier

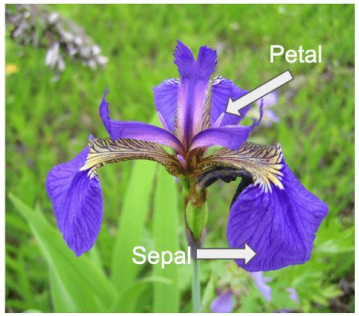
KNN Rule: Assign to a test sample the majority category label of its *k* nearest training samples



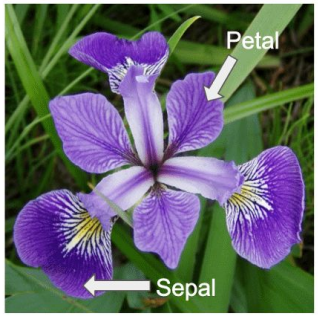
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

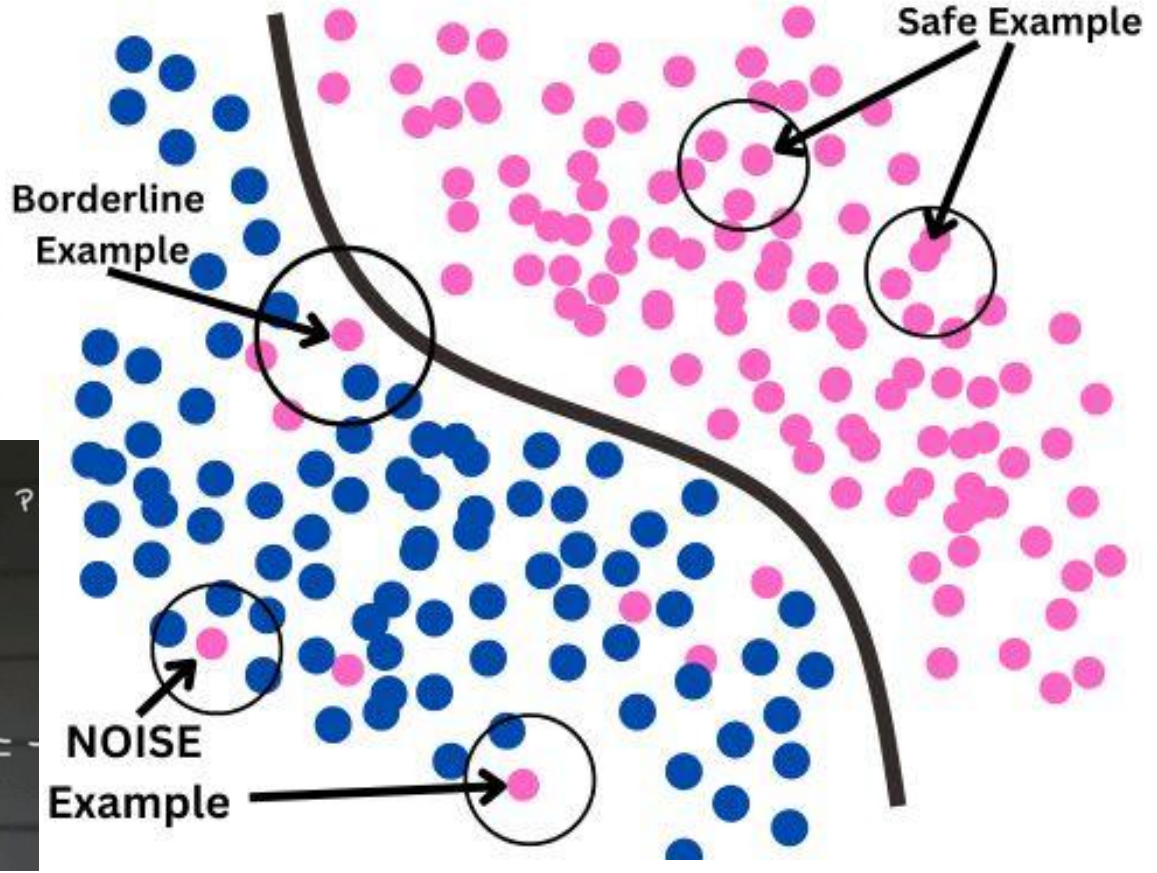
Noise in Datasets

Iris setosa



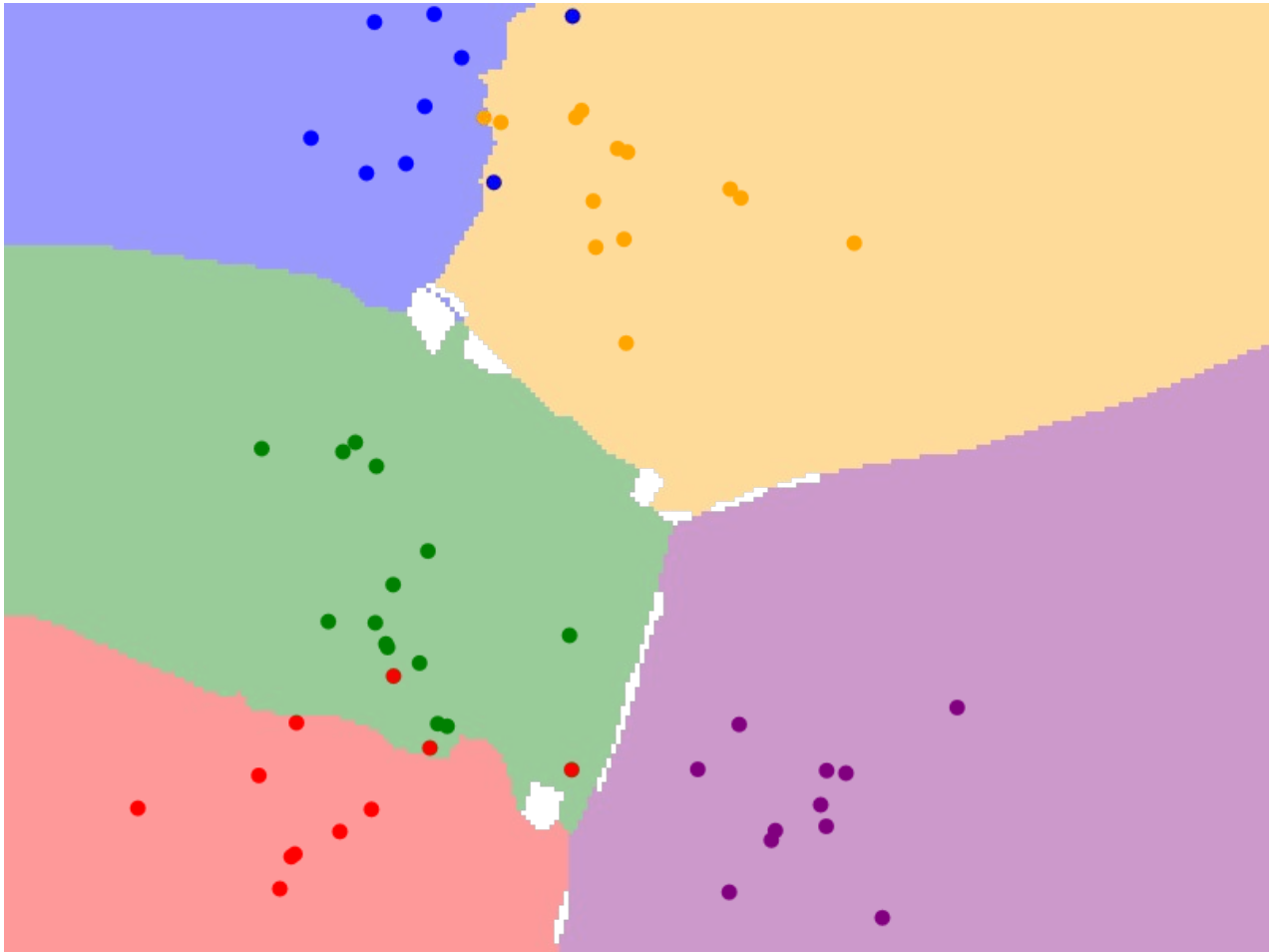
Iris versicolor



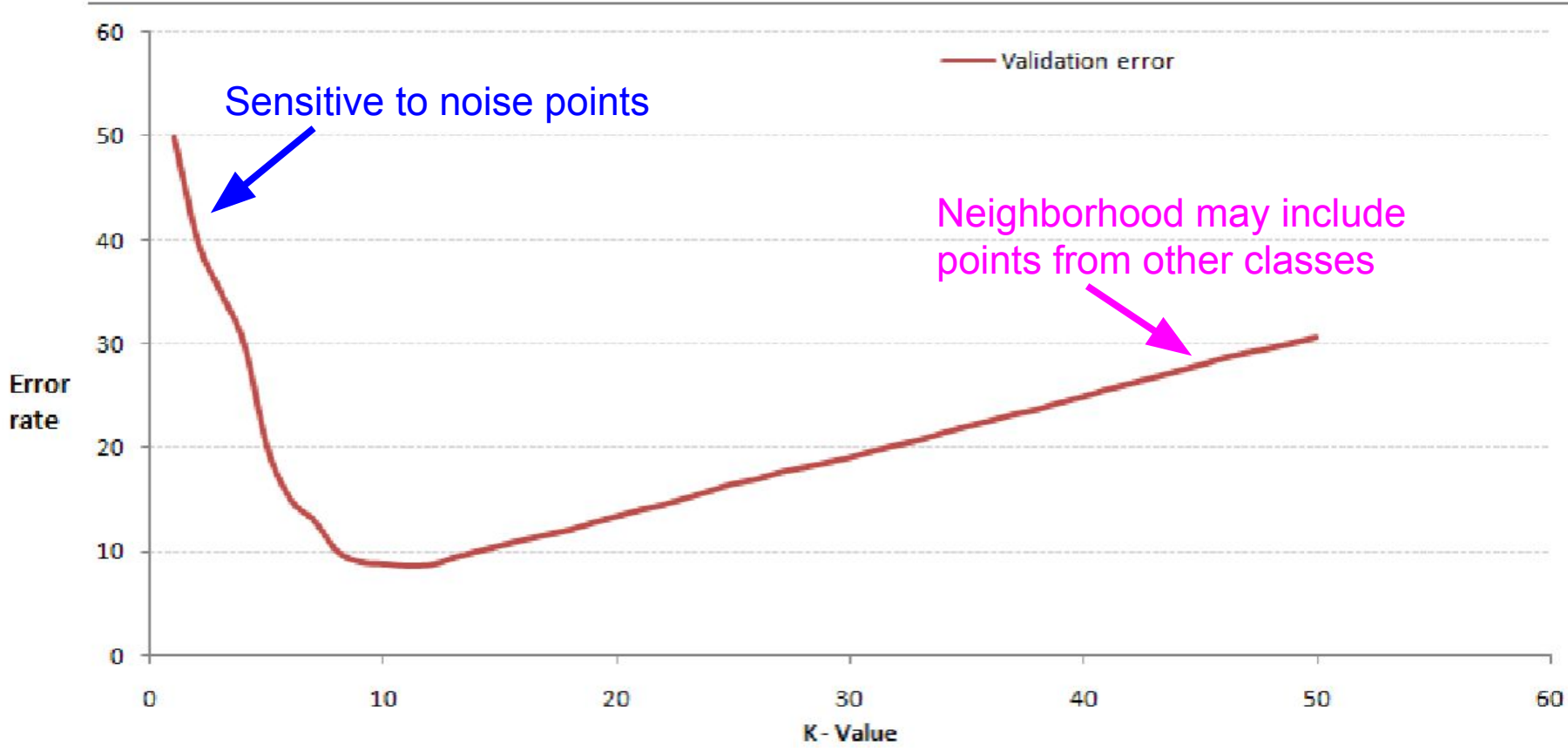


Voronoi Diagram

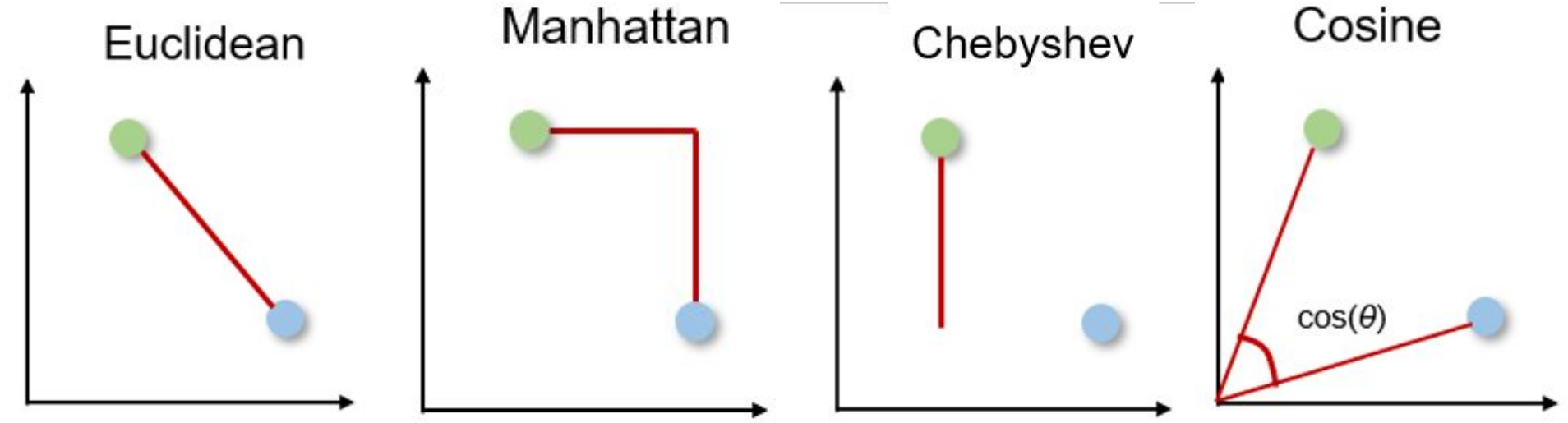
http://vision.stanford.edu/teaching/cs231n-demos/knn/

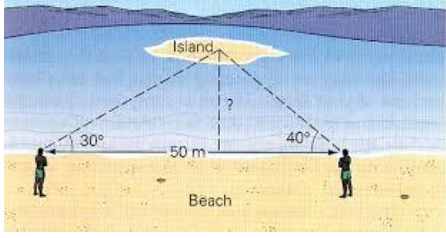


Optimal K

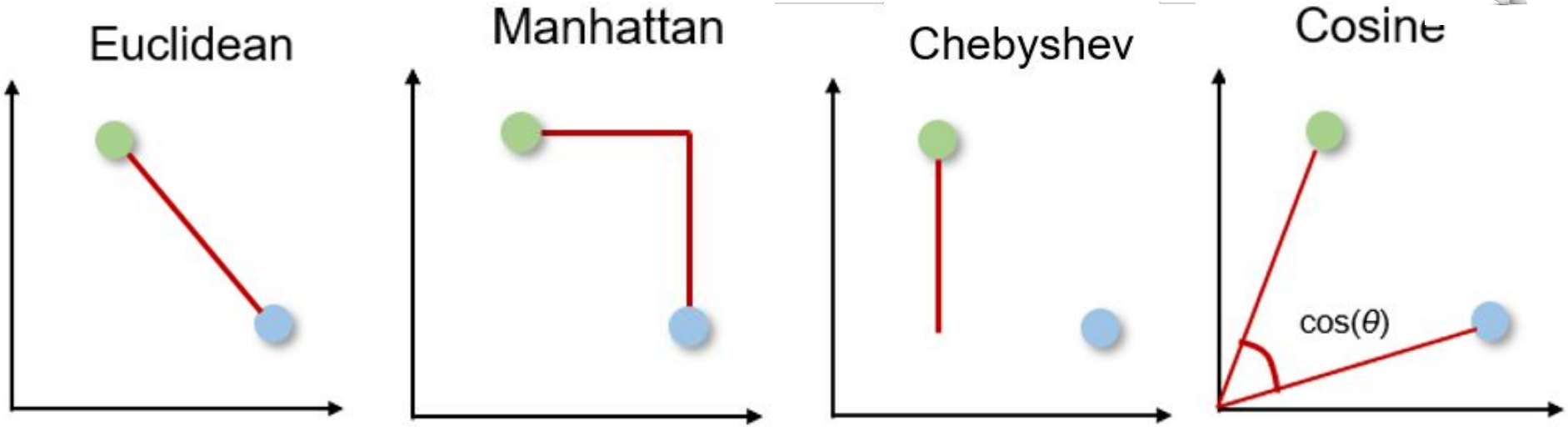

Validation error
Sensitive to noise points
Neighborhood may include points from other classes
Error rate
60
50
40
30
20
10
0
0
10
20
30
40
50
60
K - Value

Distance Measures





Distance Measures



Jumbled?

$$d = \max_i \left(|x_i - y_i| \right)$$

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

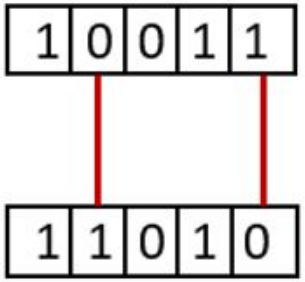
$$d = \frac{< x\, y >}{\|x\| \, \|y\|}$$

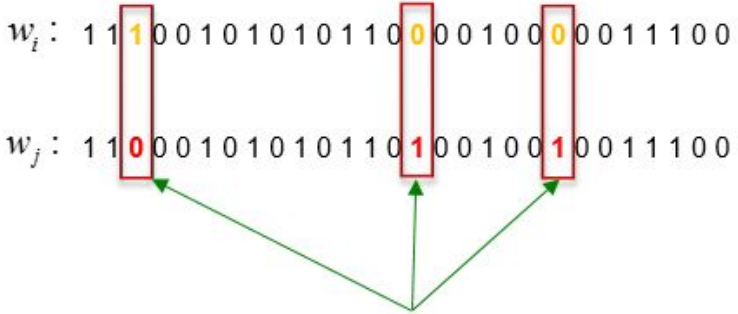
$$d = \sum_{i=1}^{n} |(x_i - y_i)|$$

Distance Measures for Binary Data

Movies						Target
Parasite	Joker	Avengers	Spotlight	The Great Beauty	There will be blood	Rating
1	0	0	0	0	0	5
0	1	0	0	0	0	4
0	0	1	0	0	0	4
1	0	0	0	1	0	2
0	0	0	1	0	0	4
0	0	0	0	1	0	3
0	0	1	0	0	0	5
0	0	0	0	0	1	4
0	0	1	0	0	0	4

Hamming





The number of bits that are different between w_i and $w_j = 3$

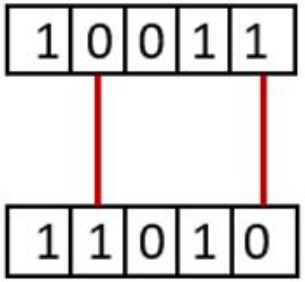
$$D(w_i, w_j) = 3$$

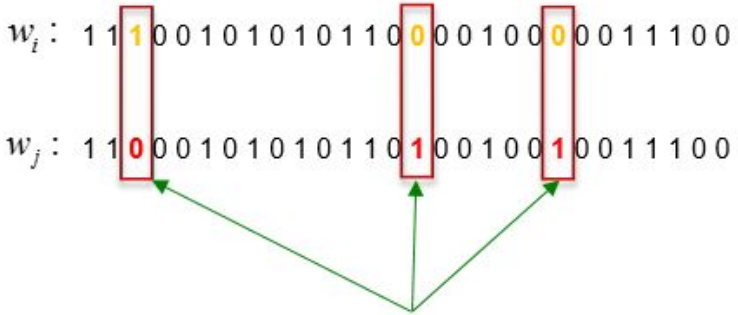
Distance Measures for Binary Data

Movies						Target
Parasite	Joker	Avengers	Spotlight	The Great Beauty	There will be blood	Rating
1	0	0	0	0	0	5
0	1	0	0	0	0	4
0	0	1	0	0	0	4
1	0	0	0	1	0	2
0	0	0	1	0	0	4
0	0	0	0	1	0	3
0	0	1	0	0	0	5
0	0	0	0	0	1	4
0	0	1	0	0	0	4

What if data is a combination of real and binary values?

Hamming

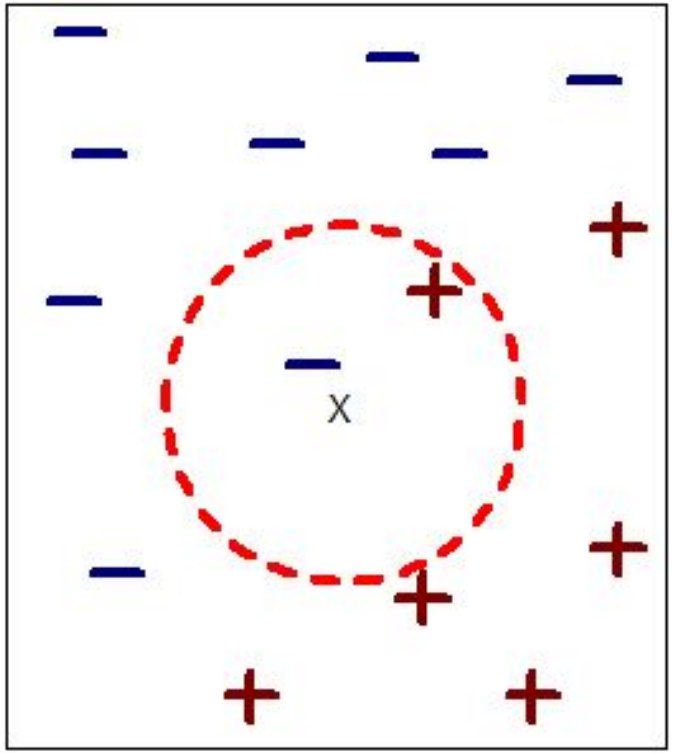




The number of bits that are different between w_i and $w_j = 3$

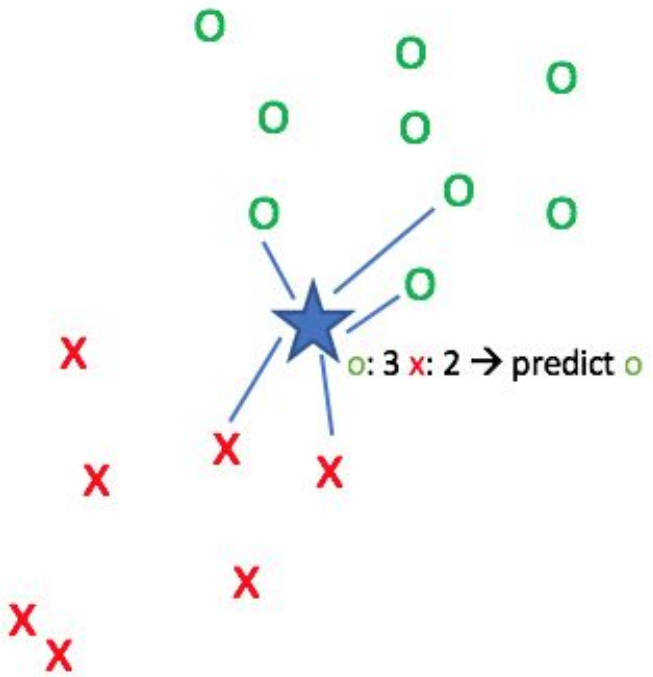
$$D(w_i, w_j) = 3$$

Distance-weighted KNN

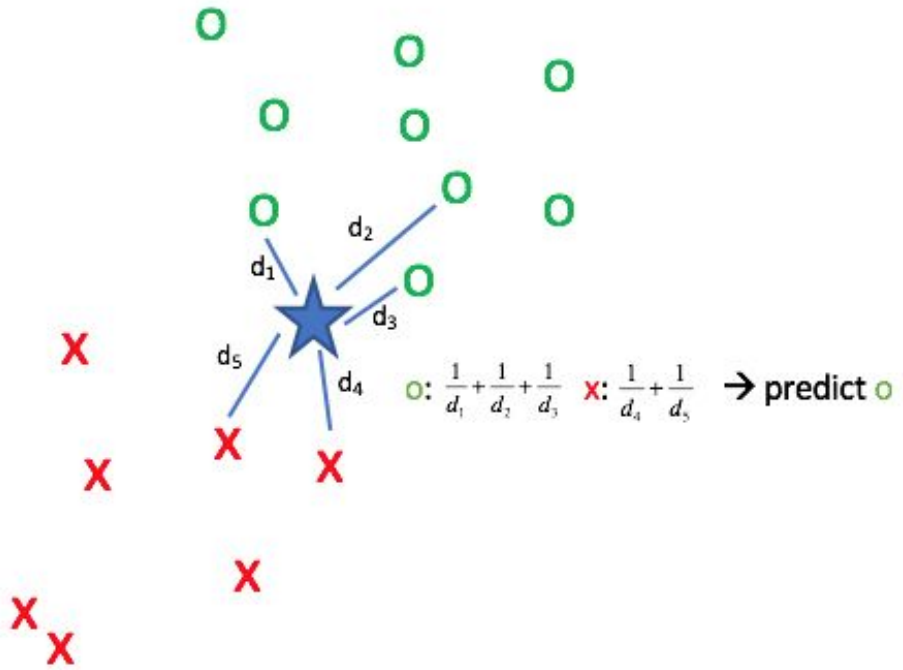


(b) 2-nearest neighbor

Distance-weighted KNN: Classification



Uniform K-NN (k=5)



Distance-weighted K-NN (k=5)

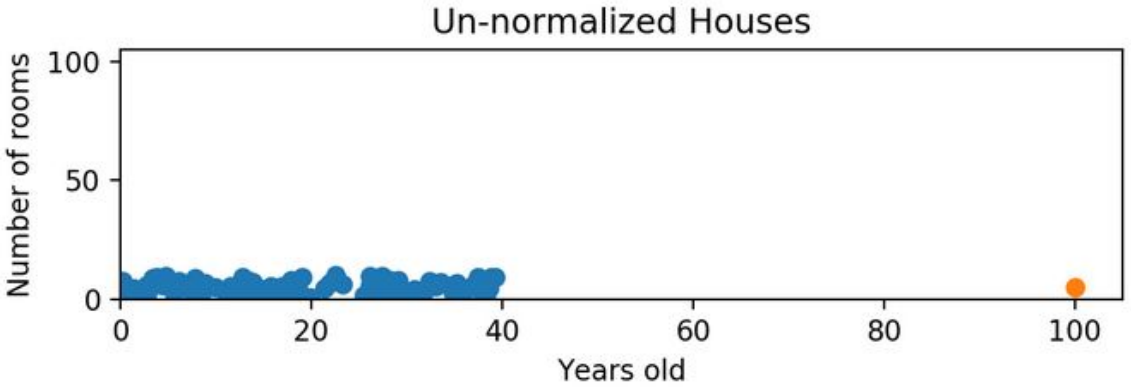
Problem of Measurement Scales

- Different features may have different measurement scales
  - E.g., patient weight in kg (range [50,200]) vs. blood protein values in ng/dL (range [-3,3])
- Consequences
  - Patient weight will have a much greater influence on the distance between samples
  - May bias the performance of the classifier

Solution for Measurement Scales: #1

Min-Max Normalization

$$\frac{value - min}{max - min}$$



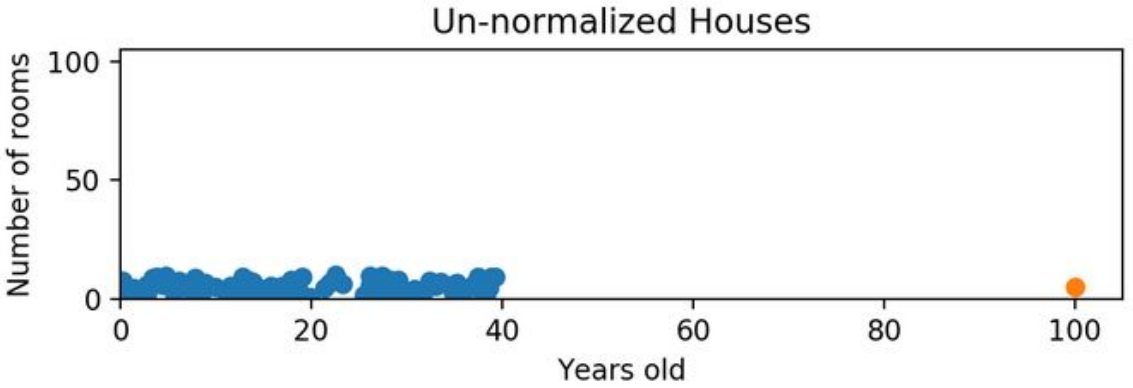
For every feature,
Minimum value – ?
Maximum value – ?
In-between Min and Max – ?

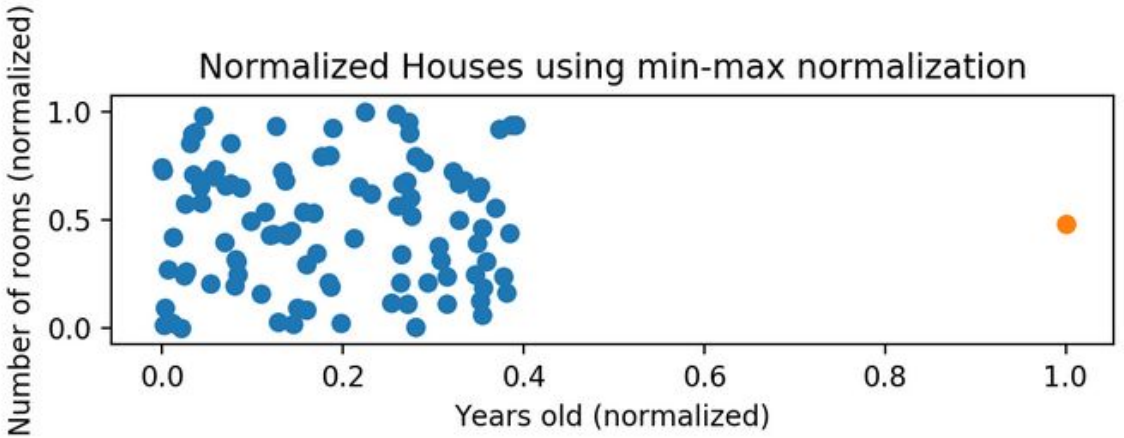
Solution for Measurement Scales: #1

Min-Max Normalization

$$\frac{value - min}{max - min}$$

For every feature,
Minimum value – 0
Maximum value – 1
In-between Min and Max – (0,1)



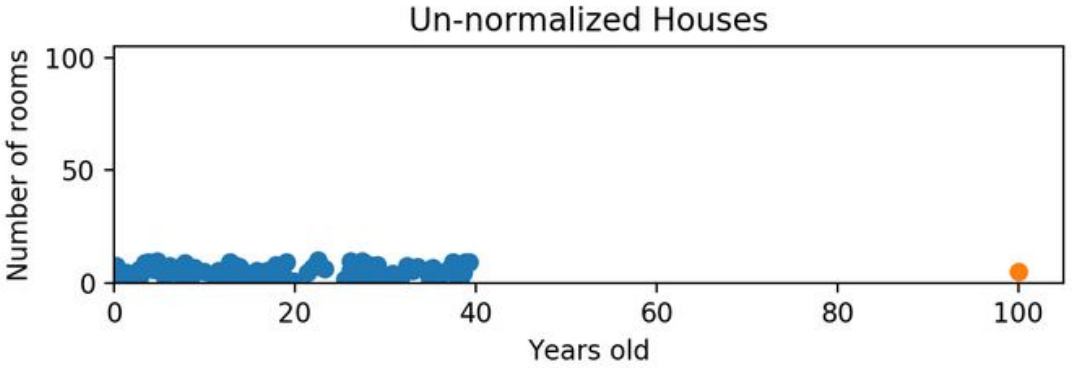


Solution for Measurement Scales: #2

Z-score Normalization

$$\frac{value - \mu}{\sigma}$$

μ - Mean of the training data
σ - Std deviation of the training data



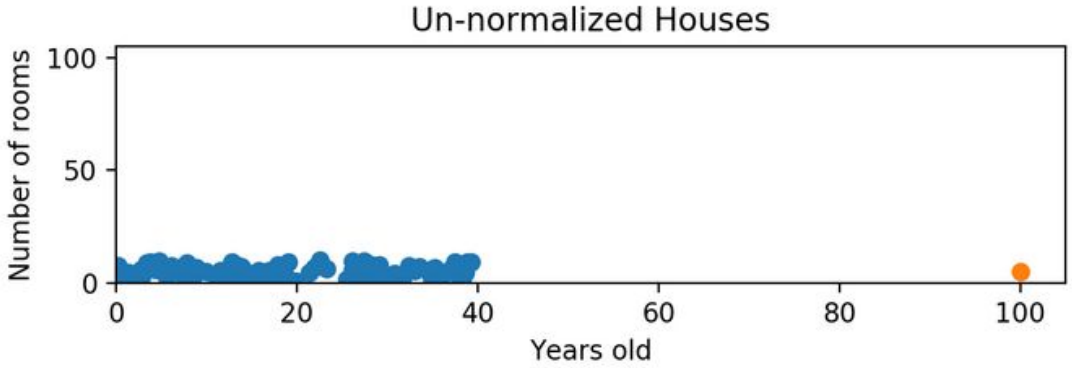
For every feature,
Mean value – ?
Below the mean – ?
Above the mean – ?
Range - ?

Solution for Measurement Scales: #2

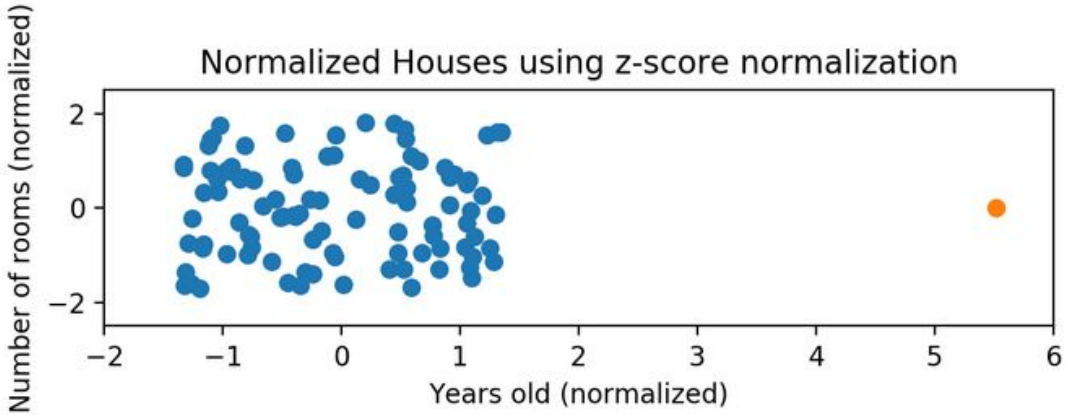
Z-score Normalization

$$\frac{value - \mu}{\sigma}$$

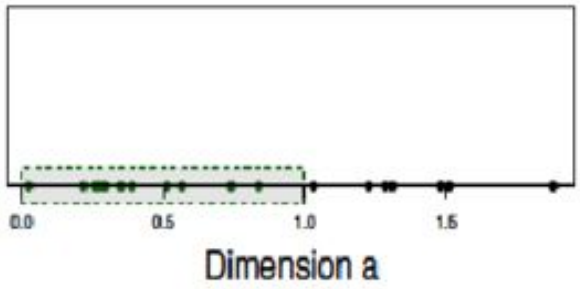
For every feature,
μ - Mean and ***σ*** - Std deviation of the data

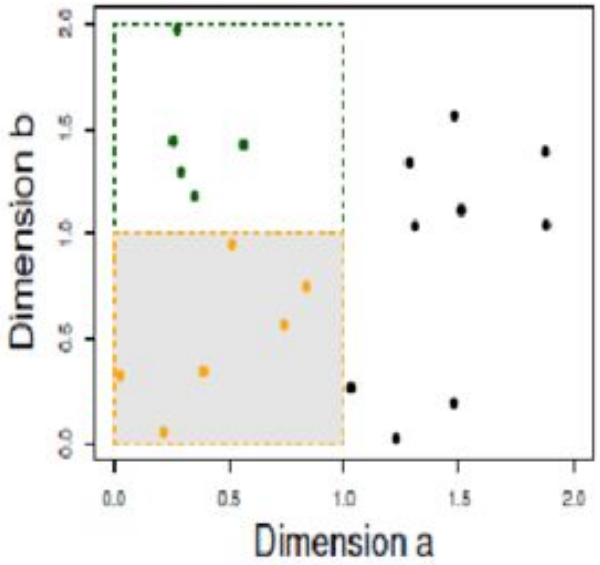


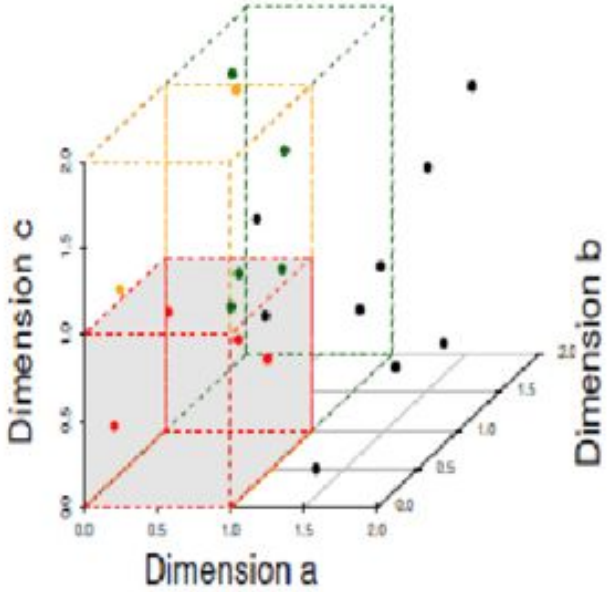
For every feature,
Mean value – 0
Below the mean – less than zero
Above the mean – more than zero



Drawback 1: Curse of Dimensionality







Drawbacks 2-3: Expensive and Storage Need