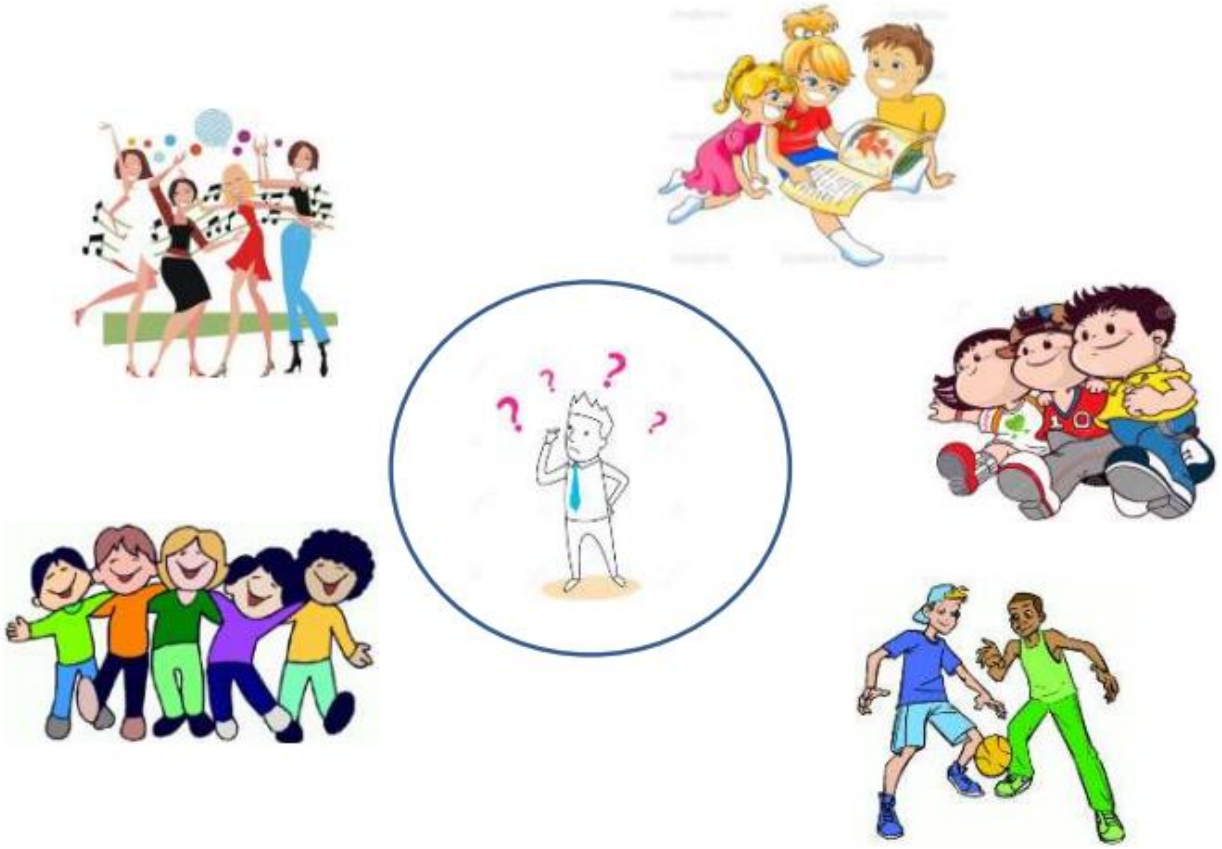
Microsoft Learn
Student Ambassadors

ML Bootcamp Session 3 : Supervised Learning – K Nearest Neighbours

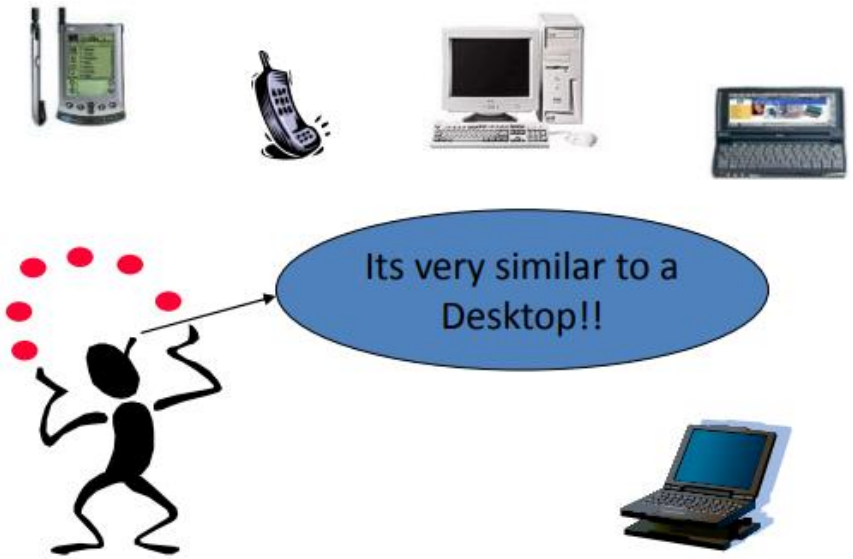
Example

- Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.



Example

Instance-based Learning

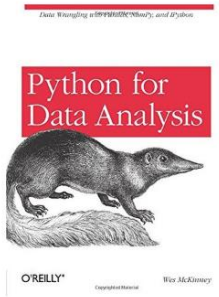


KNN Example

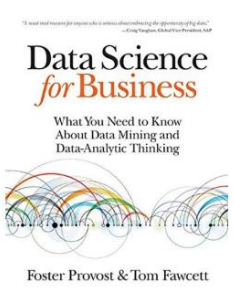
Customers Who Bought This Item Also Bought

Page 1 of 15

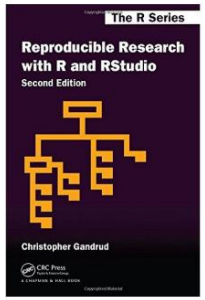
Data Science from Scratch: First Principles with Python
› Joel Grus
54
#1 Best Seller in Data Mining
Paperback
$33.99 Prime



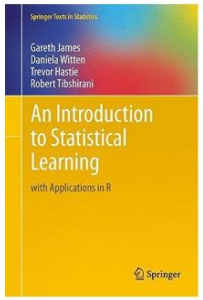
Python for Data Analysis: Data Wrangling with Pandas, NumPy, and…
› Wes McKinney
118
Paperback
$27.68 Prime



Data Science for Business: What You Need to Know about Data Mining and…
› Foster Provost
135
Paperback
$37.99 Prime

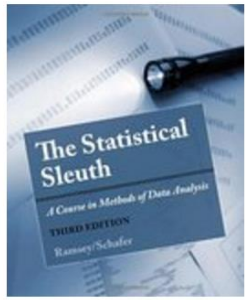


Reproducible Research with R and R Studio, Second Edition…
Christopher Gandrud
3
Paperback
$51.97 Prime



An Introduction to Statistical Learning: with Applications in R…
› Gareth James
105
Hardcover
$68.35 Prime

Data Smart: Using Data Science to Transform Information into Insight
› John W. Foreman
99
#1 Best Seller in Computer Simulation
Paperback
$28.16 Prime



The Statistical Sleuth: A Course in Methods of Data Analysis
Fred Ramsey
6
Hardcover
$284.42 Prime

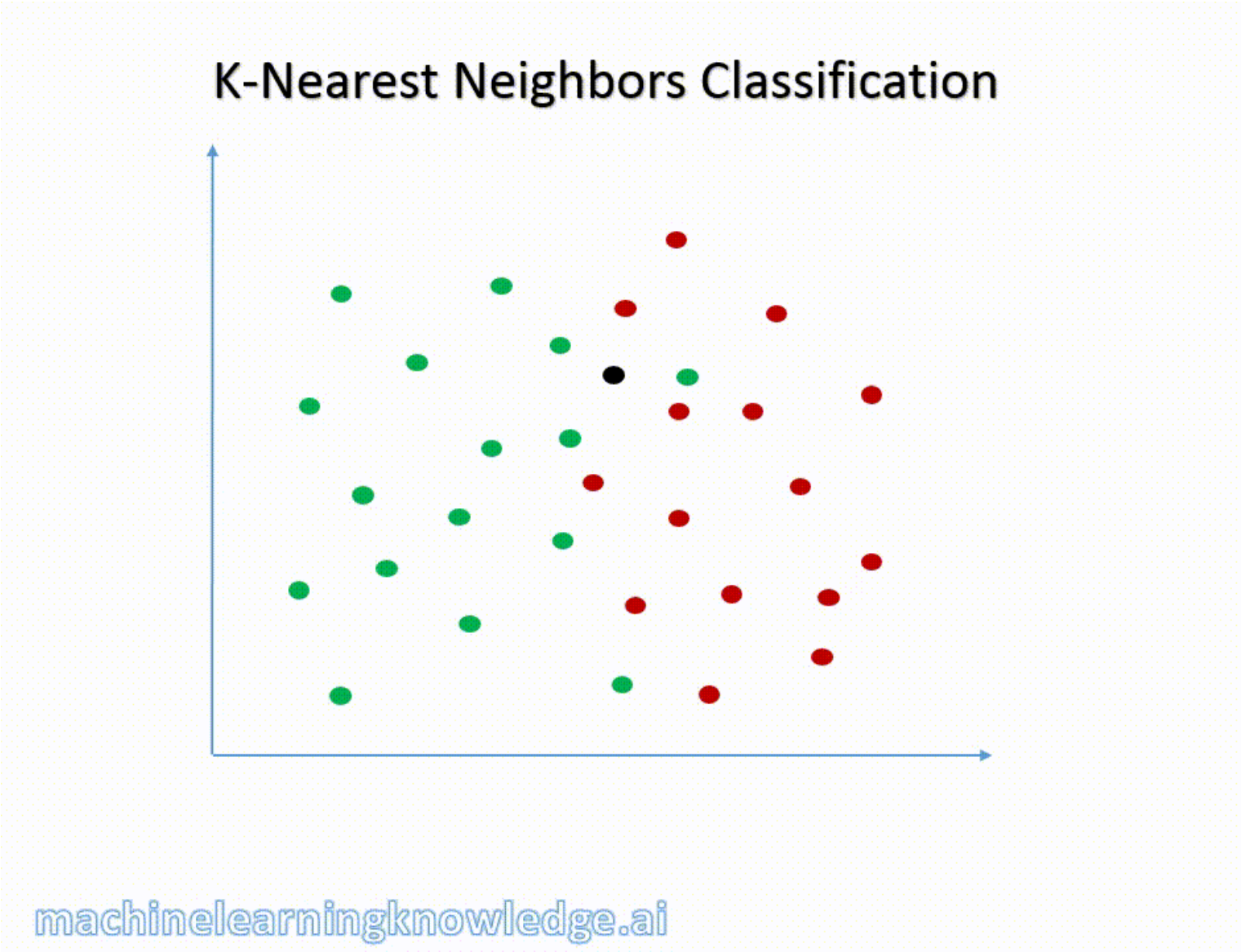
What is KNN?

A powerful classification algorithm used in pattern recognition.

K nearest neighbors stores all available cases and classifies new cases based on a similarity measure

A non-parametric lazy learning algorithm (An Instance based Learning method).

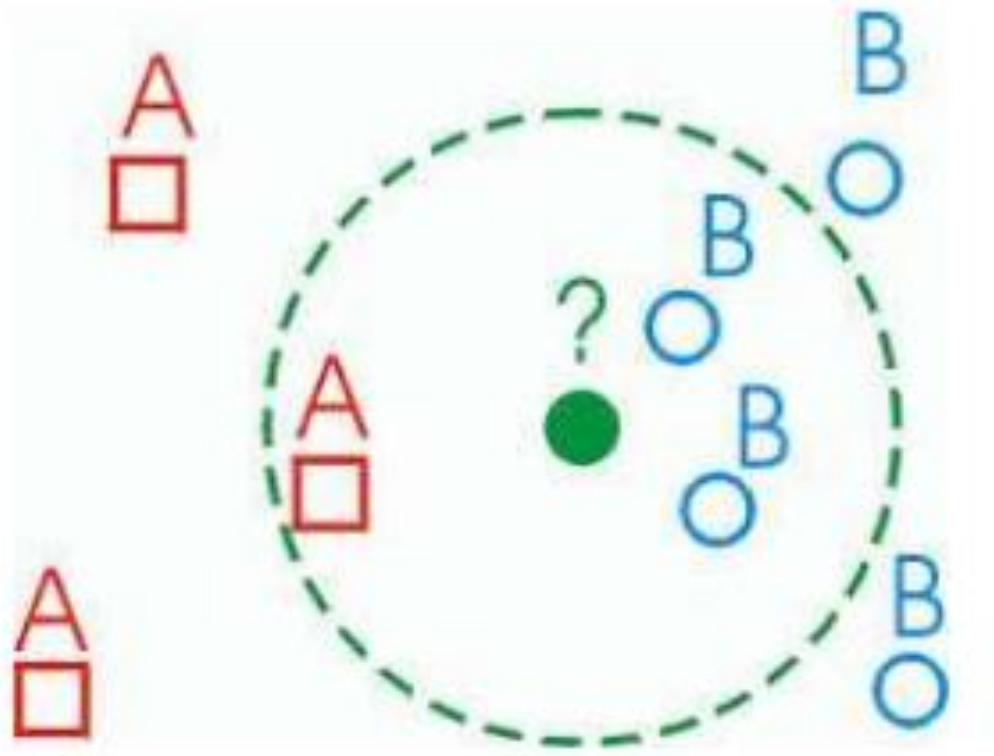
How KNN Works



KNN: Classification Approach

An object (a new instance) is classified by a majority votes for its neighbor classes.

The object is assigned to the most common class amongst its K nearest neighbors.(measured by a distant function)

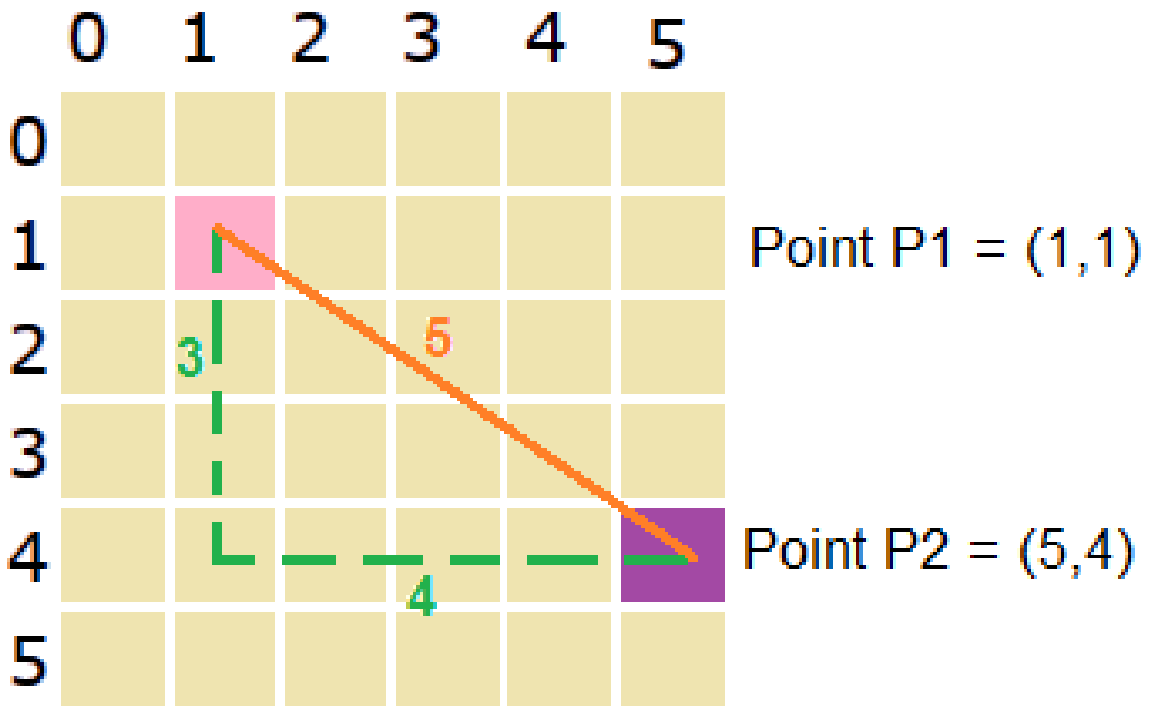


Working of KNN

- 1. Load the data
- 2. Initialize K i.e., your chosen number of neighbors
- 3. For each example in the data

  Calculate the distance between the example and all data points
- 4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- 5. Pick the first K entries from the sorted collection
- 6. Get the labels of the selected K entries
- 7. If regression, return the mean of the K labels
- 8. If classification, return the mode of the K labels

How to calculate the distance



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$

Distance Between Neighbors

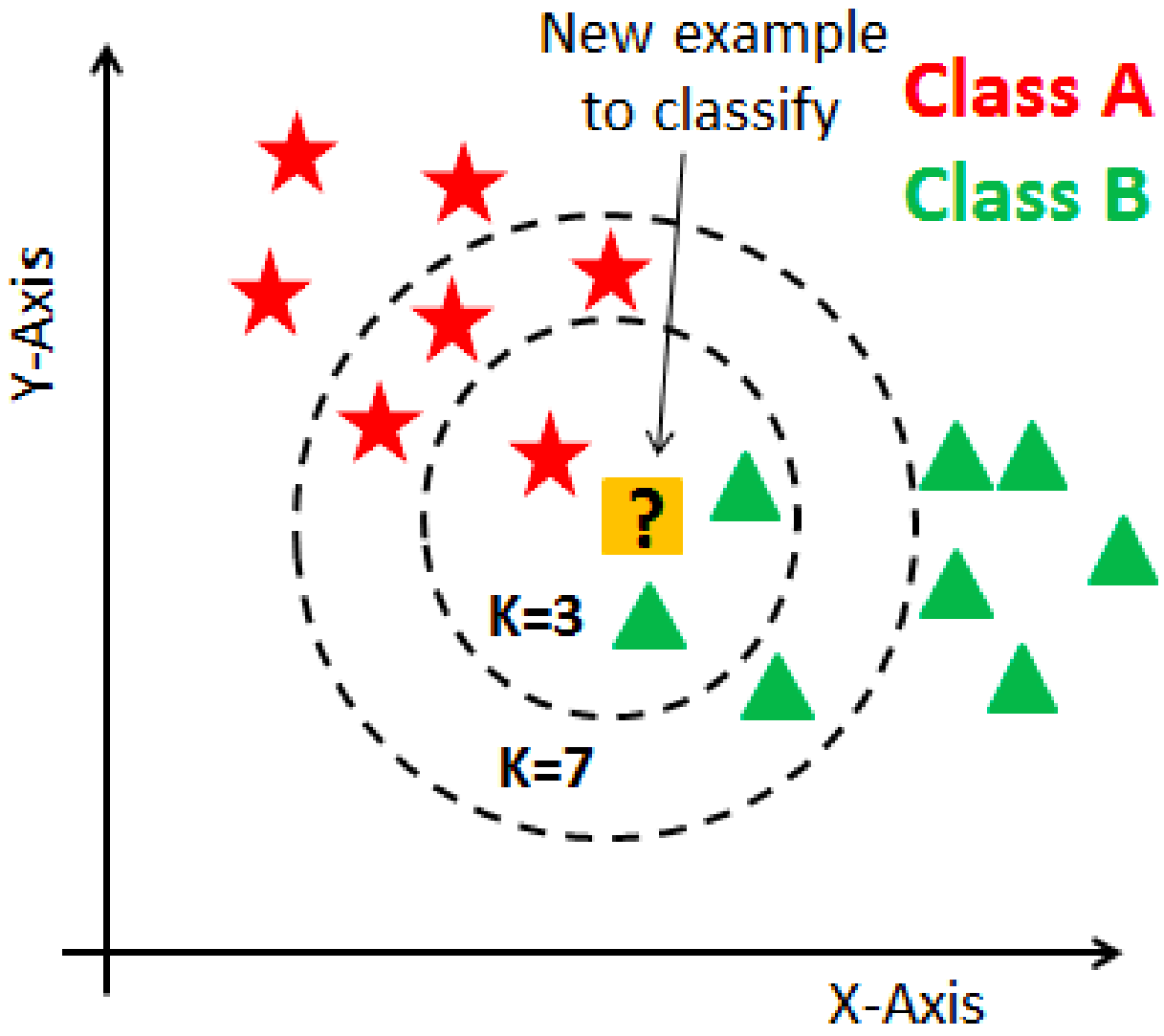
- Calculate the distance between new example (E) and all examples in the training set.
- Euclidean distance between two examples.
- X = [x1,x2,x3,..,xn]
- Y = [y1,y2,y3,...,yn]
- The Euclidean distance between X and Y is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

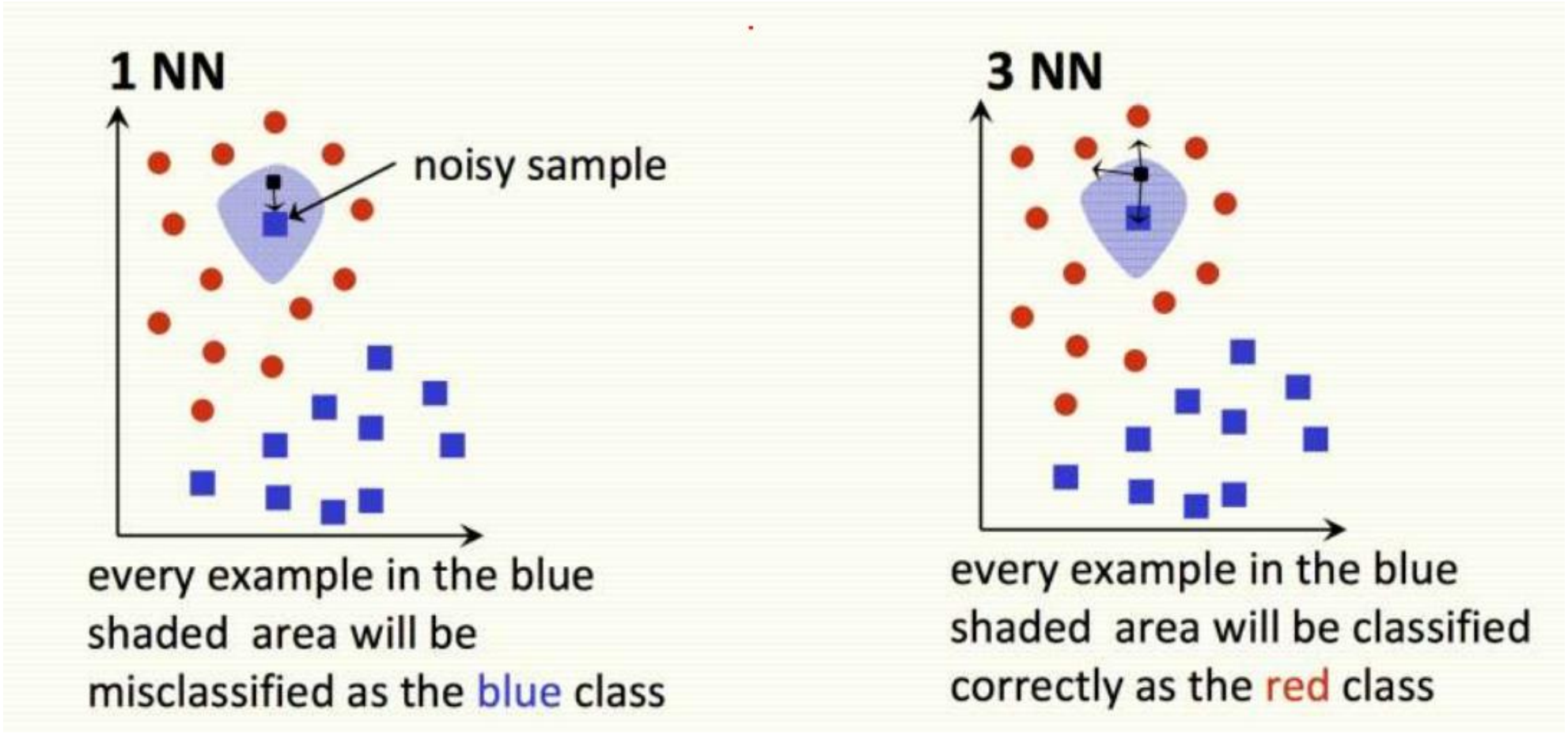
How to choose K?

- If the value of k is small then noise will have a higher dependency on the result. Overfitting of the model is very high in such cases.
- Bigger the value of K will destroy the principle behind KNN.
- Choose K as an odd number when the data has an even number of classes and even number when the data has an odd number of classes.
- Rule of thumb is K < sqrt(n), n is number of examples.

Example

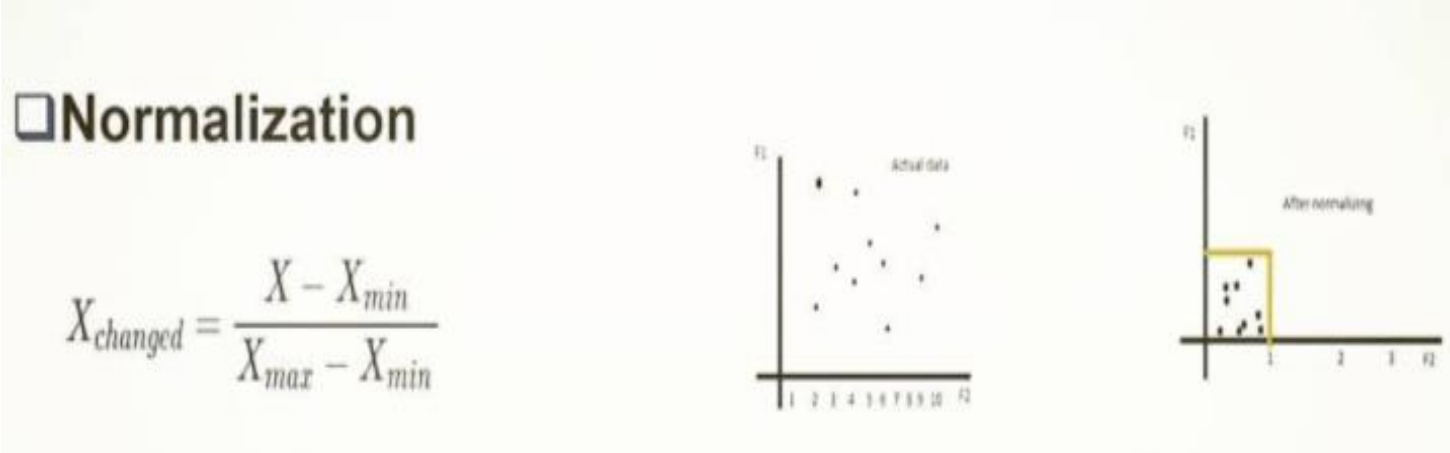


K Values


1 NN
noisy sample
every example in the blue shaded area will be misclassified as the blue class
3 NN
every example in the blue shaded area will be classified correctly as the red class

Feature Normalization

- Distance between neighbors could be dominated by some attributes with relatively large numbers. E.g., income of customers
- Arises when two features are in different scales.
- Important to normalize those features. – Mapping values to numbers between 0 – 1.



Weaknesses of KNN

- Takes more time to classify a new example.
- Need to calculate and compare distance from new example to all other examples.
- Choosing k may be tricky.
- Need large number of samples for accuracy.