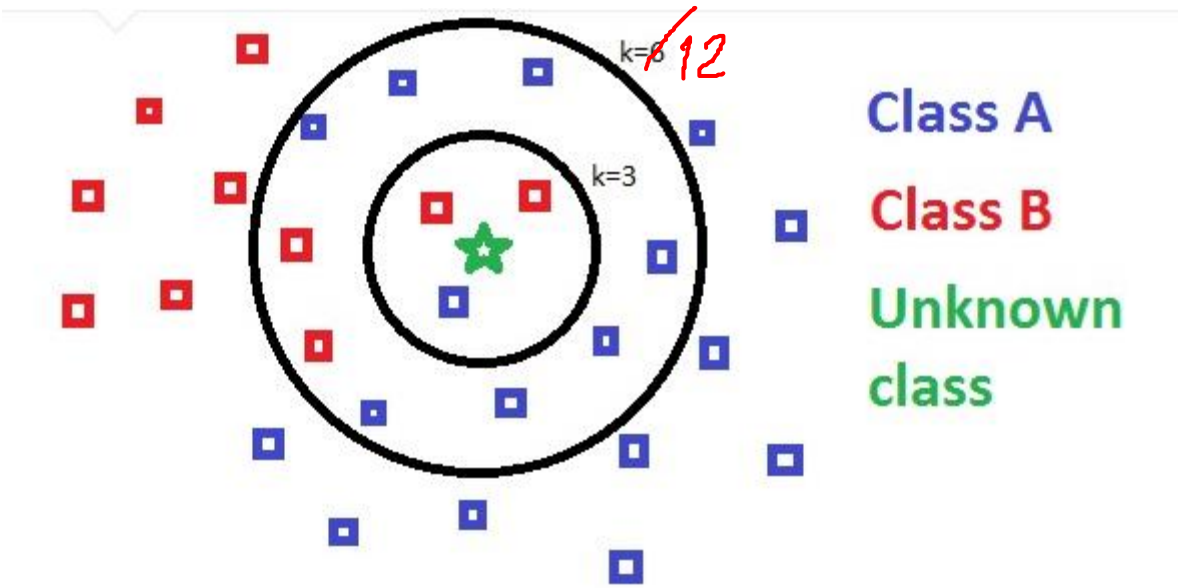
Topic 7.7 k-Nearest Neighbor Analysis

- ✓ k-nearest neighbor (k-NN) analysis is widely used in statistics and machine learning.
- ✓ In Machine learning it is a form of learning by analogy also.
- ✓ There may be k-NN Classification as well as Regression, like some other.
- ✓ k-NN classification:
  - Output is a class membership
  - An object is classified by a majority vote of its neighbors
  - k is a positive integer, typically small
- ✓ k-NN regression:
  - Output is the property value for the object
  - This value is the average of the values of *k* nearest neighbors.

✓k training samples from the sample space nearest to the given unknown sample are found.



Source: Internet

✓ Closeness may be defined by Euclidian distance in the following way:

$D(X, Y) = (\sum_{i=1:n} (x_i - y_i)^2)^{1/2}$,
where $X = (x_1, x_2, \dots , x_n)$ and $Y = (y_1, y_2, \dots , y_n)$
are two points in the n-dimensional sample space.

- Generally, it is Minkowski distance, which is widely used in ML. Manhattan and Euclidian are its variants.

- Other special purpose metrics: Mahalanobis distance, overlap metric / Hamming distance

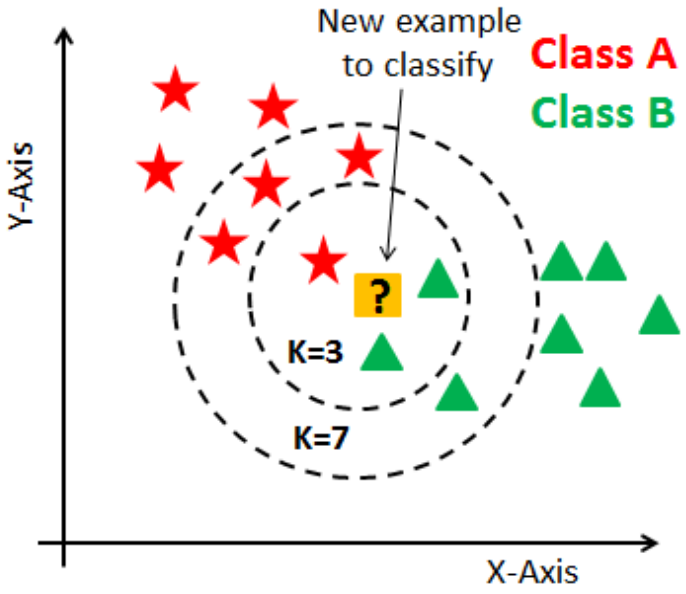
❖ k-NN : ‘lazy learner’

❖ No training phase: a classifier is not built until a new unlabeled sample is placed.

❖ In contrast, Decision tree and Neural network are ‘eager learners’.

❖ While Naïve Bayes may retain some probability calculations.

❖ A drawback of the basic "majority voting" classification occurs when the class distribution is skewed.



Source: Internet

❖ Classification accuracy can be improved through algorithms such as Large margin Nearest Neighbor.

❖ Choosing the value of k, decision boundary, feature reduction and feature selection, etc. are common concerns.