Supervised Learning (Classification with Nearest Neighbor Classifier)

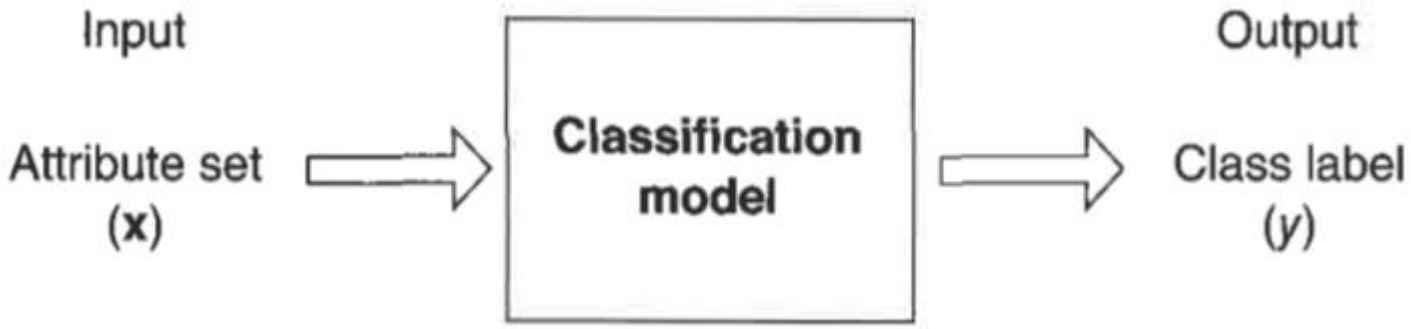
Dr. Yuzana Win (Nagasaki University, Japan)
Lecturer
Department of Computer Engineering and Information Technology

Lecture Objectives

- To introduce
  - What is Classification?
  - Classification Techniques or Methods
  - What is K-Nearest-Neighbors Algorithm (kNN)?
  - How does the KNN algorithm works?
  - Pros and Cons

What is Classification?

- Classification is the task of assigning objects to one of several predefined categories
- A classification technique (or classifier) is a systematic approach to building classification models from an input data set
- Each technique employs a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the input data.



Examples of Classification Task

- Categorizing news articles as finance, weather, entertainment, sports, etc.
- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Detecting spam email messages based upon the message header and content

Solving a Classification Problem

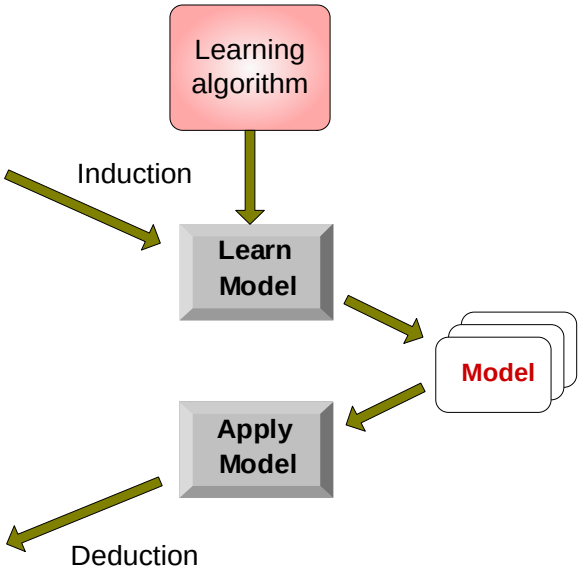
- A training set consisting of records whose class labels are known
- The training set is used to build a classification model, which is subsequently applied to the test set, which consists of records with unknown class labels
- Evaluation of the performance of a classification model is based on the counts of test records correctly and incorrectly predicted by the model.

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Classification Techniques or Methods

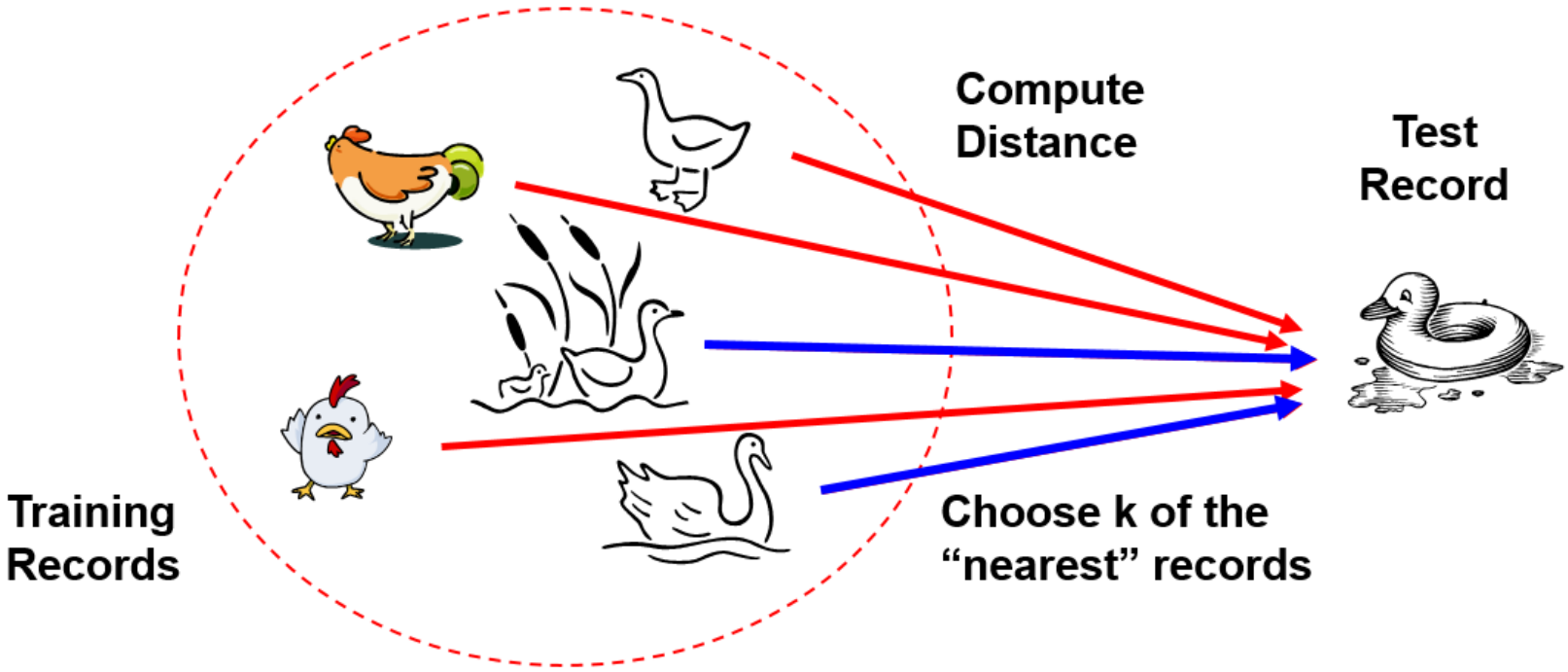
- Naïve Bayes Classifier *
- Support Vector Machines *
- Neural Networks (Deep Learning)
- Decision Tree based Methods
- Rule-based Methods
- K-Nearest Neighbors (kNN) *

Classification Techniques or Methods

- Naïve Bayes Classifier *
- Support Vector Machines *
- Neural Networks (Deep Learning)
- Decision Tree based Methods
- Rule-based Methods
- K-Nearest Neighbors (KNN) *

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



What is K-Nearest Neighbors Algorithm(KNN)

- KNN is a non-parametric and lazy learning algorithm
- Non-parametric means there is no assumption for underlying data distribution.
- Lazy algorithm means it does not need any training data points for model generation. All training data used in the testing phase.
- This makes training faster and testing phase slower

K-Nearest Neighbors Algorithm(KNN)

Algorithm Algorithm for finding K nearest neighbors.

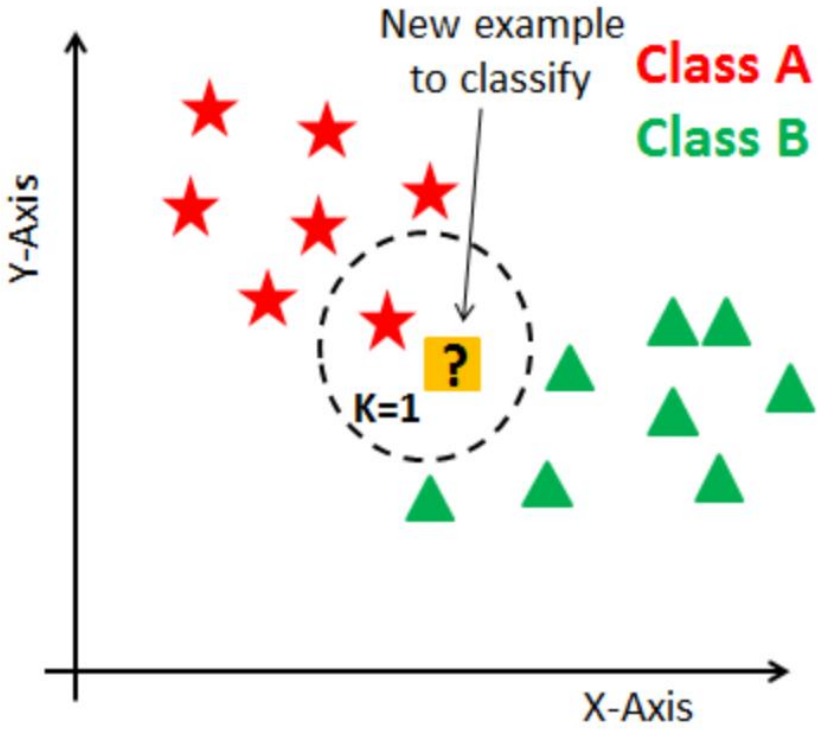
1: **for** $i = 1$ to *number of data objects* **do**
2: Find the distances of the i^{th} object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: **return** the objects associated with the first K distances of the sorted list
5: **end for**

How does the KNN algorithm works?

- In KNN, K is the number of nearest neighbors.
- The number of neighbors is the core deciding factor.
- K is generally an odd number if the number of classes is 2.
- When K=1, then the algorithm is known as the nearest neighbor algorithm.

How does the KNN algorithm work?

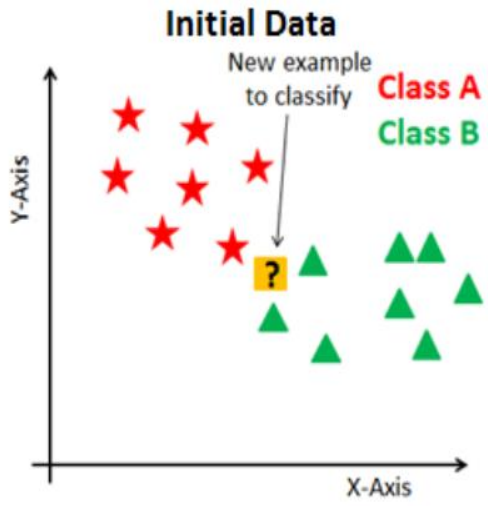
- Suppose P1 is the point, for which label needs to predict. First, you find the one closest point to P1 and then the label of the nearest point assigned to P1.

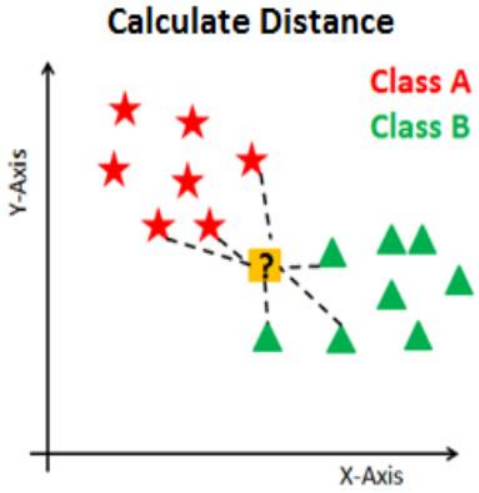


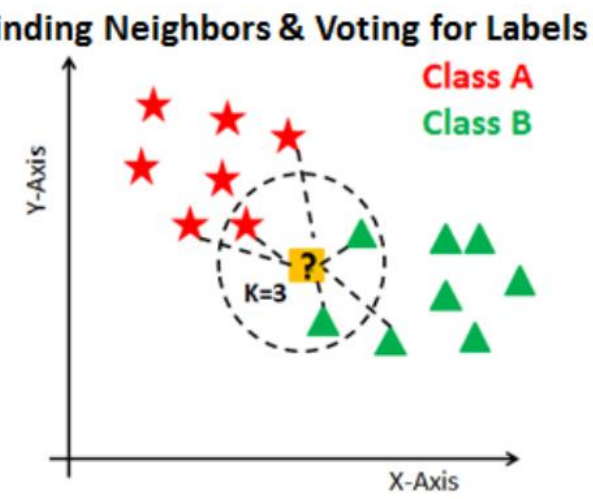
How does the KNN algorithm work?

- KNN has the following basic steps:
  - Calculate distance
    - Euclidean distance, Hamming distance, Manhattan distance and Minkowski distance
  - Find closest neighbors
    - Identify *k* nearest neighbors
  - Vote for labels
    - Use class labels of nearest neighbors to determine the class label of unknown record (e.g. by taking majority vote)

How does the KNN algorithm work?

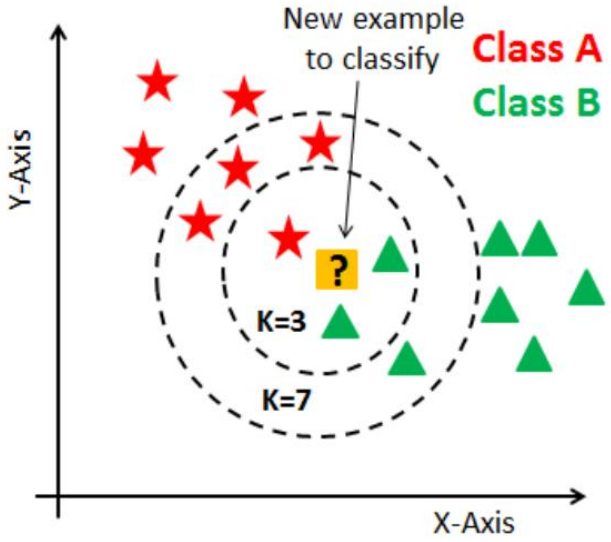




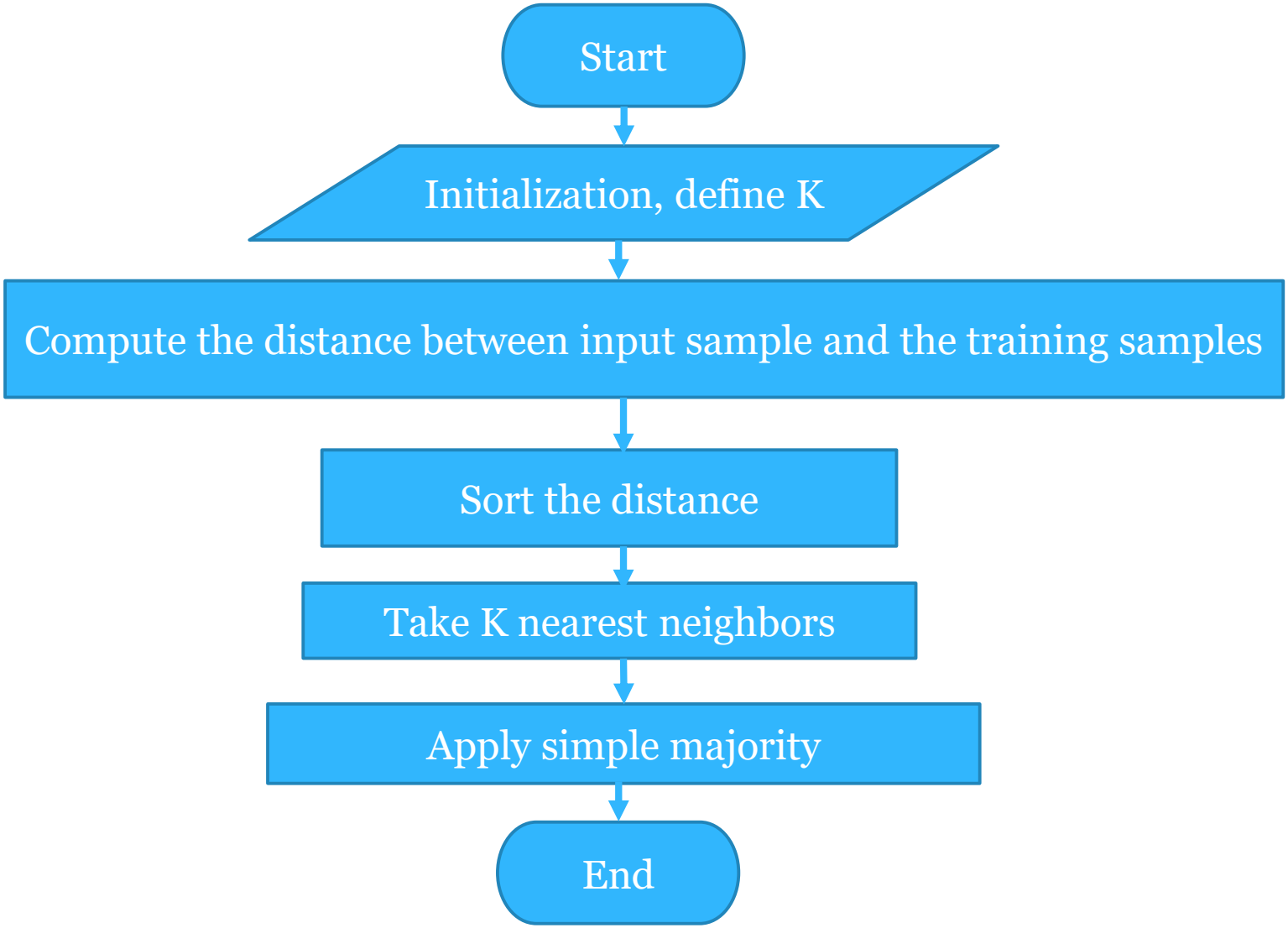


How to choose the value of K?

- Data scientists choose as an odd number if the number of classes is even.
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes
- Also check by generating the model on different values of k and check their performance.



Flow Diagram of K-Nearest Neighbors Algorithm(KNN)



Example

- The data from testing with two attribute (acid durability and strength) to classify whether a special paper tissue is good or not.
- The factory produces a new paper tissue that passes the laboratory test with X1=3 and X2=7. Guess the classification of this new tissue

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

How to implement KNN

- Step 1: Initialize and Define k
  - Assume k=3
- Step 2: Compute the distance between input sample and training sample
  - Co-ordinate of the input sample is (3,7)
  - Use the Euclidean distance and calculating Squared Euclidean distance

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance
7	7	$(7-3)^2 + (7-7)^2 = \mathbf{16}$
7	4	$(7-3)^2 + (4-7)^2 = \mathbf{25}$
3	4	$(3-3)^2 + (4-7)^2 = \mathbf{09}$
1	4	$(1-3)^2 + (4-7)^2 = \mathbf{13}$

How to implement KNN

- Step 3: Sort the distance and determine the nearest neighbors based on the K^{th} minimum distance

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance	Rank minimum distance	Is it included in 3-Nearest Neighbour?
7	7	16	3	Yes
7	4	25	4	No
3	4	09	1	Yes
1	4	13	2	Yes

How to implement KNN

- Step 4: Take 3-Nearest Neighbors
  Gather the category Y of the nearest neighbors

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Squared Euclidean distance	Rank minimum distance	Is it included in 3-Nearest Neighbour?	Y = Category of the nearest neighbour
7	7	16	3	Yes	Bad
7	4	25	4	No	-
3	4	09	1	Yes	Good
1	4	13	2	Yes	Good

How to implement KNN

- Step 5: Apply simple majority
  Use simple majority of the category of the nearest neighbors is the prediction value of the query instance

- According to the results, 2 “good” and 1 “bad” are obtained

- Therefore, the new paper tissue passes the laboratory test with X1=3 and X2=7 is in the “good” category

Advantages of KNN classifier

- Very simple and intuitive
- Good classification if the number of samples is large enough
- Can be useful in nonlinear data
- No need to train a model for generalization
- The training phase of K-nearest neighbor classification is much faster compared to other classification algorithms

Disadvantages of KNN classifier

- Choosing K may be tricky
- Test stage is computationally expensive
- No training stage, all the work is done during the test stage
- KNN is not suitable for large dimensional data

Applications of KNN Classifier

- Classifications
- Getting missing values
- Pattern recognition
- Gene expression
- Protein-protein prediction
- 3D structure of protein
- Measure document similarity

Conclusion

- KNN is eager learning or lazy learning algorithm
- Simple and easy to understand and explain
- Very flexible decision boundaries
- Requires a lot computation cost

Next Week Lecture

- Supervised Learning: Classification with Naïve Bayes Classifier

Thank You