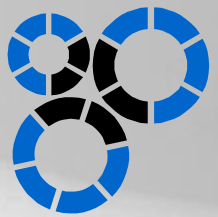
Machine Learning
CSE - 465

Lecture - 09

Lecture 09
K-Nearest Neighbor Classifier

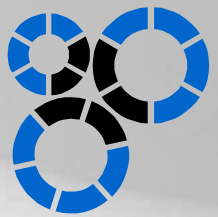
Outline

- Introduction to K- Nearest Neighbor Classifier
- How to determine a good value for k?
- KNN Example
- Advantages and disadvantages of KNN



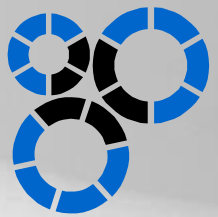
Introduction to KNN

- Nearest-neighbor classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it.
- The training tuples are described by n attributes.
- When k = 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space.



How to determine a good value for k?

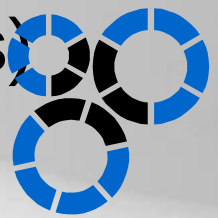
- Starting with k = 1, we use a test set to estimate the error rate of the classifier.
- The k value that gives the minimum error rate may be selected.
- If infinite number of samples available, the larger is k, the better is classification.



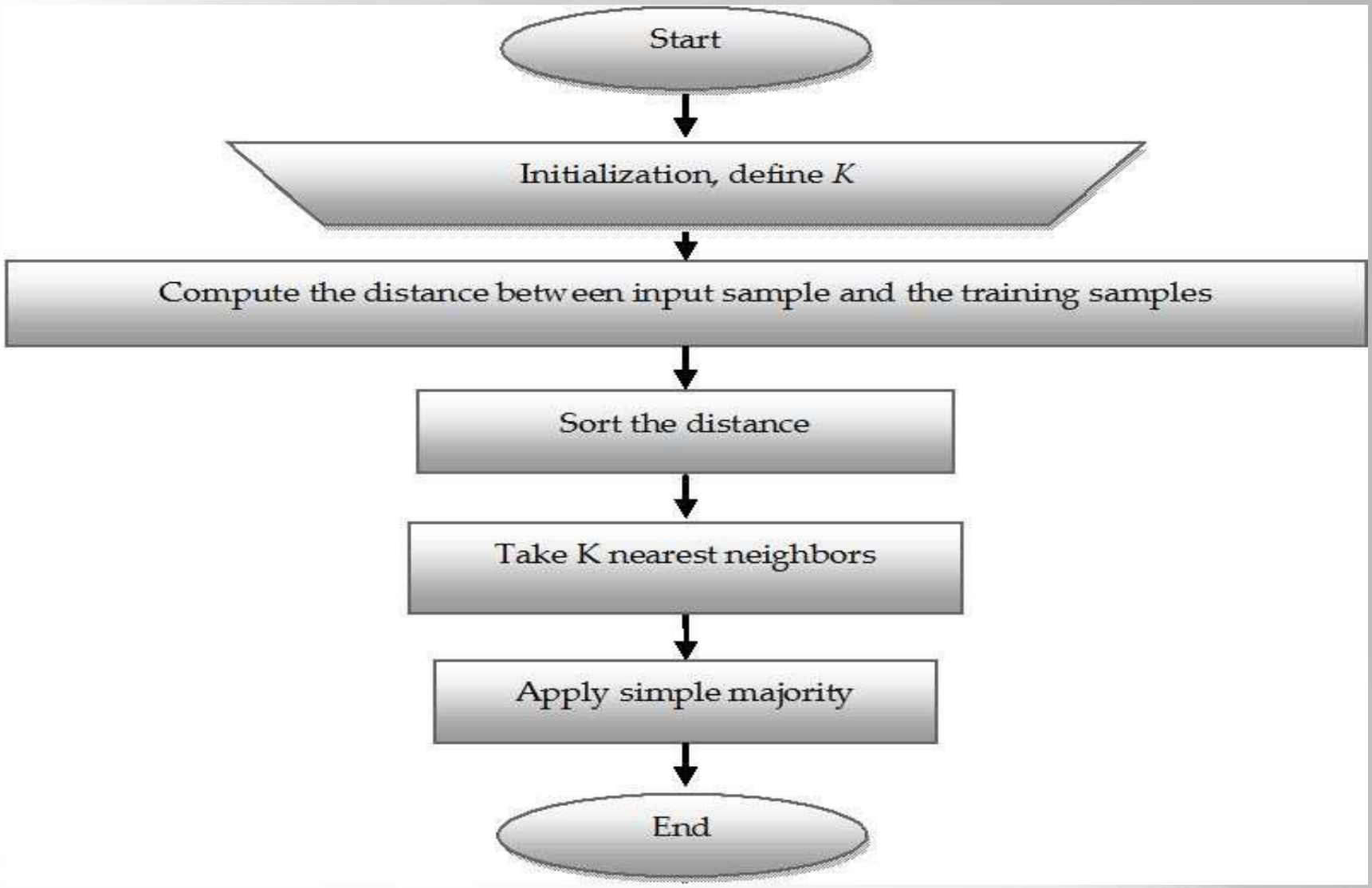
How to determine a good value for k?

- k = 1 is often used for efficiency, but sensitive to “noise”
- Larger k gives smoother boundaries, better for generalization, but only if locality is preserved. Locality is not preserved if end up looking at samples too far away, not from the same class.
- Interesting relation to find k for large sample data :

  k = sqrt(n)/2 (where n is number of examples)



KNN Algorithm



KNN Example

- We are testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not.
- Here are four training samples :

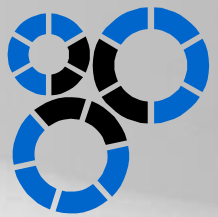
X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

- Now 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.

KNN Example

- Step 1 : Initialize and Define k.
  - Lets say, k = 3 (Always choose k as an odd number if the number of attributes is even to avoid a tie in the class prediction)

- Step 2 : Compute the distance between input sample and training sample
  - Co-ordinate of the input sample is (3,7).
  - Instead of calculating the Euclidean distance, we calculate the Squared Euclidean distance.



KNN Example

- Step 2 (Continued)

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	$(1-3)^2 + (4-7)^2 = \mathbf{13}$

KNN Example

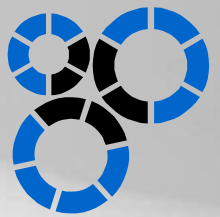
- Step 3 : Sort the distance and determine the nearest neighbours based of the Kth minimum distance :

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

KNN Example

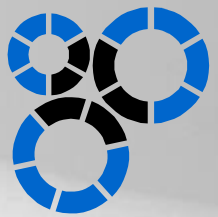
- Step 4 : Take 3-Nearest Neighbors:
- Gather the category Y of the nearest neighbors.

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



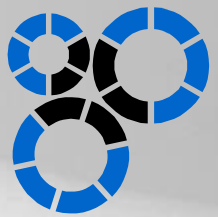
KNN Example

- Step 5 : Apply simple majority
- Use simple majority of the category of the nearest neighbors as the prediction value of the query instance.
- We have 2 “good” and 1 “bad”.
- Thus we conclude that the new paper tissue that passes the laboratory test with X1 = 3 and X2 = 7 is included in the “good” category.



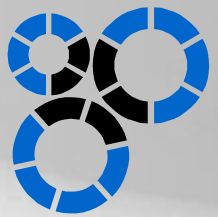
Advantages of KNN

- Can be applied to the data from any distribution for example, data does not have to be separable with a linear boundary
- Very simple and intuitive
- Good classification if the number of samples is large enough



Disadvantages of KNN

- Choosing k may be tricky
- Test stage is computationally expensive
- No training stage, all the work is done during the test stage
- This is actually the opposite of what we want. Usually we can afford training step to take a long time, but we want fast test step



Thank You