Eager Learners vs Lazy Learners

- Eager learners, when given a set of training tuples,
- will construct a generalization model before receiving
- new (e.g., test) tuples to classify.
- Lazy learners simply stores data (or does only a little minor processing) and waits until it is given a
- test tuple.

Lazy learners store the training tuples or “instances,” they are also referred to as instance based learners, even though all learning is essentially based on

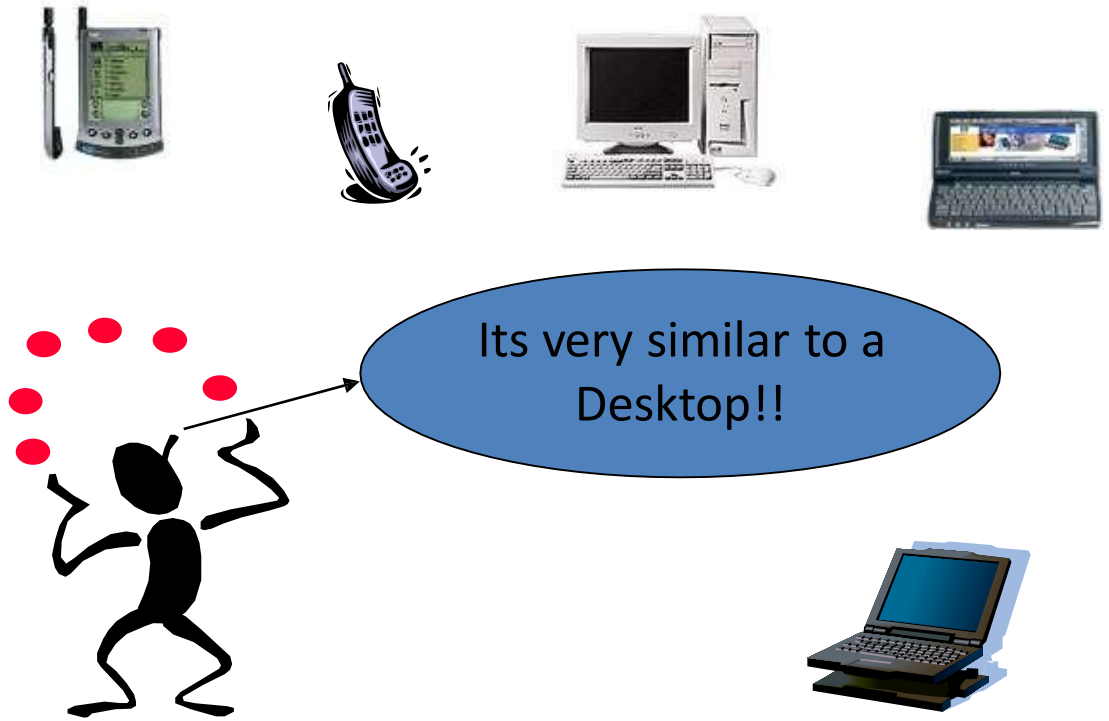
-**k- Nearest Neighbor Classifier**

-Case Based Classifier

Simple Analogy..

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

Instance-based Learning



k- Nearest Neighbor Classifier

History

- It was first described in the early 1950s.
- The method is labor intensive when given large training sets.
- Gained popularity, when increased computing power became available.
- Used widely in area of pattern recognition and statistical estimation.

What is k- NN??

- 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.

Closeness

- The Euclidean distance between two points or tuples, say,

X1 = (x11,x12,...,x1n) and X2 =(x21,x22,...,x2n), is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x_{1i} - x_{2i})^2}.$$

- Min-max normalization can be used to transform a value v of a numeric attribute A to v0 in the range [0,1] by computing

$$v' = \frac{v - min_A}{max_A - min_A},$$

Distance measure for Continuous Variables

Distance functions

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

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

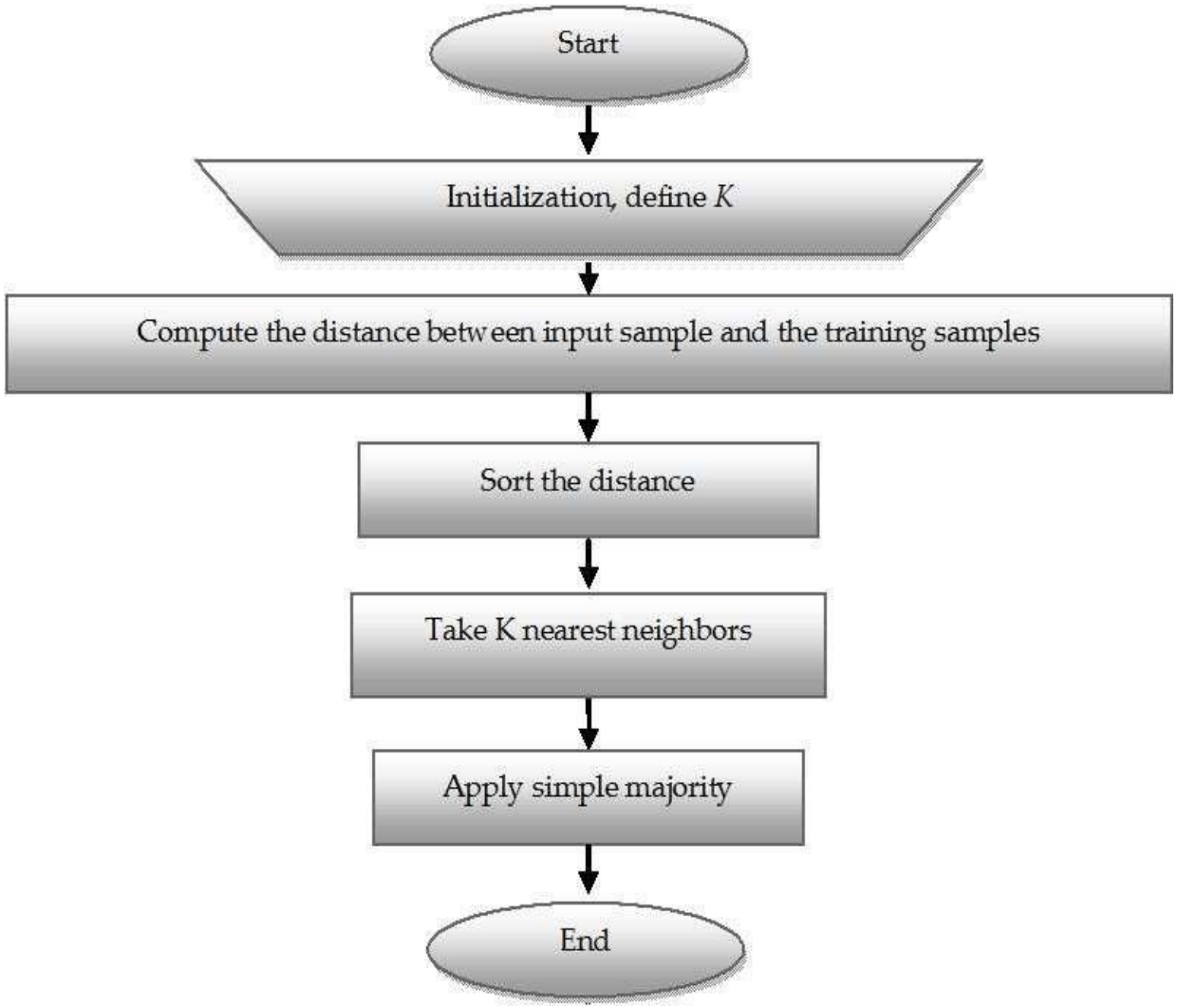
Minkowski $$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

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.

KNN Algorithm and Example

KNN Classifier Algorithm



Example

- We have data from the questionnaires survey and objective 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.

- 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.

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

- 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**

- Step 4 : **Take 3-Nearest Neighbours:**
- Gather the category Y of the nearest neighbours.

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**

- Step 5 : **Apply simple majority**

- Use simple majority of the category of the nearest neighbours 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 classifier :**
- 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 classifier :**
- 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

Applications of KNN Classifier

- Used in classification
- Used to get missing values
- Used in pattern recognition
- Used in gene expression
- Used in protein-protein prediction
- Used to get 3D structure of protein
- Used to measure document similarity

CASE STUDY

Car manufacturer company that has manufactured a new SUV car. The company wants to give the ads to the users who are interested in buying that SUV.
So for this problem, we have a dataset that contains multiple user's information through the social network.

The dataset contains lots of information but the **Estimated Salary** and **Age** we will consider for the independent variable and the **Purchased variable** is for the dependent variable. Below is the dataset

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

Steps to implement the K-NN algorithm:

- Data Pre-processing step
- Fitting the K-NN algorithm to the Training set
- Predicting the test result
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.