

# Machine Learning: Classification versus Clustering

by Elena Battini Sönmez  
İstanbul Bilgi University

by Elena Battini Sönmez, İstanbul Bilgi University

# The Classification problem:

- We start with a database of objects whose classes are already known
  - The database is known as the training database since it allows us to train a model, which learns to distinguish among those objects
- We take a new sample, and we want to know its class

# Example of classification:

- Suppose we have a database storing info of different people, together with their emotions:



- We want to be able to use this database to assign an emotion to a test face: which is the emotion of this picture?



by Elena Battini Sönmez, İstanbul Bilgi University

# The k-Nearest Neighbours:

- The k-Nearest Neighbours (**k-NN**) classification algorithm considers the k-neighbours of the test sample and assigns it to the majority of the class
- Question 1: What is a digital image?
- Question 2: What makes two items count as similar, and how do we measure similarity?

# What is a digital image?



08	02	22	97	38	15	00	40	00	75	04	05	07	78	52	12	50	77	51	74
49	49	99	40	17	81	18	57	60	87	17	40	98	43	69	48	04	56	62	00
81	49	31	73	55	79	14	29	93	71	40	67	58	88	30	03	49	13	36	65
52	70	95	23	04	60	11	42	63	44	65	56	01	32	56	71	37	02	36	91
22	31	16	71	51	67	85	59	41	92	36	54	22	40	40	28	66	33	13	80
24	47	33	62	99	03	45	02	44	75	33	53	78	36	84	20	35	17	12	50
32	98	81	28	64	23	67	10	26	38	40	67	59	54	70	66	18	38	64	70
67	26	20	68	02	62	12	20	95	63	94	39	63	08	40	91	66	49	94	21
24	55	58	05	66	73	99	26	97	17	78	78	96	83	14	88	34	89	63	72
21	36	23	09	75	00	76	44	20	45	35	14	00	61	33	97	34	31	33	95
78	17	53	28	22	75	31	67	15	94	03	80	04	62	16	14	09	53	56	92
16	39	05	42	96	35	31	47	55	58	88	24	00	17	54	24	36	29	85	57
86	56	00	48	35	71	89	07	05	44	44	37	44	60	21	58	51	54	17	58
19	80	81	68	05	94	47	69	28	73	92	13	86	52	17	77	04	89	55	40
04	52	08	83	97	35	99	16	07	97	57	32	16	26	26	79	33	27	98	66
85	47	68	87	57	62	20	72	03	46	33	67	46	55	12	32	63	93	53	69
04	42	16	73	38	45	39	11	24	94	72	18	08	46	29	32	40	62	76	36
20	69	36	41	72	30	23	88	34	70	82	69	82	67	59	85	74	04	36	16
20	73	35	29	78	31	90	01	74	31	49	71	41	54	53	16	23	57	05	54
01	70	54	71	83	51	56	69	16	92	33	48	61	43	52	01	89	21	62	48

What the computer sees

image classification

82% cat  
15% dog  
2% hat  
1% mug

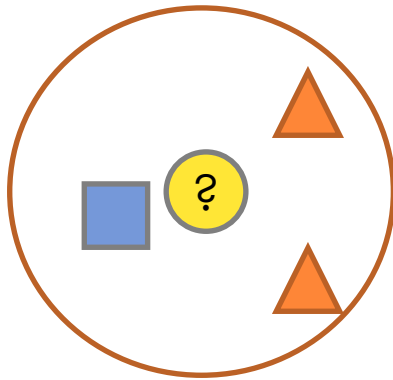
# What makes two items count as similar?

## Euclidean distance:

- The k-NN algorithm interprets each object in the database as a point in the space; that is, each attribute is a feature, a coordinate in the plane
- The **similarity** of two points is measured as the distance between them

$$\text{Euclidean\_dist}((x,y),(a,b)) = \sqrt{(x-a)^2 + (y-b)^2}$$

# k-NN Algorithm:

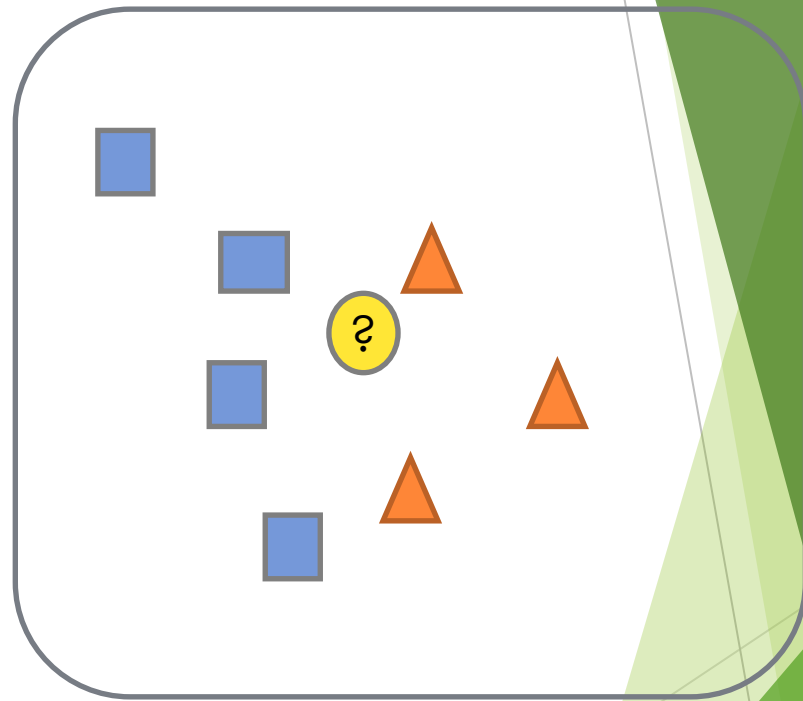


- It requires:
  1. The set of stored labeled records (training set)
  2. A distance metric to compute the distance between records
  3. The value of  $k$ , the number of nearest neighbors to consider
- To classify an unknown record (test sample):
  - Compute distance to all other training records
  - Identify  $k$  nearest neighbors
  - Use class labels of nearest training samples to assign the class (e.g., by taking majority vote) to the test sample

# Challenges of k-NN:

- Choosing the value of  $k$ :
  - If  $k$  is too small, sensitive to noise points
  - If  $k$  is too large, neighborhood may include points from other classes
  - Choose an odd value for  $k$ , to eliminate ties

Q: Give the class for  $K=1, 3, 5$





# Problems of k-NN:

- Computationally intensive, especially when the size of the training set grows
- High dimension
- Accuracy can be severely degraded by the presence of noisy or irrelevant features

# Clustering:

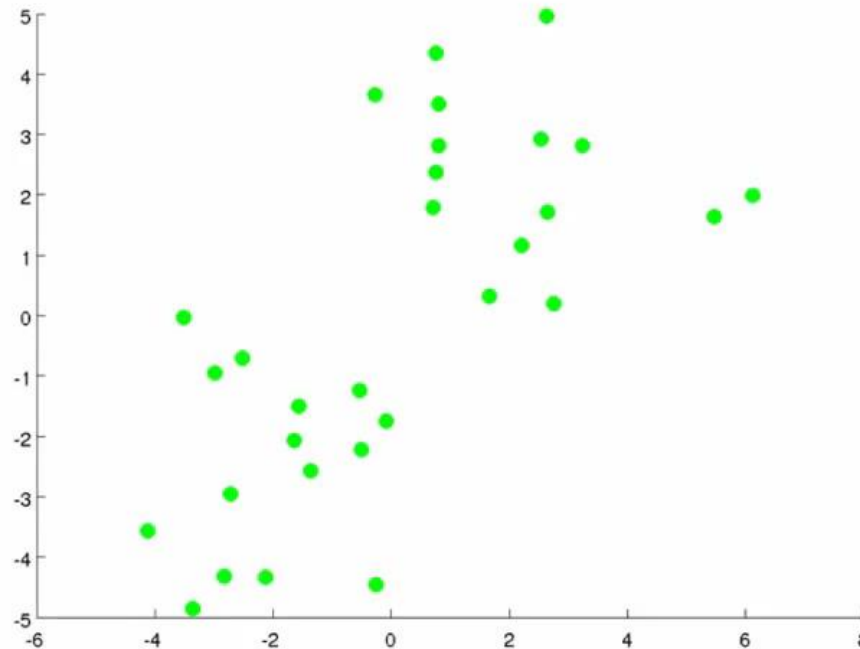
- The process of organizing objects into groups whose members are similar in some way  
*A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters*
- The goal of clustering is to determine the intrinsic grouping in a set of **unlabeled** data

# Example of clustering:

- *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records
- *Biology*: clustering of plants and animals given their features
- *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds
- *WWW*: document clustering; clustering weblog data to discover groups of similar access patterns

# K-means algorithm (1/6):

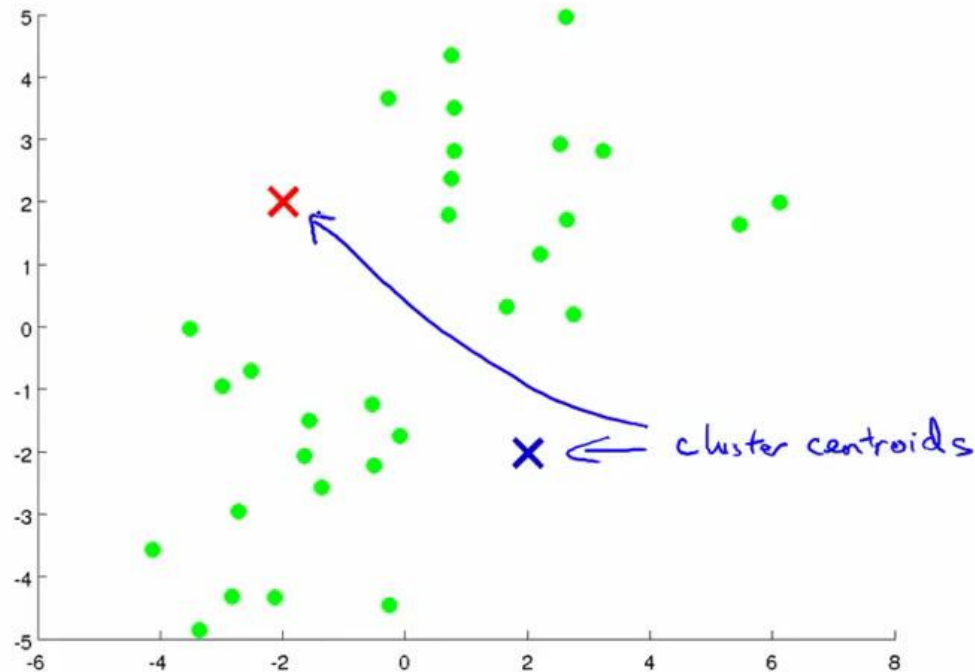
(Coursera, Machine Learning course - week 8)



Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University

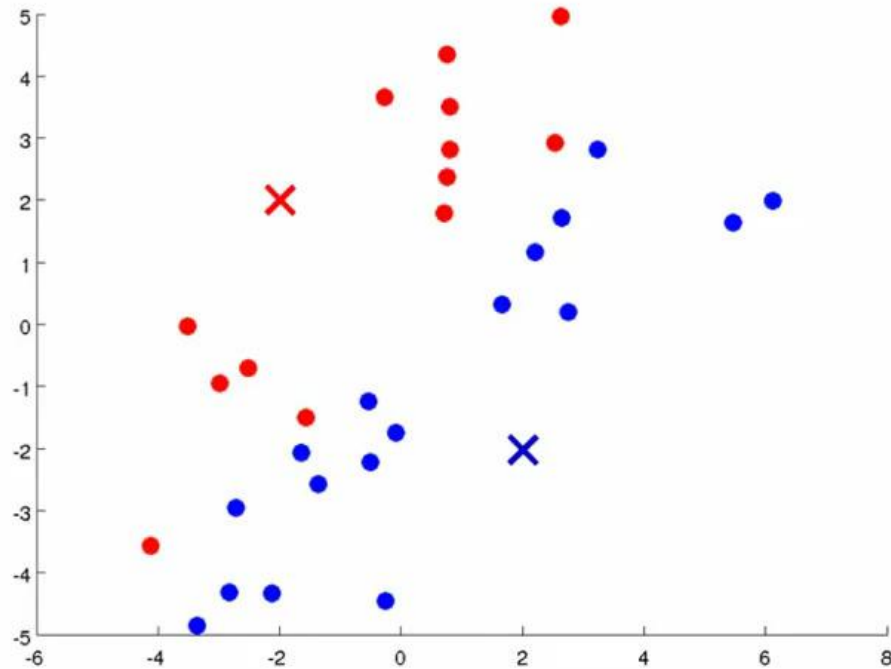
# K-means algorithm (2/6):



Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University

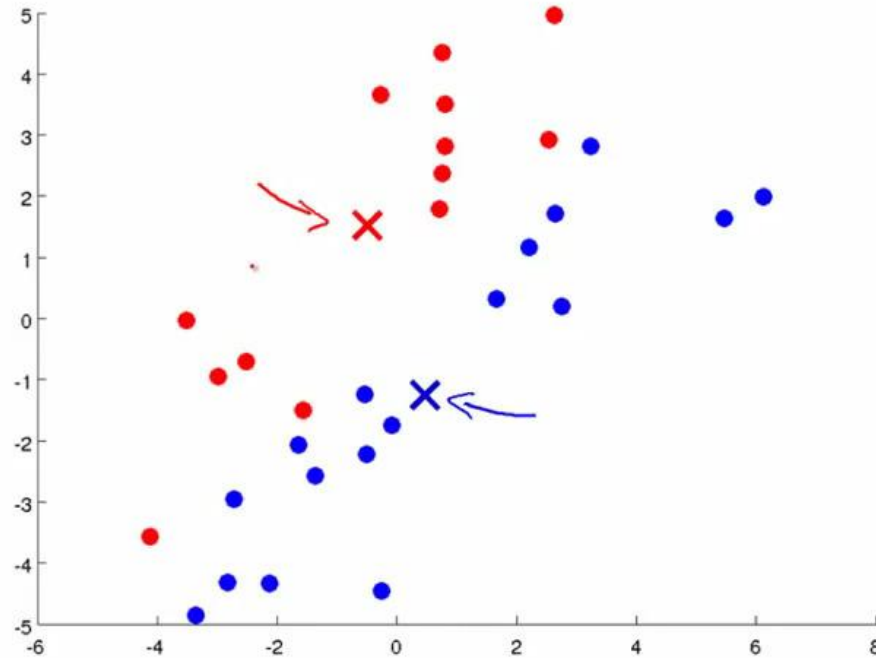
# K-means algorithm (3/6):



Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University

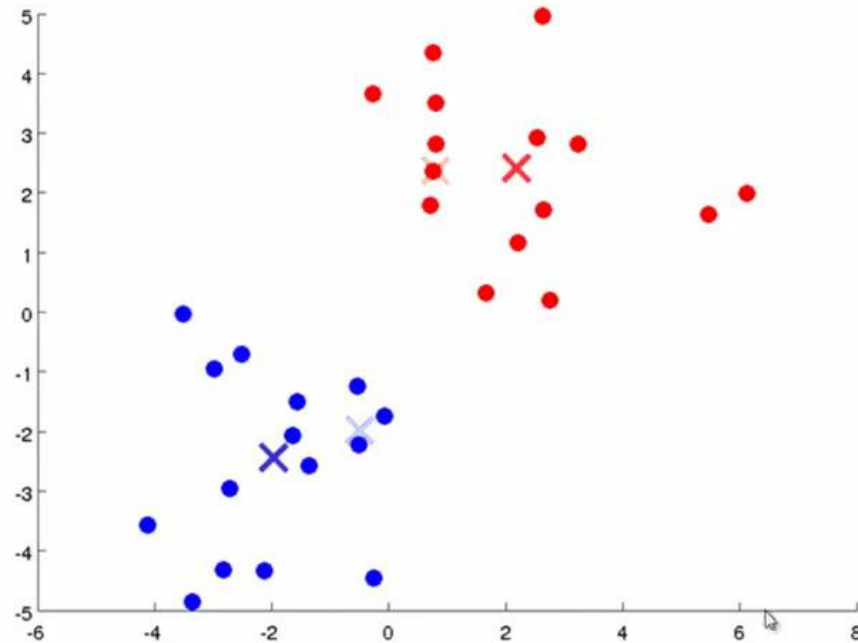
# K-means algorithm (4/6):



Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University

# K-means algorithm (5/6):



Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University



# K-means algorithm (6/6):

## K-means algorithm

Randomly initialize  $K$  cluster centroids  $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

Repeat {

  for  $i = 1$  to  $m$

$c^{(i)} :=$  index (from 1 to  $K$ ) of cluster centroid  
    closest to  $x^{(i)}$

  for  $k = 1$  to  $K$

$\mu_k :=$  average (mean) of points assigned to cluster  $k$

}

Andrew Ng

by Elena Battini Sönmez, İstanbul Bilgi University

# Requirements of clustering algorithms:

- scalability
- dealing with different types of attributes
- discovering clusters with arbitrary shape
- ability to deal with noise and outliers
- high dimensionality