

CLUSTERING

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: <http://jelenajovanovic.net>

OUTLINE

- What is clustering?
- Application domains
- K-Means clustering method
- Expectation Maximization (EM) clustering method

WHAT IS CLUSTERING?

Clustering is an *unsupervised* ML task

- its input is a set of instances/observations (described with a set of attributes) that should be grouped based on their similarity
- there is no data about the desired/correct group for any of the the instances from the dataset

WHAT IS CLUSTERING?

This grouping of instances should be done in such a manner that for each instance the following is true:

- the instance is more *similar* to the instances from its group (cluster), than to instances from other groups (clusters)

HOW TO ESTIMATE SIMILARITY?

Similarity between instances is computed using one of

- similarity measures (e.g., Correlation coefficient, Cosine similarity), or
- distance measures (e.g., Euclidian distance, Manhattan distance)

HOW TO ESTIMATE SIMILARITY?

Euclidian distance:

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

Manhattan distance:

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

p – number of attributes that describe the instances in the dataset

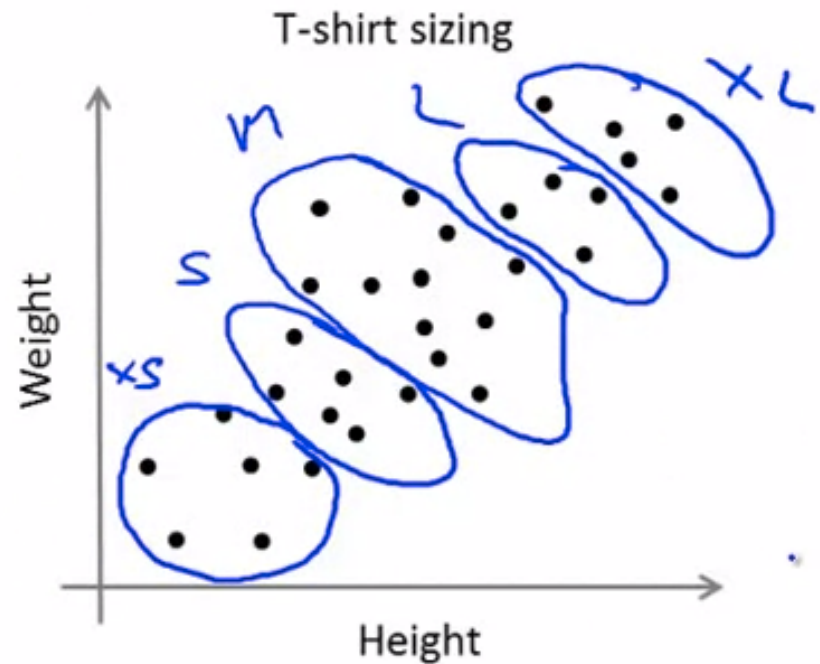
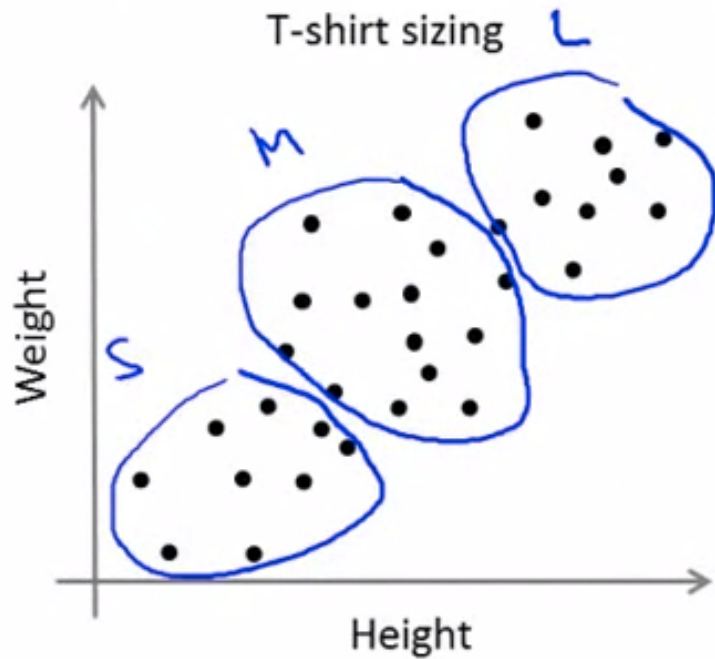
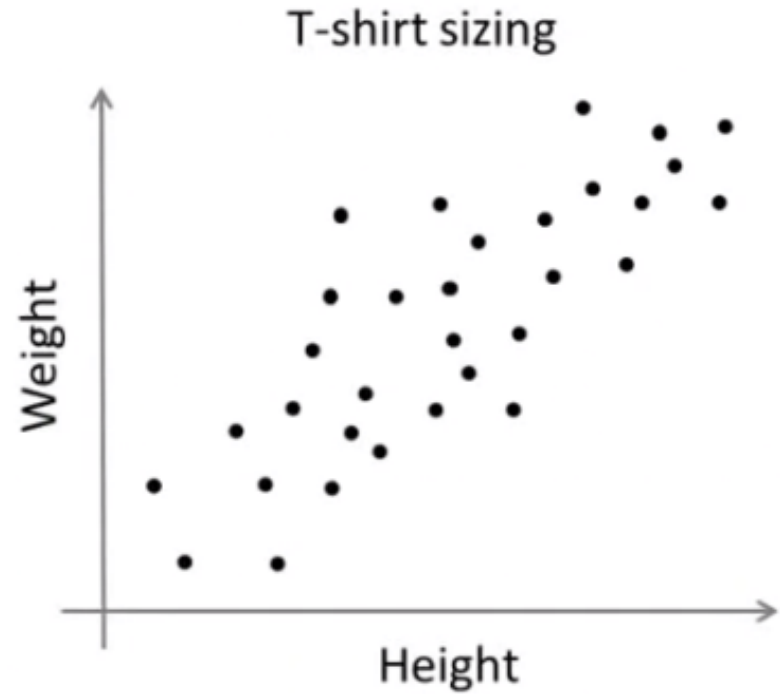


HOW TO EVALUATE THE RESULTS?

Unlike the classification task, for this task, there is no unique “correct” or best solution

- how good/suitable a solution is, that depends upon the specific domain and the application case
- the same solution might be differently evaluated in different application cases
- if it is to be done properly, domain experts need to evaluate the solution(s) produced by the model

An example illustrating different valid solutions for the same input dataset



APPLICATION DOMAINS

- Market segmentation
- Detection of groups/communities in social networks
- Identifying patterns in user tracking data -> allows for learning about the ways people use an application
- Grouping of objects (e.g., images or documents) to facilitate search / discovery
- ...

K-MEANS ALGORITHM



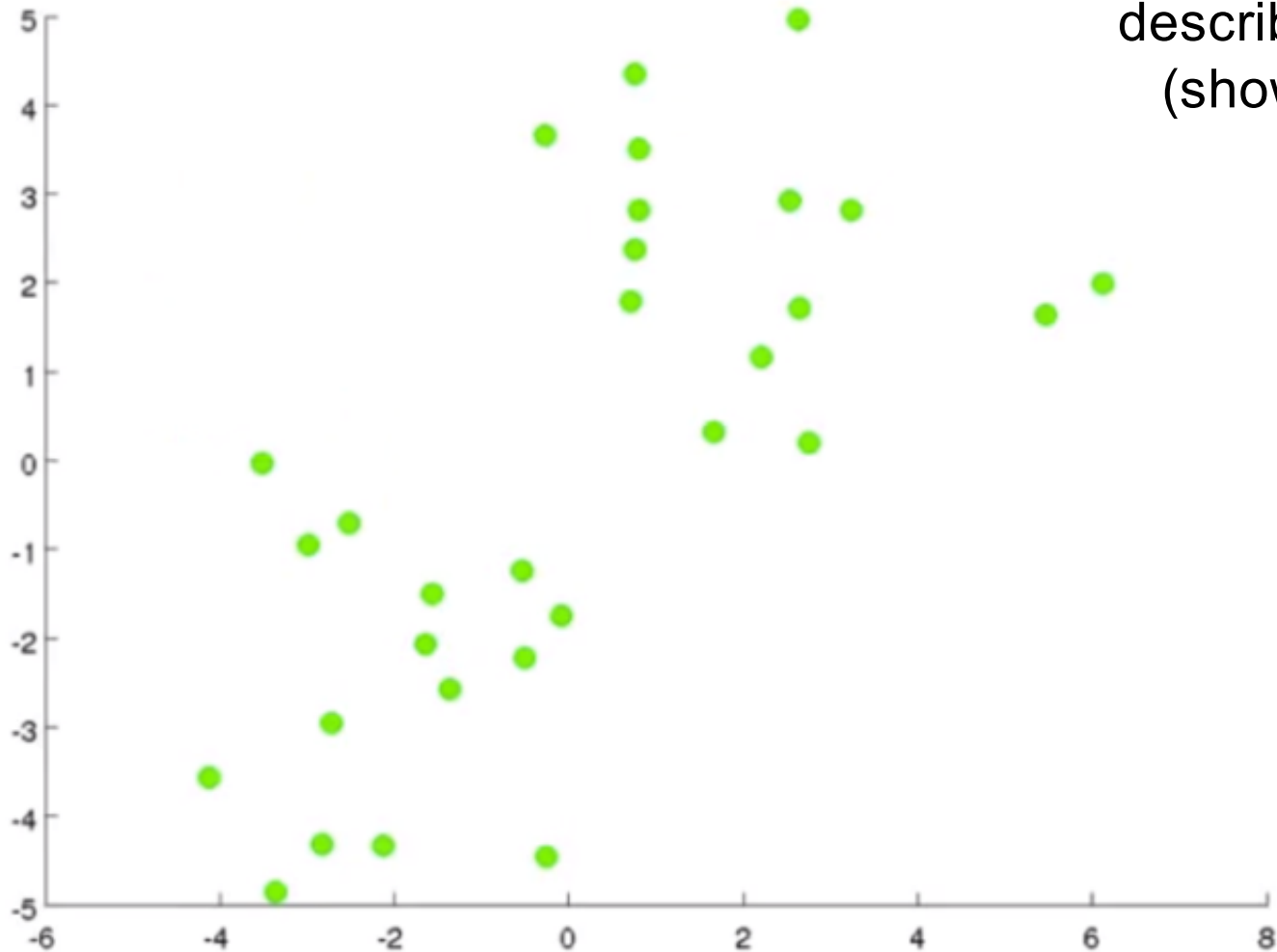
K-MEANS

One of the simplest and widely used clustering algorithm

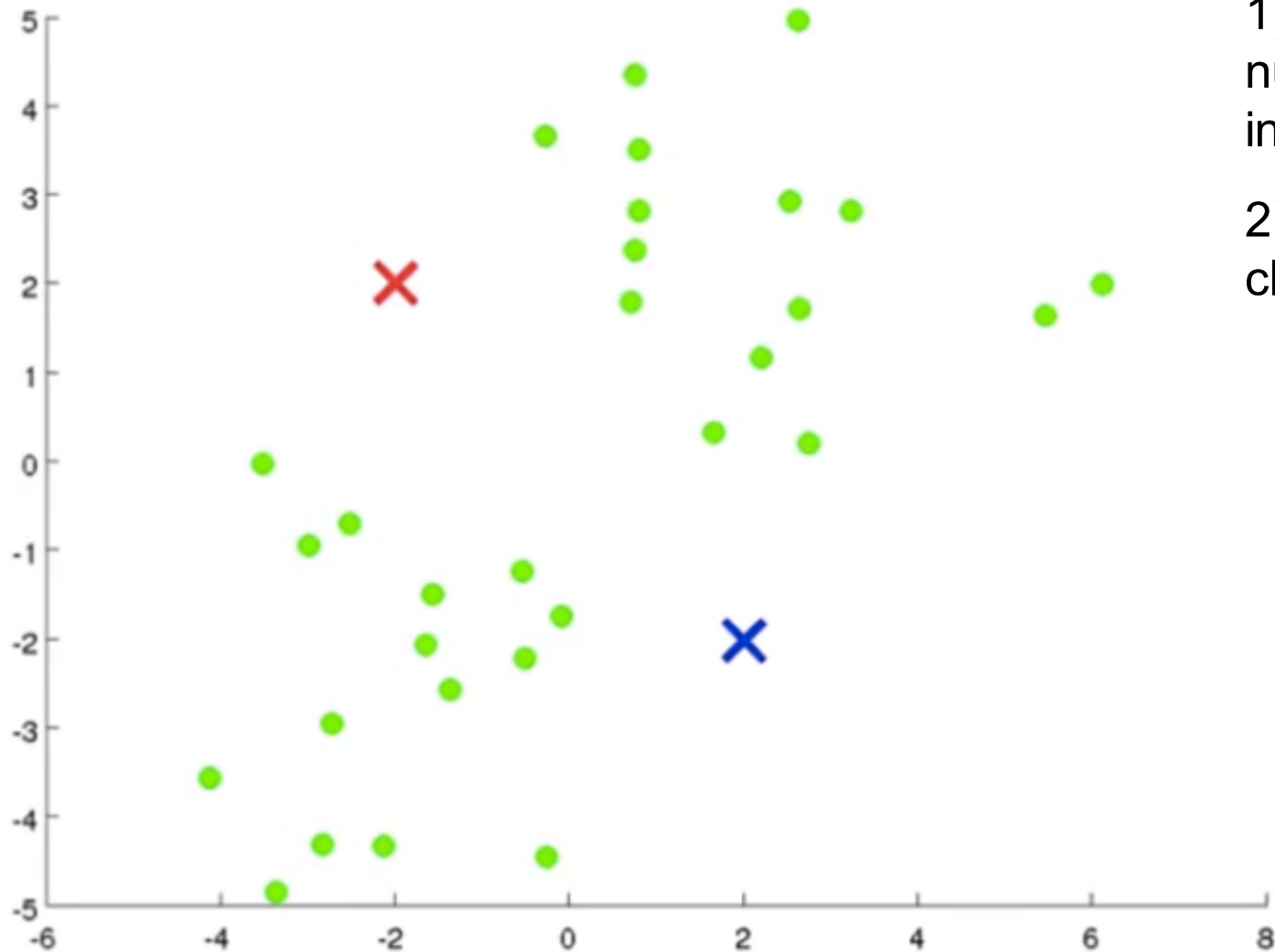
It can be best understood through examples, so we will first have a look at an example

K-MEANS: AN EXAMPLE

Let's suppose we have a dataset with instances described with 2 attributes (shown on x and y axes)



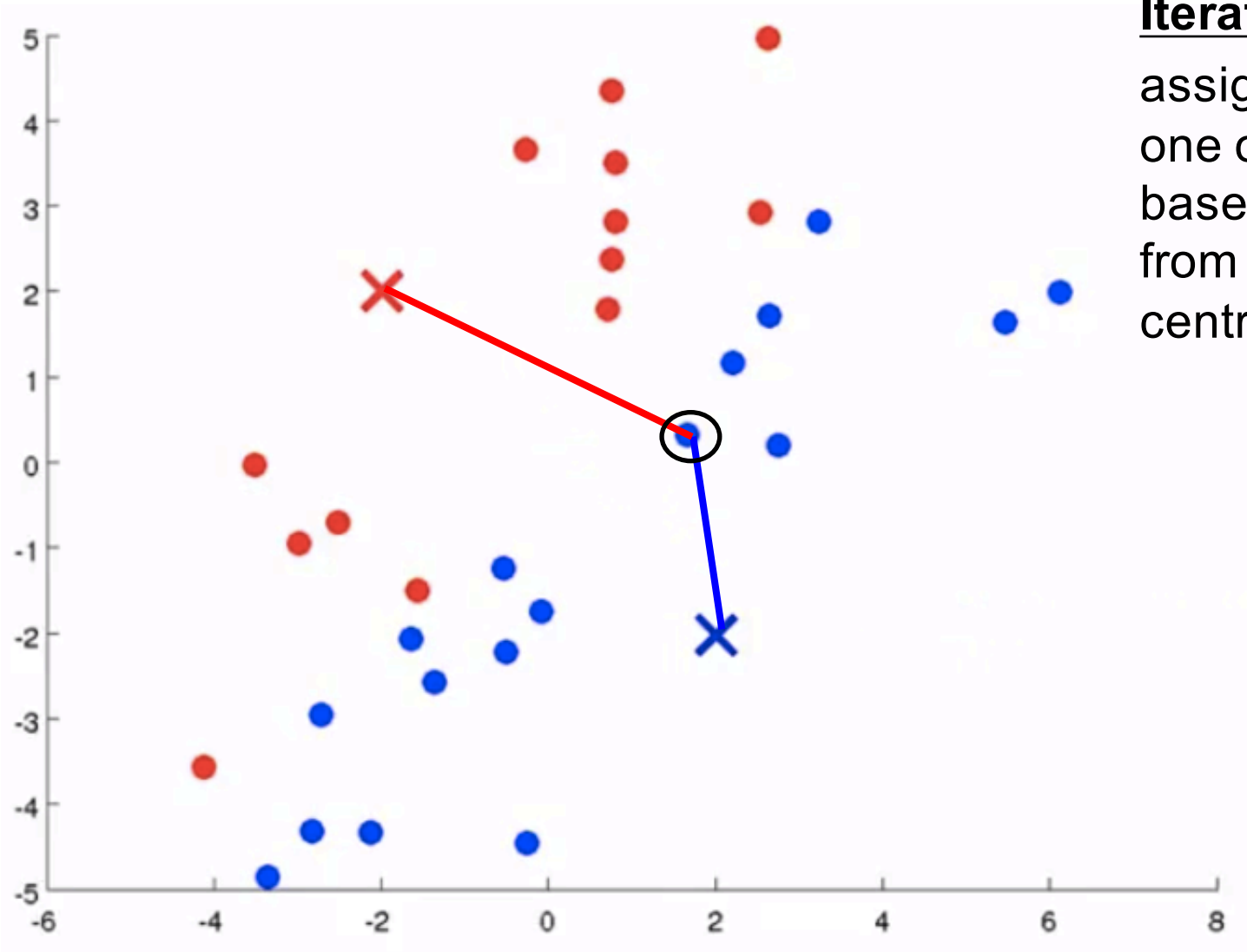
K-MEANS: AN EXAMPLE



Initialization:

- 1) Choose the number of clusters, in this case $K=2$
- 2) Randomly select cluster centroids

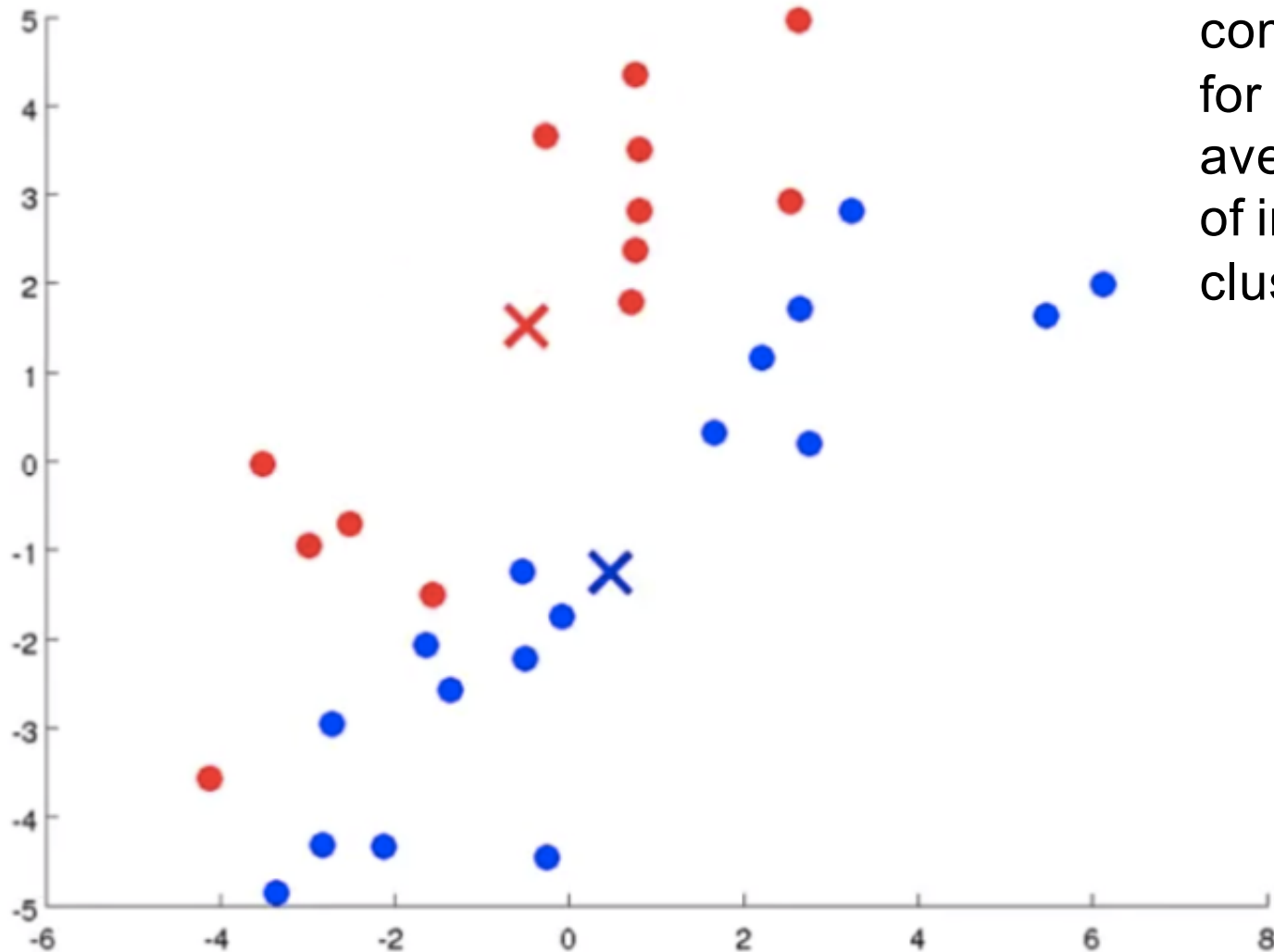
K-MEANS: AN EXAMPLE



Iteration 1, Step 1:

assigning instances to one of the clusters based on their distance from the clusters' centroids

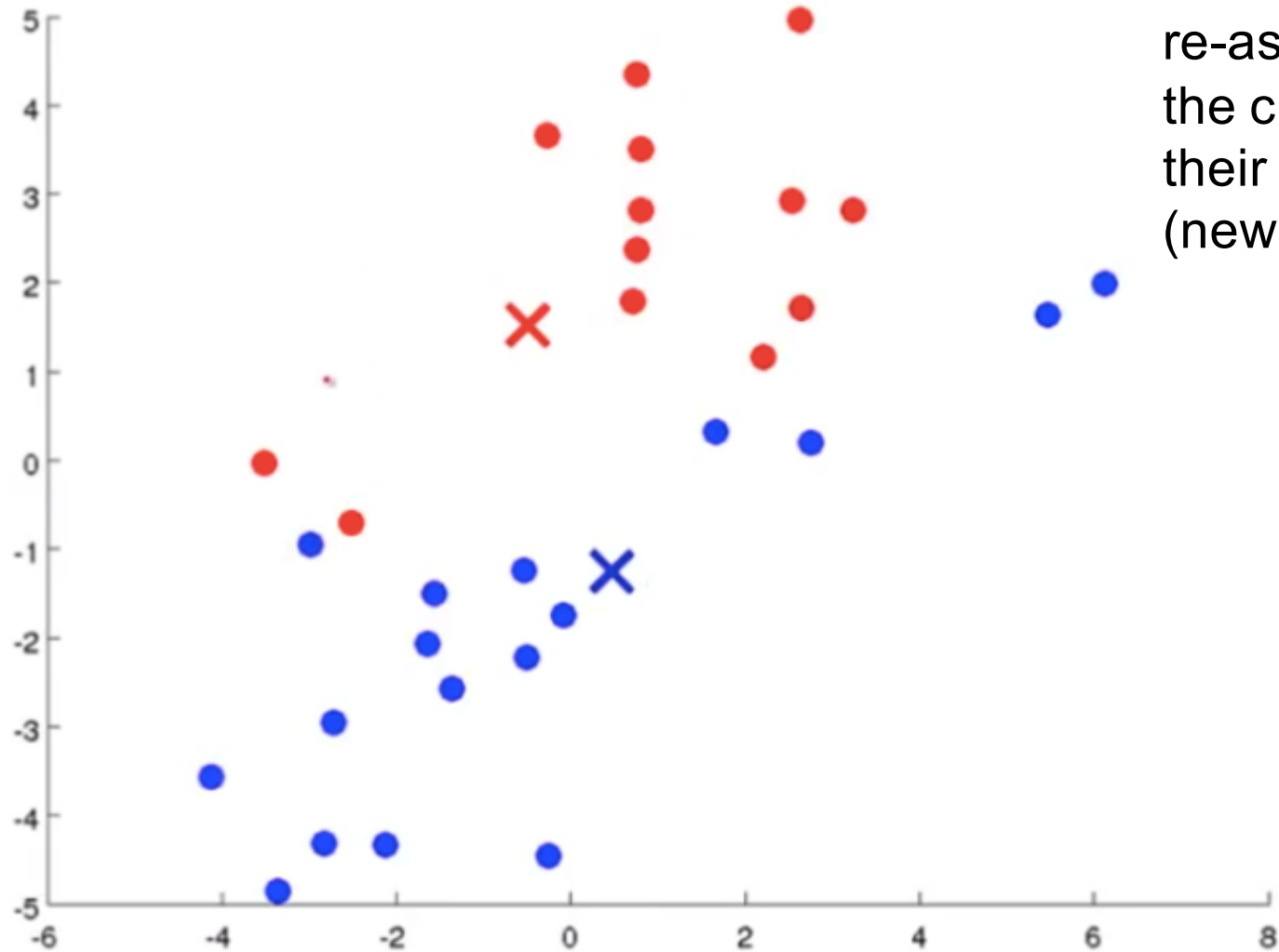
K-MEANS: AN EXAMPLE



Iteration 1, Step 2:

compute new centroid for each cluster, by averaging the values of instances within the cluster

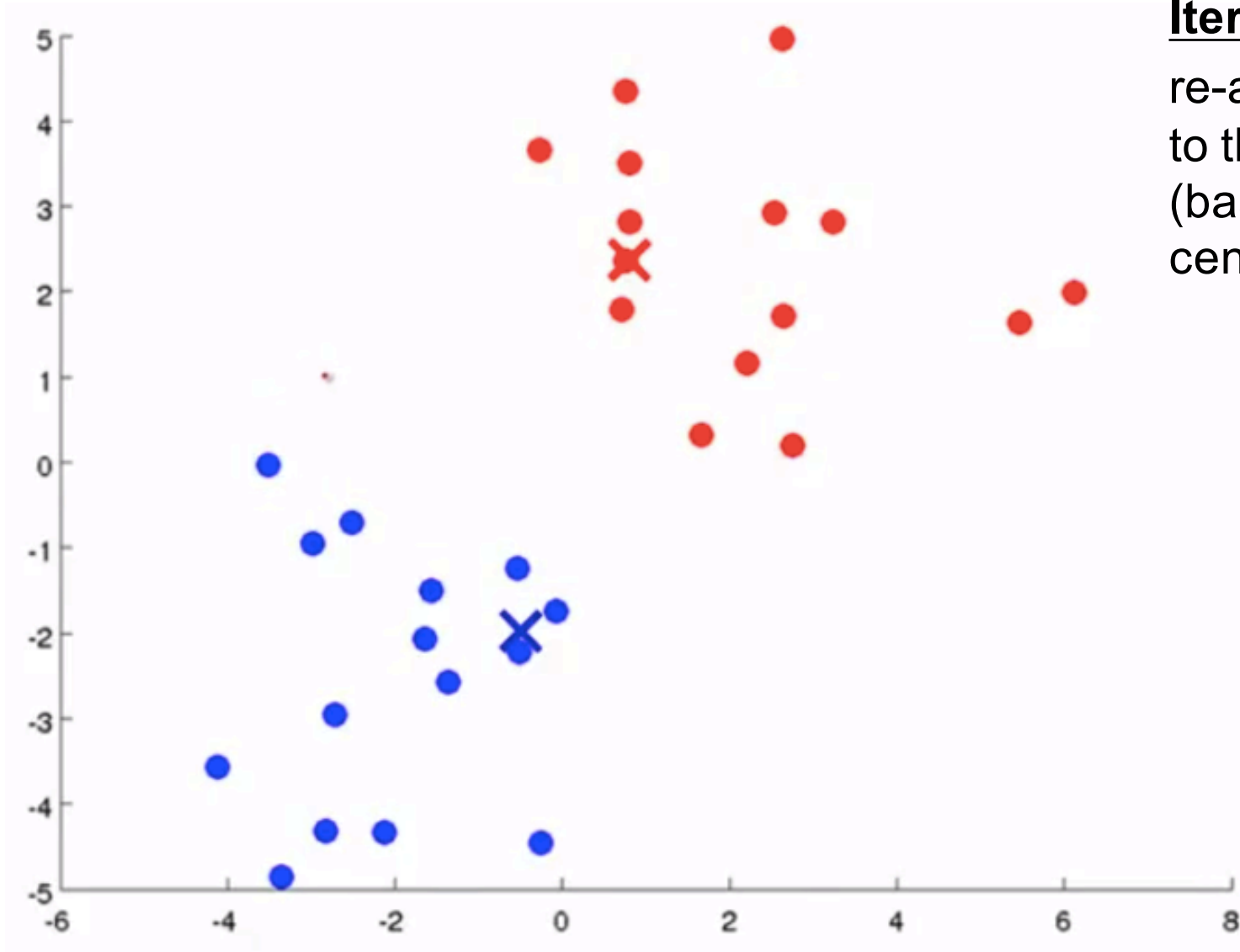
K-MEANS: AN EXAMPLE



Iteration 2, Step 1:

re-assign instances to the clusters based on their distance from the (new) cluster centroids

K-MEANS: AN EXAMPLE

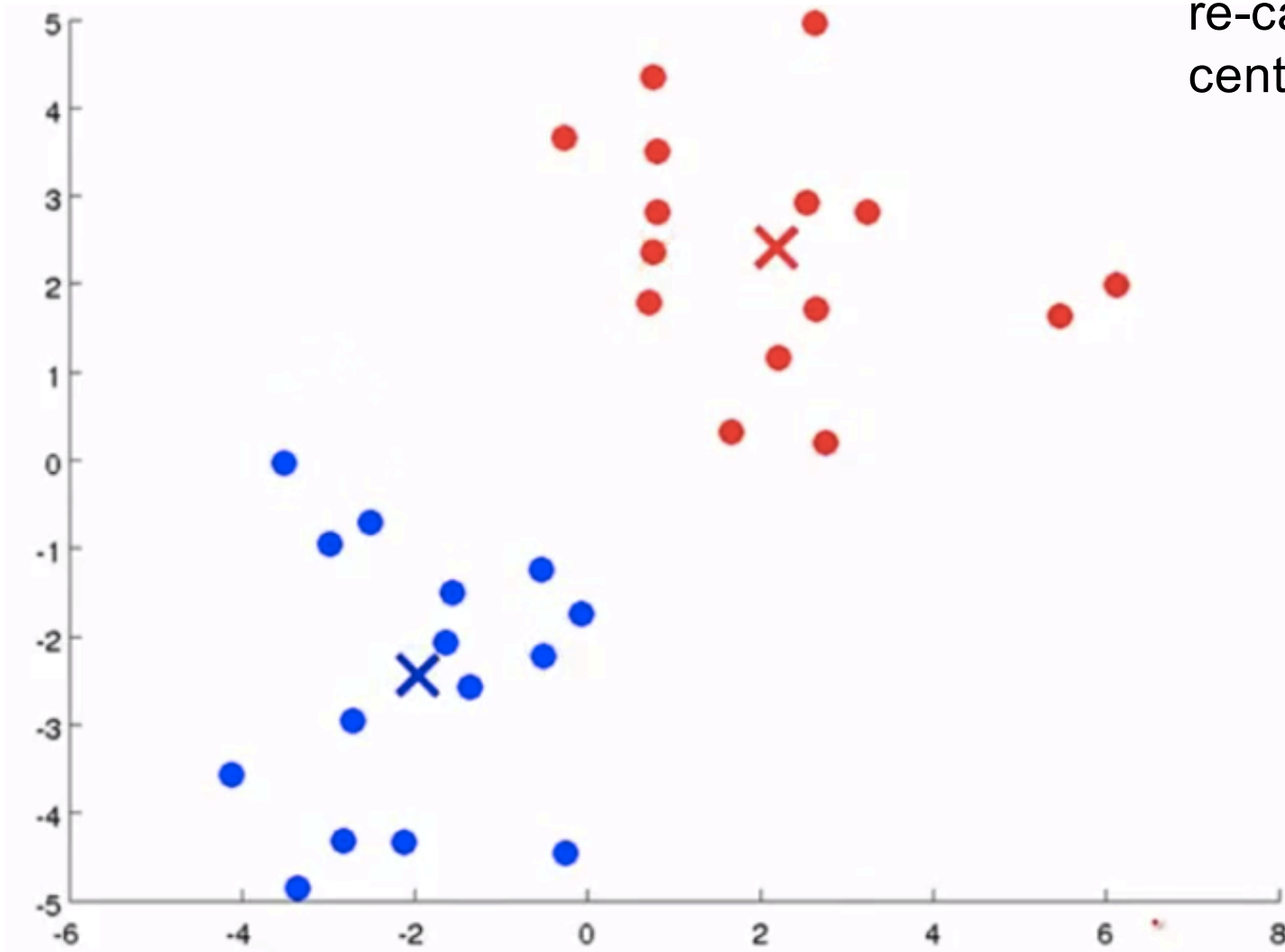


Iteration 3, Step 1:
re-assign instances
to the clusters
(based on the new
centroids)

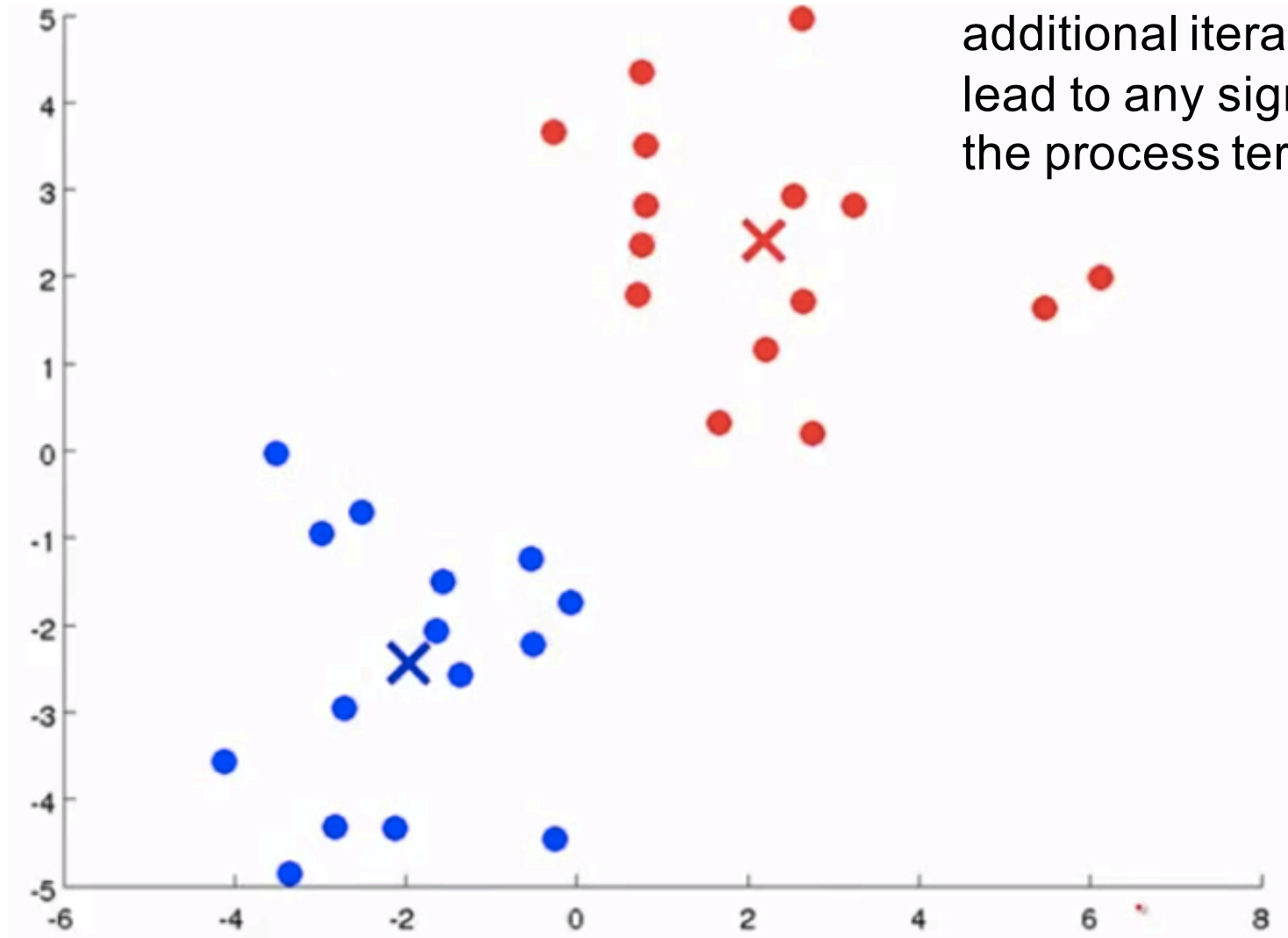
K-MEANS: AN EXAMPLE

Iteration 3, Step 2:

re-calculate cluster centroids



K-MEANS: AN EXAMPLE



The algorithm is converging:
additional iterations will not
lead to any significant change;
the process terminates

K-MEANS: THE ALGORITHM

Input:

- K – the number of clusters
- data set with m instances; each instance in this set is described with a vector of n attributes (x_1, x_2, \dots, x_n)
- max - max number of iterations (optional parameter)

K-MEANS: THE ALGORITHM

Steps:

- 1) Initial, random selection of a centroid for each cluster
 - centroids are chosen from the given dataset, i.e., K instances are randomly taken from the dataset and set as centroids
 - 2) Repeat:
 - 1) *Cluster assignment*: for each instance i from the dataset, $i = 1, m$, identify the closest centroid and assign the instance to the corresponding cluster
 - 2) *Repositioning of centroids*: for each cluster, compute a new centroid by averaging the values of instances assigned to that cluster
- until the algorithm starts converging or the number of iterations reaches *max*

K-MEANS: THE COST FUNCTION

The objective of the K-means algorithm is to *minimize the cost function J*:

$$J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K) = \frac{1}{m} \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

$x^{(i)}$ – i -th instance in the training dataset, $i=1, m$

$\mu_{c^{(i)}}$ – centroid of the cluster to which the instance $x^{(i)}$ has been assigned

$c^{(i)}$ – index of the cluster to which the instance $x^{(i)}$ is currently assigned

μ_j – centroid of the cluster j , $j=1, K$

This function is also known as *distortion function*

K-MEANS: THE COST FUNCTION

$$\min_{\substack{c^{(1)}, \dots, c^{(m)}, \\ \mu_1, \dots, \mu_K}} J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_K)$$

K-means algorithm minimizes the cost function \mathbf{J} in the following manner:

- the *Cluster assignment* phase minimizes \mathbf{J} with respect to $c^{(1)}, \dots, c^{(m)}$, holding μ_1, \dots, μ_K fixed
- the *Repositioning of centroids* phase minimizes \mathbf{J} with respect to μ_1, \dots, μ_K , holding $c^{(1)}, \dots, c^{(m)}$ fixed

K-MEANS: EVALUATION

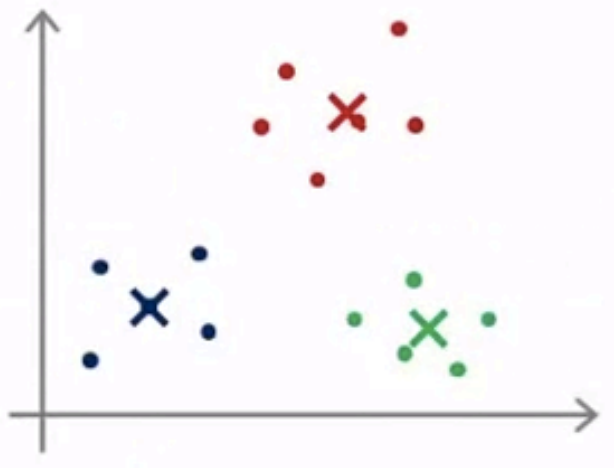
Criteria for evaluating the quality of the resulting clusters:

- Distance between the centroids
 - the more distant the centroids are, the lower is the overlap between the clusters, and the quality of the clusters is considered higher
- St. deviation of instances from the centroid
 - the lower the st. deviation of instances from the cluster centroid, the more tightly they are grouped, and clusters are considered better
- Within cluster sum of squares
 - sum of squared differences between individual data points in a cluster and the cluster center; the smaller, the better

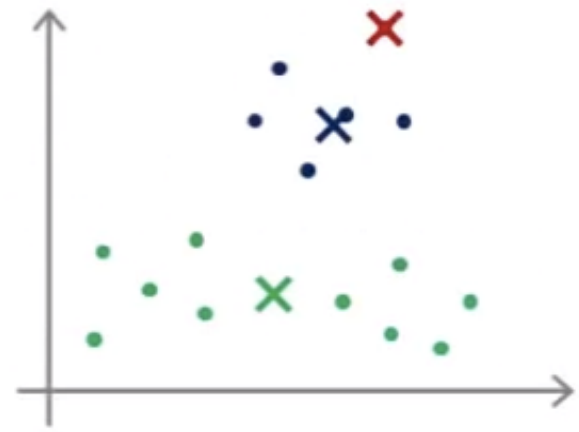
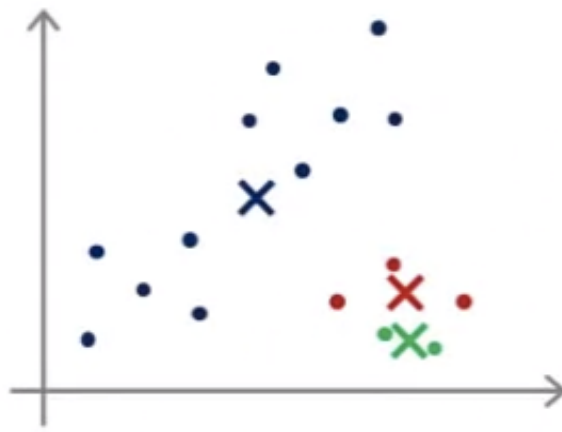
K-MEANS:

INITIAL SELECTION OF CENTROIDS

- Depending on how initial cluster centroids were chosen, the K-means algorithm would converge quicker or slower
- “Unlucky” selection of initial centroids may lead K-Means to get stuck in the so called *local optima* and produce poor results
 - this is a local minimum of the *cost function*



“Lucky” initialization



“Unlucky” initializations that lead to a local minimum

K-MEANS:

MULTIPLE RANDOM INITIALIZATIONS

It allows for avoiding situations that lead K-means in a local minimum

Consists of the following:

```
for i = 1 to n { //n is often in the range 50-1000
  Randomly select the initial set of centroids;
  Apply the K-Means algorithm;
  Compute the cost function
}
```

Choose the instance of the algorithm that produces the lowest value of the cost function

This approach gives good results if the number of clusters is relatively low (2 - 10); should not be used if the number of clusters is higher

Another option: [K-means++ algoritam](#)

K-MEANS: HOW TO CHOOSE K ?

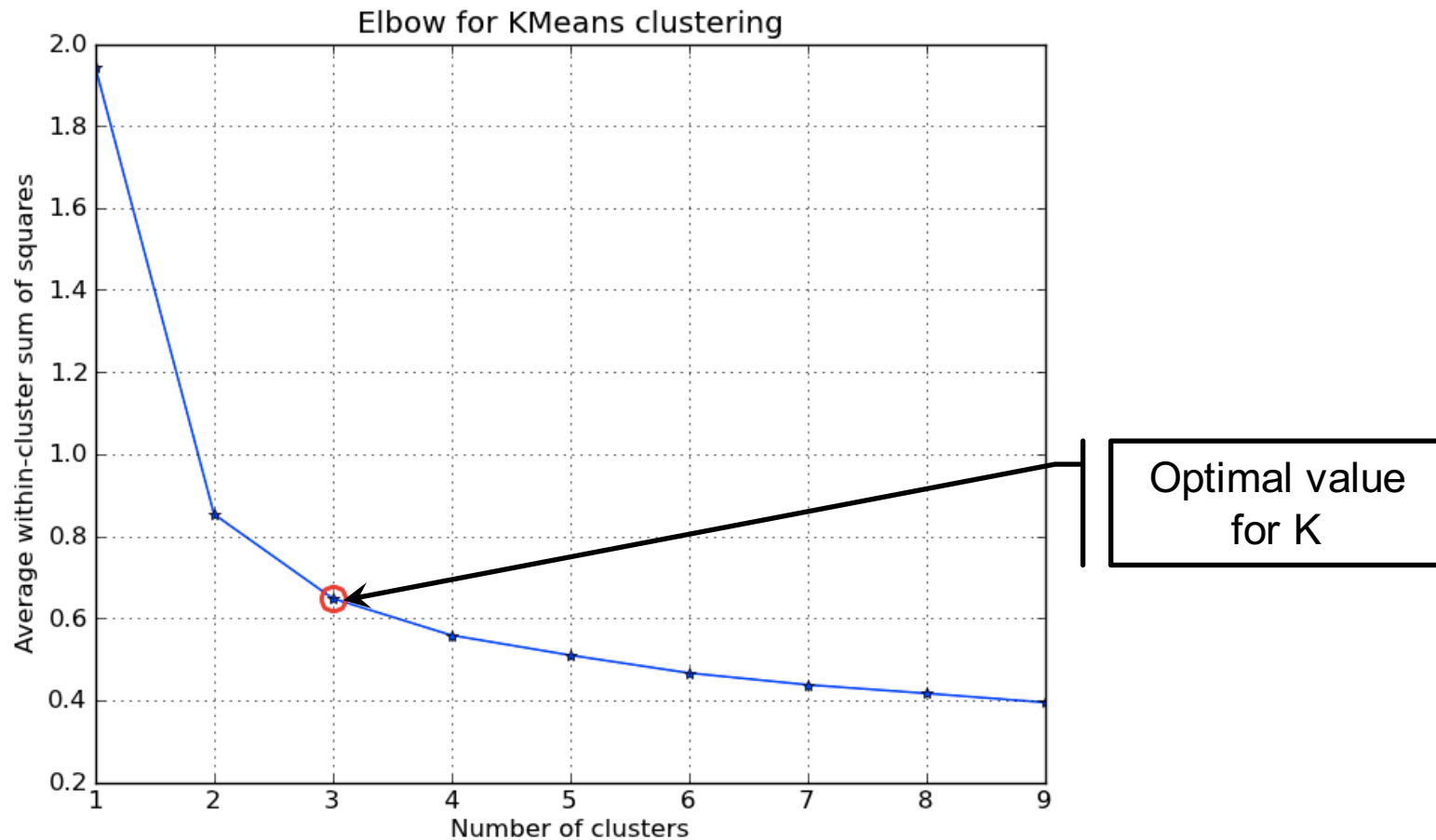
How to determine the number of clusters K?

- In case we have domain knowledge about the phenomenon described by the data
 - Make an assumption about the number of clusters (K) based on the domain knowledge
 - Test the models with K-1, K, K+1 clusters and compare the error*
- If we lack domain knowledge about the studied phenomenon
 - Start with a small number of clusters, and in multiple iterations test multiple models, where each model will have one cluster more than the previous one
 - In each iteration, compare the error* of the current and the previous model, and when the error reduction becomes insignificant, terminate the process

*E.g., within cluster sum of squared errors can be used for the comparison

K-MEANS: HOW TO CHOOSE K ?

When we lack domain knowledge about the studied phenomenon



EXPECTATION
MAXIMIZATION (EM)
ALGORITHM



PROBABILISTIC CLUSTERING

EM is used for probabilistic clustering

From a probabilistic perspective

- instances should not be placed categorically in one cluster or the other,
- instead, they have a certain probability of belonging to each cluster

The rationale: no finite amount of evidence is enough to make a completely firm decision on how to do the clustering

PROBABILISTIC CLUSTERING

This type of clustering has the following assumptions:

- Each cluster is described by probability distribution(s)
 - there might be one distribution, common to all the attributes describing cluster members, or multiple distributions, one per each attribute
 - these distributions determine the probability of attribute values for the cluster members (instances)
- In addition, not all clusters have equal likelihood: there is a probability distribution that reflects the clusters' *prior* probability

PROBABILISTIC CLUSTERING

Let's consider the simplest form of probabilistic clustering:

- instances are described with just one numeric attribute that is Normally distributed in all clusters (K clusters)
- each cluster (C_i) has its specific mean (μ_i) and st. deviation (σ_i)
– i.e., specific parameters of the Normal distribution
- p_i is the *prior probability* of the cluster C_i

PROBABILISTIC CLUSTERING

Let's consider the simplest form of probabilistic clustering (cont.):

Suppose we've been given a set of instances that originate from the previously described K clusters; however, we do not know:

- the specific cluster that each instance originates from
- parameters of the model $(\mu_i, \sigma_i, p_i, i=1, K)$.

The task/problem to be solved:

based on the given set of instances, estimate

- the parameters of the model $(\mu_i, \sigma_i, p_i, i=1, K)$
- for each instance, the probability of belonging to each of the K clusters

EM ALGORITHM

To solve the described problem, we can apply a procedure similar to the one used for the K-means algorithm:

- 1) start by defining the number of clusters (K) and randomly choosing the model parameters $(\mu_i, \sigma_i, p_i, i=1, K)$
- 2) for the given parameter values, compute, for each instance, the probability of belonging to each of the K clusters
- 3) use the computed probabilities to re-estimate the parameter values

Repeat steps 2) and 3) until the parameter values start to converge

This procedure is the gist of the EM algorithm

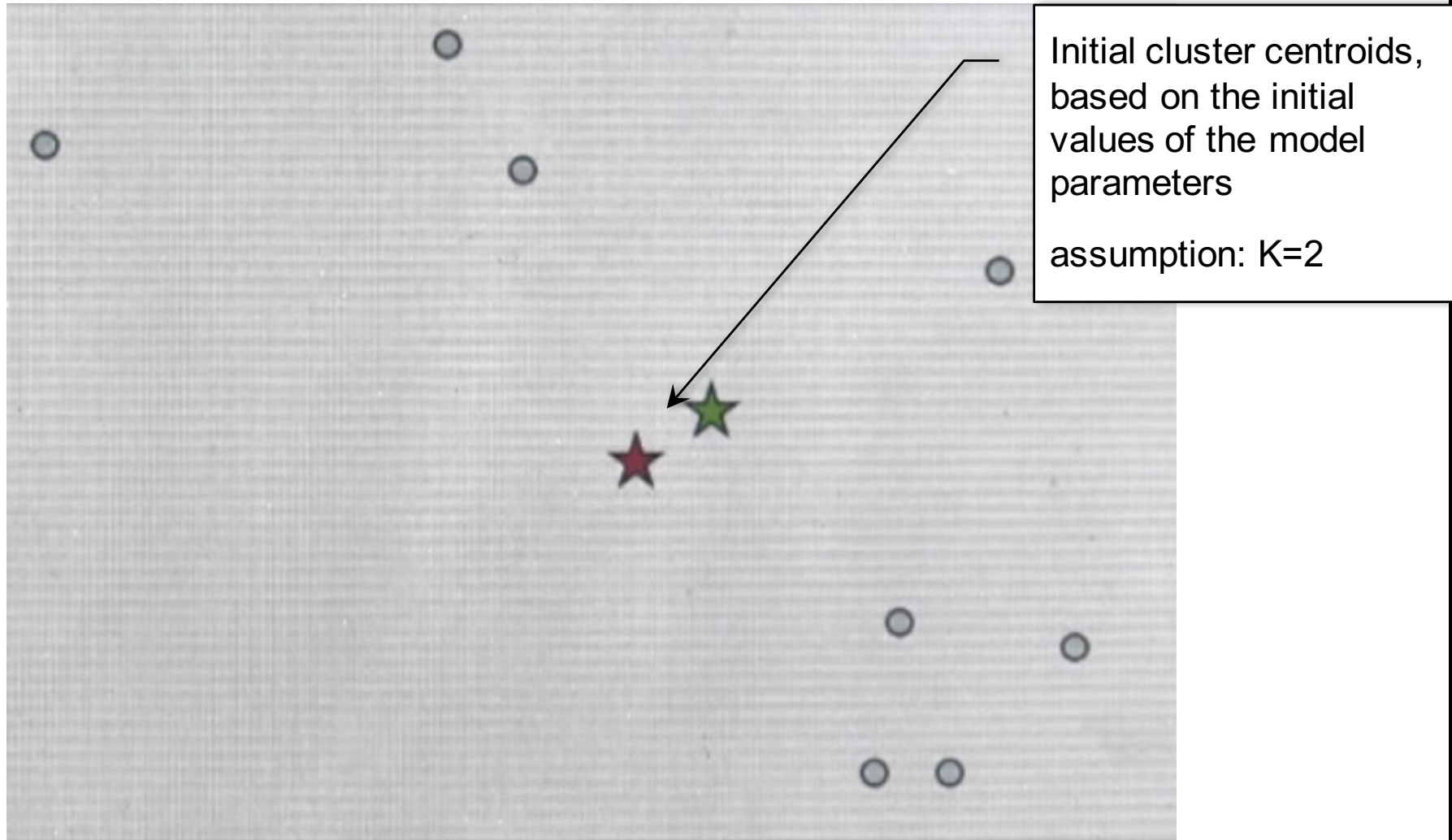
EM ALGORITHM

The EM algorithm consists of 2 key steps:

- **E (expectation) step** – calculation of the cluster probabilities for each instance from the dataset; in this step we assume that we know the values of all the model parameters;
- **M (maximization) step** – calculation of the model parameters; we aim to “maximize” the likelihood of the model given the available data

These steps are repeated until the algorithm starts to converge

EM ALGORITHM: INITIALIZATION



The example is taken from the AI course: <https://www.udacity.com/course/cs271>

EM ALGORITHM: E STEP

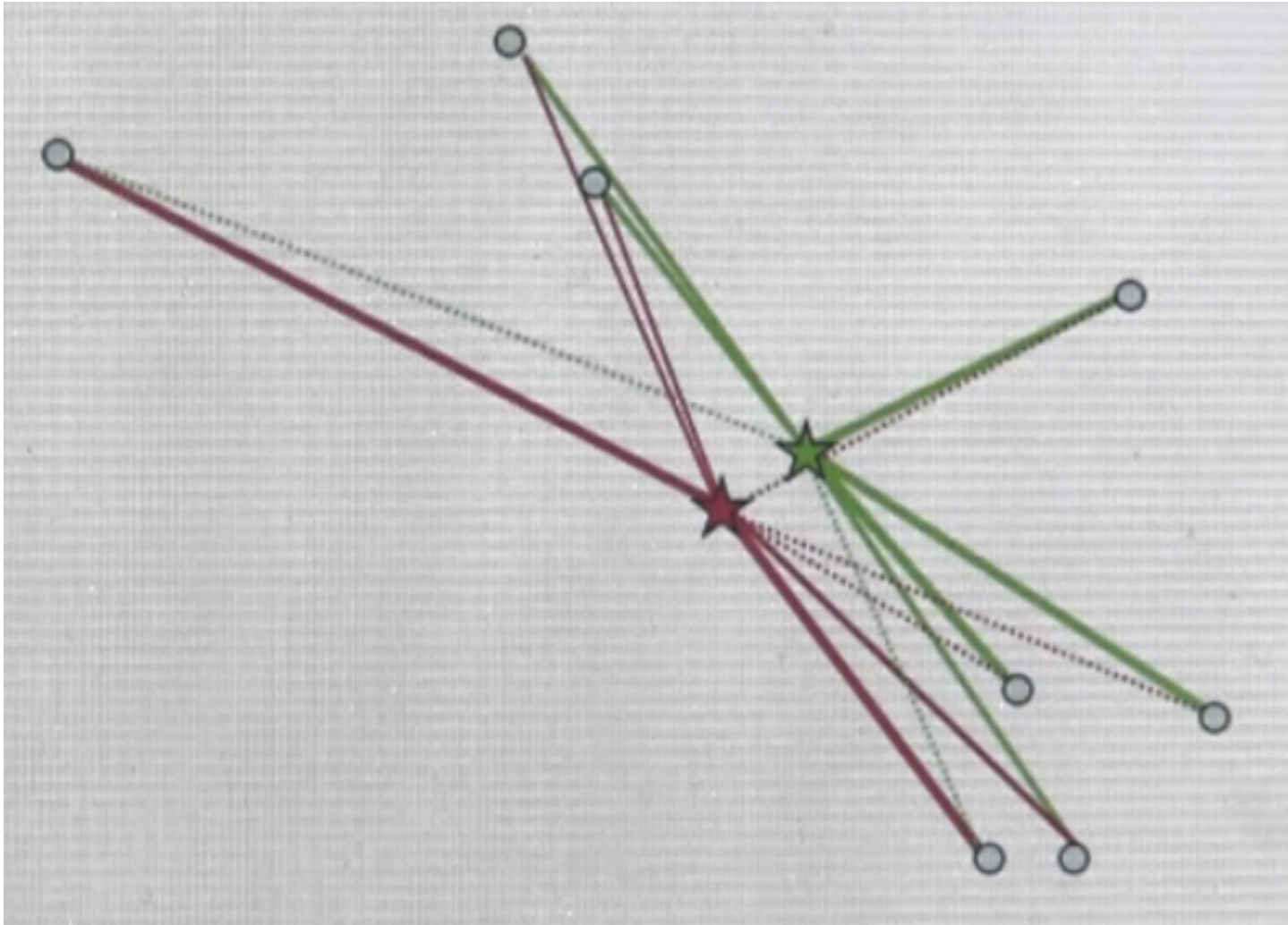
For each instance from the dataset x_j ($j=1,n$), we compute the expectation that it belongs to the cluster C_i ($i=1,K$)

$$e_{ij} = p_i * P(x_j | C_i)$$

$P(x_j | C_i)$ is computed using the probability density formula of the Normal distribution $f(x; \mu, \sigma)$

Reminder: in this step we assume that the values of all the model parameters – $\mu_i, \sigma_i, p_i, i=1,K$ – are known

EM ALGORITHM: E STEP



The thickness of a line indicates the probability that an instance belongs to a certain cluster, i.e., it reflects the computed e_{ij} value

EM ALGORITHM: M STEP

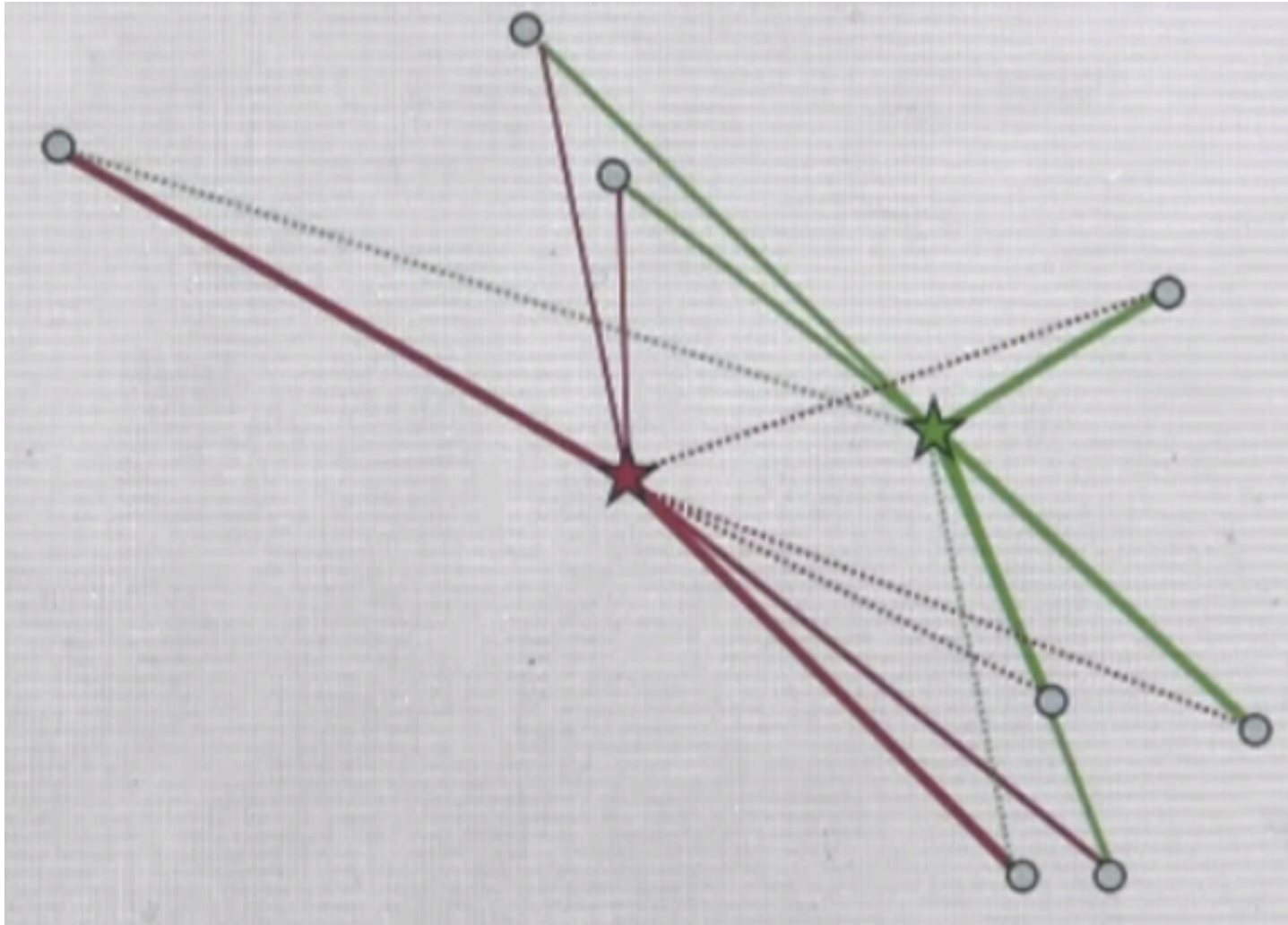
In this step, values of all the model parameters are re-computed

$$\text{prior probability: } p_i = \sum_j \frac{e_{ij}}{n}$$

$$\text{mean: } \mu_i = \frac{\sum_j e_{ij} * x_j}{\sum_j e_{ij}}$$

$$\text{variance: } \sigma_i^2 = \frac{\sum_j e_{ij} * (x_j - \mu_i)^2}{\sum_j e_{ij}}$$

EM ALGORITHM: M STEP



Cluster centroids change their position based on the newly computed values of the model parameters

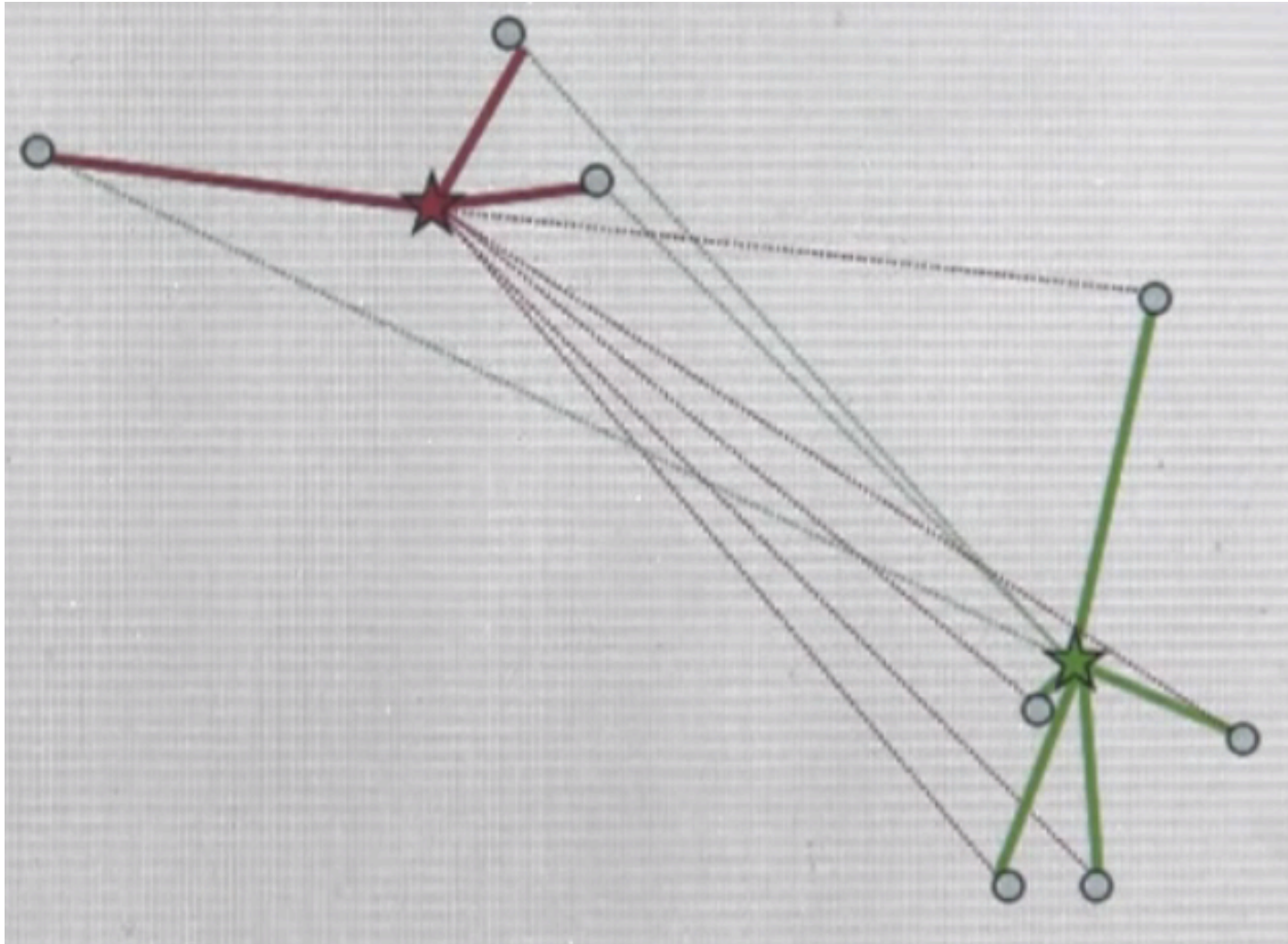
EM ALGORITHM: CONVERGENCE

The two steps of the EM algorithm are repeated until the increase in the overall log-likelihood of the model becomes negligible:

$$\log P(x) = \log \sum_i (p_i * P(x|C_i))$$

Typically, the log-likelihood will increase very sharply over the first few iterations, and then converge rather quickly to a point that is virtually stationary

EM ALGORITHM: CONVERGENCE



The state of convergence of the model parameters

EM ALGORITHM

EM algorithm is guaranteed to converge to a maximum of the log-likelihood function

However, this is a local maximum that may not necessarily be the same as the global maximum

For a higher chance of obtaining the global maximum, the whole procedure should be repeated several times, with different initial guesses for the parameter values

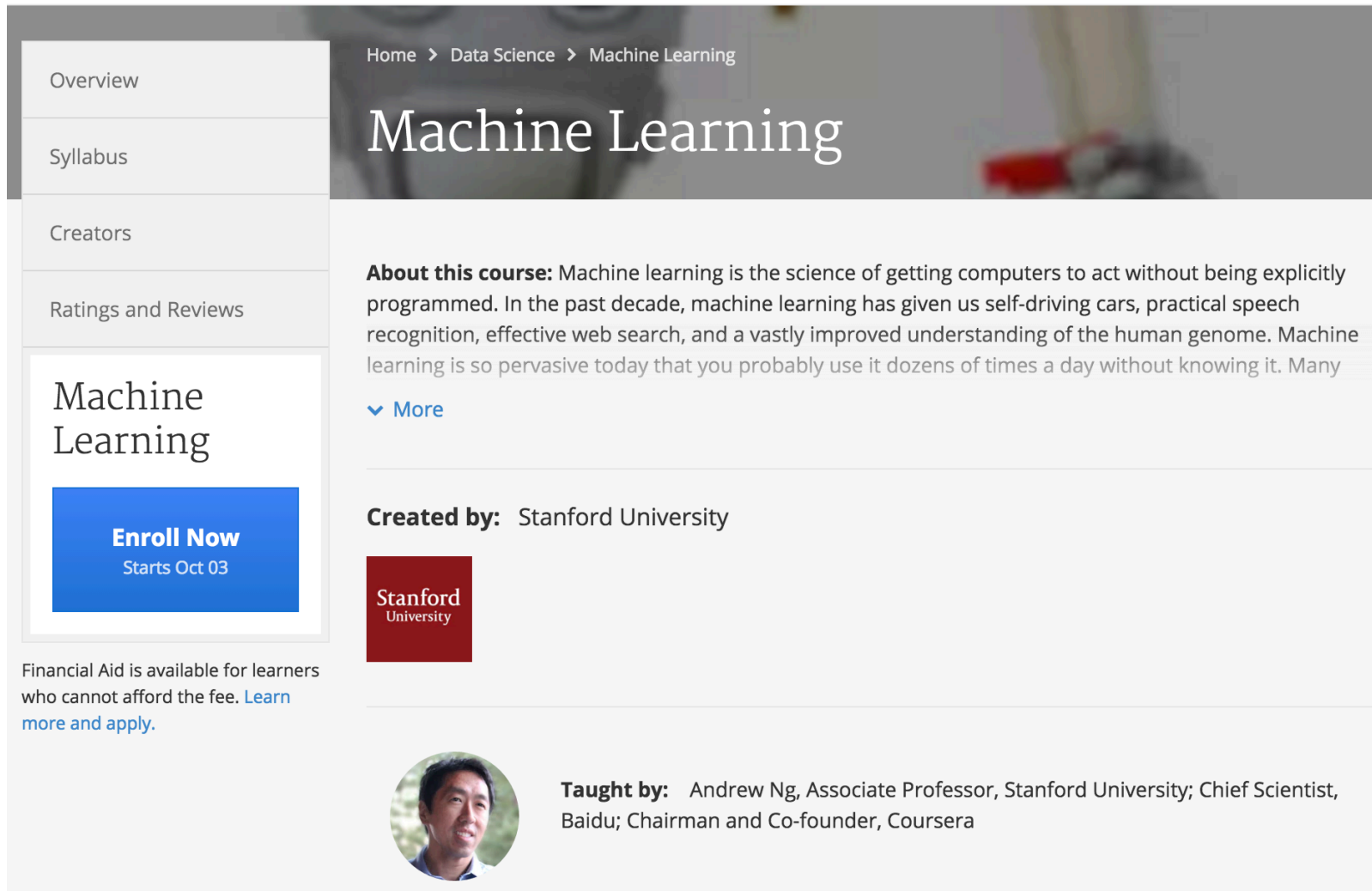
At the end, we choose the configuration that produces the largest overall log-likelihood

EM ALGORITHM

We've considered the simplest EM application case; but EM can be equally well applied to more complex problems

- Instances can be described with more than one numeric attribute as long as independence between attributes is assumed
 - individual probabilities for all the attributes are multiplied to obtain the joint probability for the instance, just as in the Naive Bayes method
- Attributes can be nominal, as well
 - In that case, Normal distribution has to be abandoned
 - nominal attribute with v possible values is characterized by v numbers representing the probability of each value

ACKNOWLEDGEMENT AND RECOMMENDATION



The screenshot shows the Coursera course page for 'Machine Learning'. The page has a navigation menu on the left with options: Overview, Syllabus, Creators, and Ratings and Reviews. The main content area features a breadcrumb trail 'Home > Data Science > Machine Learning' and a large title 'Machine Learning'. Below the title is a paragraph about the course: 'About this course: Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many'. A blue button labeled 'Enroll Now' with 'Starts Oct 03' is visible. Below the button, there is a note about financial aid: 'Financial Aid is available for learners who cannot afford the fee. Learn more and apply.' The course is created by Stanford University, with a Stanford University logo. The instructor is Andrew Ng, with a circular profile picture and text: 'Taught by: Andrew Ng, Associate Professor, Stanford University; Chief Scientist, Baidu; Chairman and Co-founder, Coursera'.

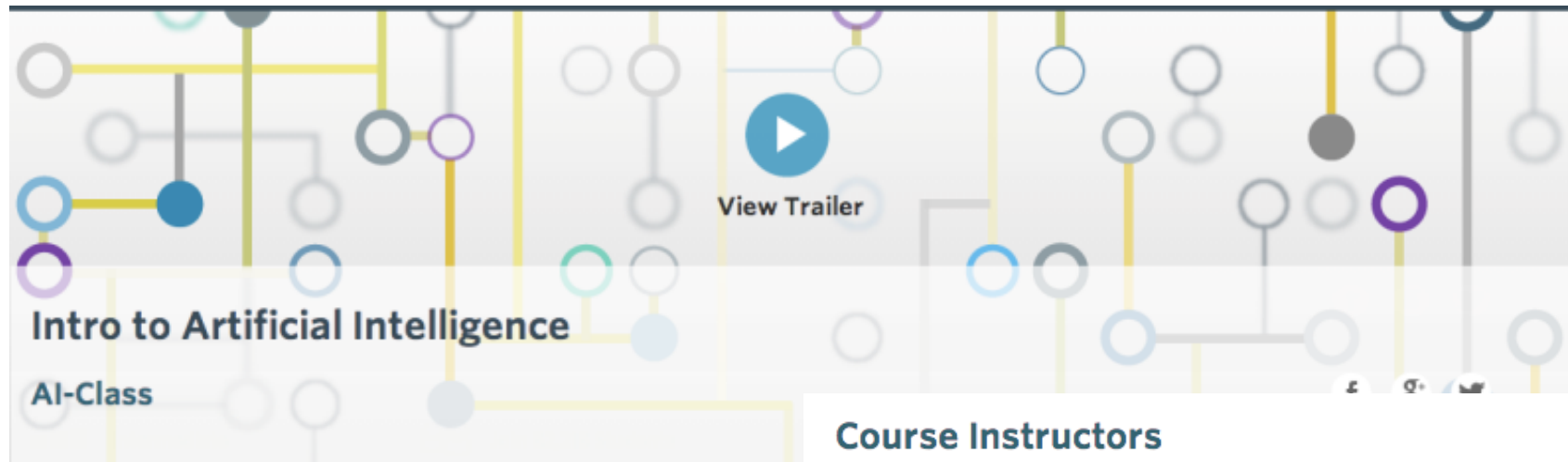
Coursera:

<https://www.coursera.org/learn/machine-learning>

Stanford YouTube channel:

http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599

ACKNOWLEDGEMENT AND RECOMMENDATION



Intermediate

Join 52,215 Students

Class Summary

The objective of this class is to teach you modern AI. You will learn about the basic techniques and tricks of the trade. We also aspire to excite you about the field of AI.

Course Instructors



Peter Norvig

INSTRUCTOR

Peter Norvig is Director of Research at Google Inc. He is also a Fellow of the American Association for Artificial Intelligence and the Association for Computing Machinery. Norvig is co-author of the popular textbook *Artificial Intelligence: A Modern Approach*. Prior to joining Google he was the head of the Computation Sciences Division at NASA Ames Research Center.



Sebastian Thrun

INSTRUCTOR

Sebastian Thrun is a Research Professor of Computer Science at Stanford University, a Google Fellow, a member of the National Academy of Engineering and the German Academy of Sciences. Thrun is best known for his research in robotics and machine learning, specifically his work with self-driving cars.

URL: <https://www.udacity.com/course/intro-to-artificial-intelligence--cs271>

(Anonymous) questionnaire for your critiques, comments, suggestions:

<http://goo.gl/cqdp3l>