Guided Tour of Machine Learning in Finance

Week 3: Unsupervised Learning

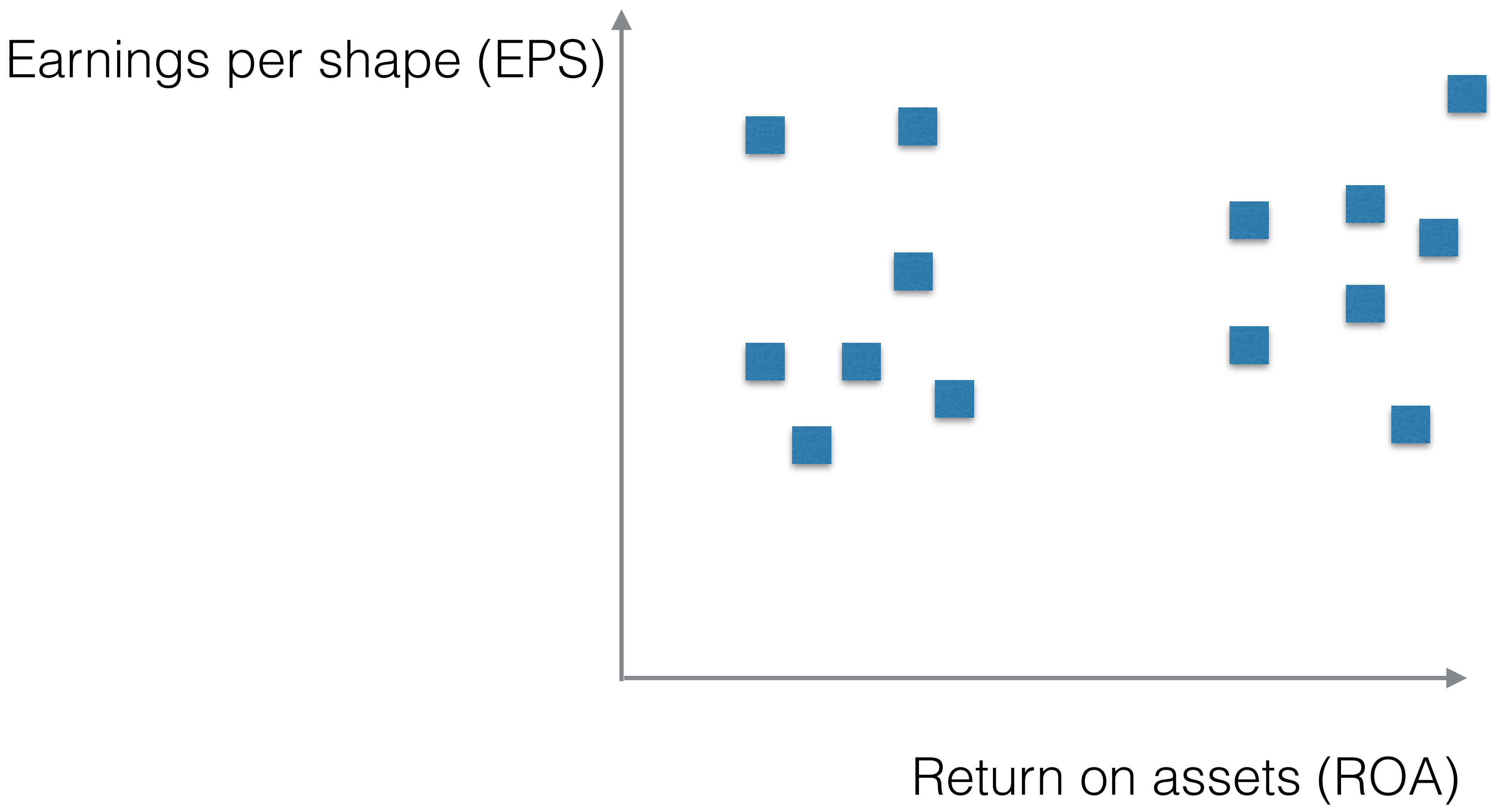
3-2-2: K-means clustering

Igor Halperin

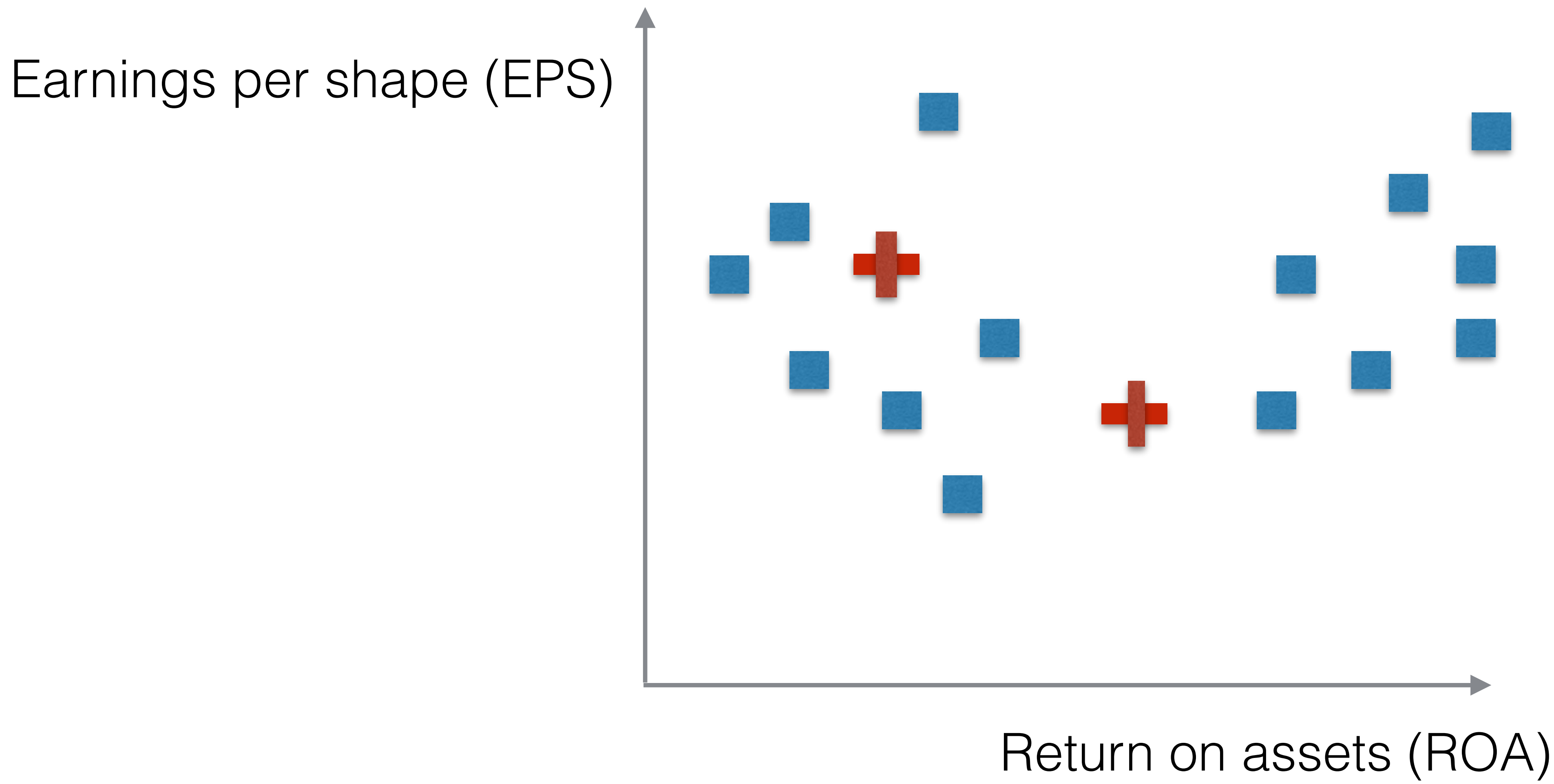
NYU Tandon School of Engineering, 2017

K-means clustering

Cluster a two-dimensional view of a set of companies



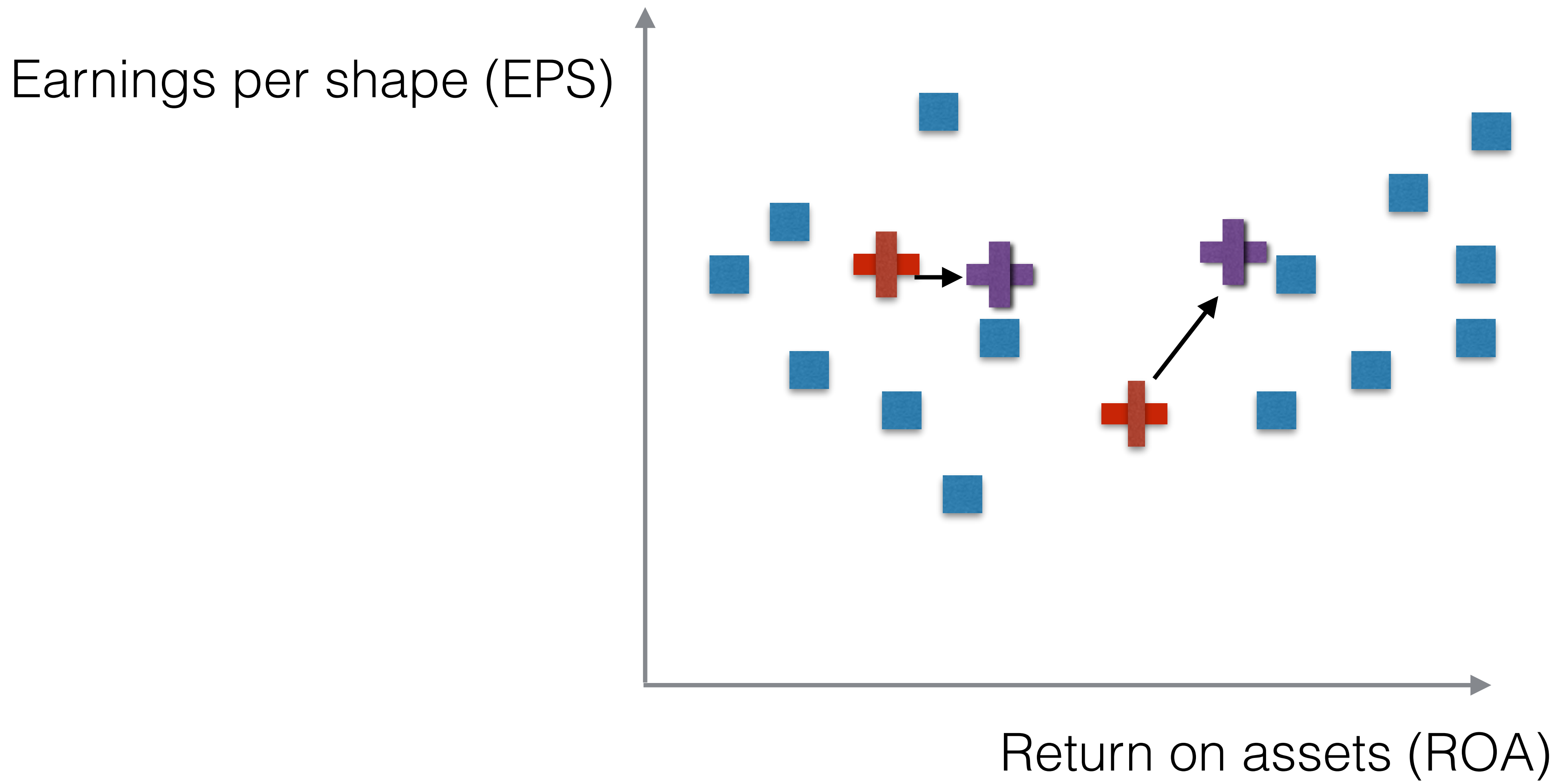
K-means clustering



1. Initialization:

- choose $K: \ K = 2$
- choose K random points in the input space
- assign the cluster centers μ_k to these points

K-means clustering: learning

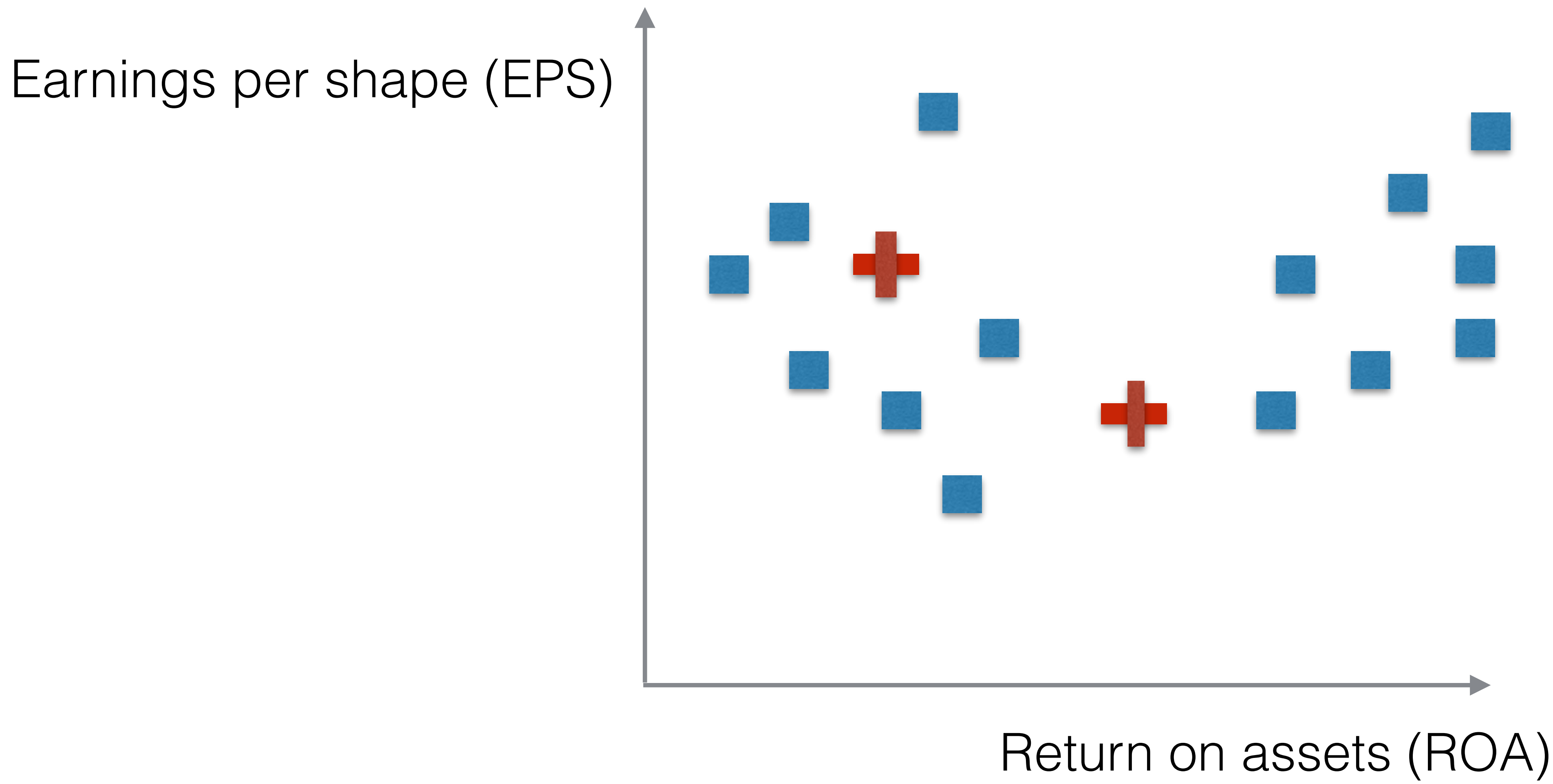


2. Learning:

- repeat
  - for each datapoint $\mathbf{x}_i$:
    - compute the distance $d(\mathbf{x}_i, \mu_k)$ to each cluster center μ_k
    - assign the datapoint to cluster C_k with the smallest distance $d(\mathbf{x}_i, \mu_k)$
  - for each cluster C_k, re-compute its centroid

$$\mu_k = \frac{1}{N_{C_k}} \sum_{i \in C_k} \mathbf{x}_i$$

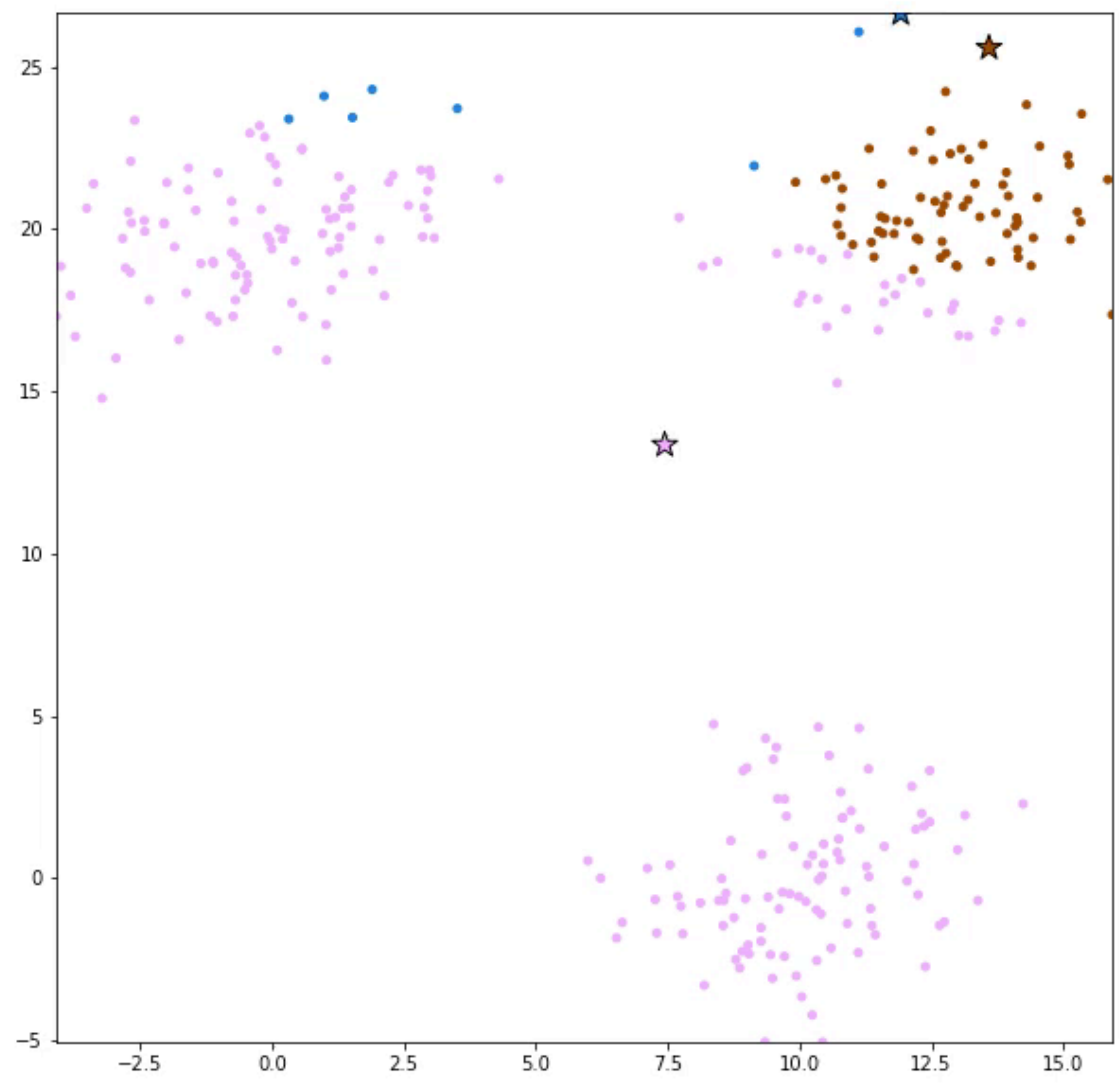
K-means clustering: usage



2. Usage of a trained model:

- for each test datapoint $\mathbf{x}_i$:
  - compute the distance $d(\mathbf{x}_i, \mu_k)$ to each cluster center μ_k
  - assign the datapoint to cluster C_k with the smallest distance $d(\mathbf{x}_i, \mu_k)$

K-means for analysis of companies



For a well-separated sets of points, the K-means finds clusters in a small number of steps.

K-means: energy-based formulation

We can alternatively re-formulate the K-means algorithm as the problem of minimization of the following loss function (distortion function):

$$L = \min_{r_{nk}\,,\,\mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \mu_k \right\|^2$$

where $r_{nk} = 1 \; if \; \mathbf{x}_i \in C_k \, , else \; r_{nk} = 0$

Minimization of the loss function:
- choose initial values of $\{\mu_k\}$
- repeat:
    - Minimize w.r.t. $\{r_{nk}\}$ for fixed values of $\{\mu_k\}$
    - Minimize w.r.t. $\{\mu_k\}$ for fixed values of $\{r_{nk}\}$

K-means: energy-based formulation

We can alternatively re-formulate the K-means algorithm as the problem of minimization of the following loss function (distortion function):

$$L = \min_{r_{nk},\, \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \mu_k \right\|^2$$

where $r_{nk} = 1 \; if \; \mathbf{x}_i \in C_k, \, else \; r_{nk} = 0$

Expectation Maximization (**EM**) algorithm

Minimization of the loss function:
- choose initial values of centroids $\{\mu_k\}$
- repeat:
    - Minimize w.r.t. $\{r_{nk}\}$ for fixed values of $\{\mu_k\}$ (**M-step**)
    - Minimize w.r.t. $\{\mu_k\}$ for fixed values of $\{r_{nk}\}$ (**E-step**)

K-means: energy-based formulation

We can alternatively re-formulate the K-means algorithm as the problem of minimization of the following loss function (distortion function):

$$L = \min_{r_{nk},\, \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \mu_k \right\|^2$$

where $r_{nk} = 1 \; if \; \mathbf{x}_i \in C_k, \, else \; r_{nk} = 0$

Expectation Maximization (**EM**) algorithm

Minimization of the loss function:
- choose initial values of centroids $\{\mu_k\}$
- repeat:
    - Minimize w.r.t. $\{r_{nk}\}$ for fixed values of $\{\mu_k\}$ (**M-step**)
    - Minimize w.r.t. $\{\mu_k\}$ for fixed values of $\{r_{nk}\}$ (**E-step**)

M-step: $r_{nk} = 1 \; if \; k = \arg\min_j \left\| \mathbf{x}_n - \mu_j \right\|^2; \; and \; 0 \; otherwise, \forall n$

K-means: energy-based formulation

We can alternatively re-formulate the K-means algorithm as the problem of minimization of the following loss function (distortion function):

$$L = \min_{r_{nk},\, \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \mu_k \right\|^2$$

where $r_{nk} = 1 \; if \; \mathbf{x}_i \in C_k, \, else \; r_{nk} = 0$

Expectation Maximization (**EM**) algorithm

Minimization of the loss function:
- choose initial values of centroids $\{\mu_k\}$
- repeat:
    - Minimize w.r.t. $\{r_{nk}\}$ for fixed values of $\{\mu_k\}$ (**M-step**)
    - Minimize w.r.t. $\{\mu_k\}$ for fixed values of $\{r_{nk}\}$ (**E-step**)

M-step: $r_{nk} = 1 \; if \; k = \arg\min_j \left\| \mathbf{x}_n - \mu_j \right\|^2; \; and \; 0 \; otherwise, \forall n$

E-step: $$\frac{\partial L}{\partial \mu_k} = 0 \quad \Rightarrow \mu_k = \frac{\sum_{n=1} r_{nk} \mathbf{x}_n}{\sum_{n=1} r_{nk}}, \, k = 1, \ldots, K$$

K-means: notes on the algorithm

- The K-means algorithm is **deterministic**: repeated experiments with the same datasets and the same initial positions of centroids give the same result
- The optimization problem of loss minimization in the K-means algorithm is **non-convex**: repeated experiments with the same datasets but different initial positions of centroids generally give different results
- **Cluster stability**: clustering results may vary for a different dataset from the same data-generating distribution. Centroid positions are sensitive to outliers
- Control of **convergence speed and local minima**: more advanced initialization algorithms, e.g. the K++ Means.
- **Inputs for K-means**: data normalization and/or whitening can help the algorithm
- For large datasets: **Mini-Batch** or **Online K-means** clustering
- **Fake clusters**: the K-means will find clusters even when there are no clusters (e.g. the data is white noise)
- What is the **right value of K** to choose?
- Why the **Euclidean distance** as a measure of similarity?

Control question

Select all correct answers

1. The name “K-means” means that the number of clusters in this algorithm is always set to be K times the mean of the data.
2. The K-means clustering is a Probabilistic (Soft) clustering, where cluster probabilities are estimated empirically using different initializations.
3. The K-means algorithm has a unique solution due to a strict convexity of its objective function.
4. The K-means is simultaneously a Flat and Hard deterministic clustering algorithm, whose solution is non-unique due to non-convexity of its loss function.

Correct answer: 4