



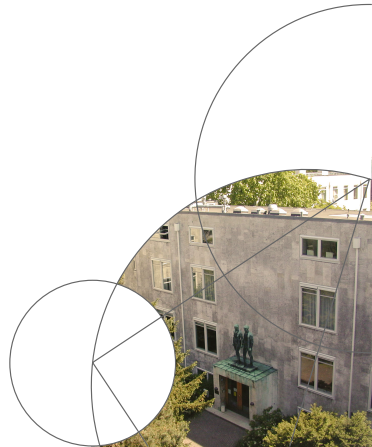
Faculty of Science



k -means Clustering

Data Science

Christian Igel
Department of Computer Science



Outline

- ➊ Unsupervised Learning
- ➋ Clustering
- ➌ k -means Clustering
- ➍ Deriving k -means
- ➎ k -means Clustering for Image Segmentation
- ➏ Summary



Outline

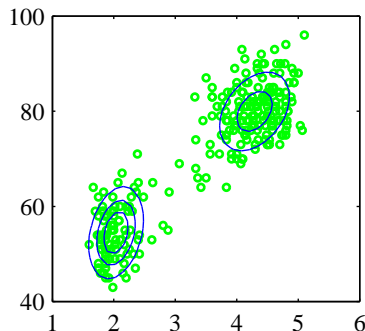
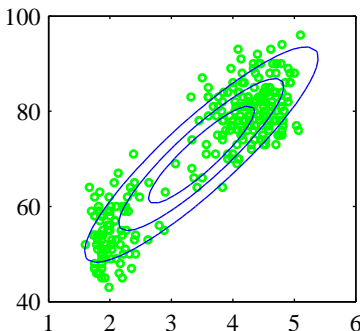
- 1 Unsupervised Learning
- 2 Clustering
- 3 k -means Clustering
- 4 Deriving k -means
- 5 k -means Clustering for Image Segmentation
- 6 Summary



Unsupervised learning

Unsupervised learning means

- learning (important aspects of) a data distribution p ,
- finding new *representations* of data that foster learning, generalisation, and communication.



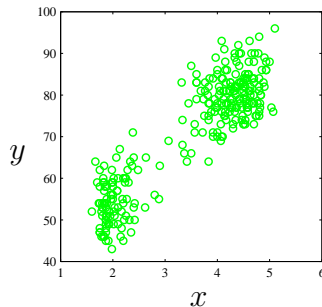
Unsupervised learning tasks

- Density estimation
 - Creating “closed-form” compact representation of data
 - Generative modeling
 - Classification/regression
 - Outlier detection
- Clustering
 - Unsupervised classification
 - Summarization by prototypes
- Feature extraction/visualization
 - Finding sub-space with highest variance and enabling best reconstruction
 - Finding regions with high density (k -means).



Example: Old Faithful

- Hydrothermal geyser in Yellowstone National Park, Wyoming, USA.



- x -axis duration of eruption in minutes
- y -axis time to next eruption in minutes



Outline

- ① Unsupervised Learning
- ② Clustering
- ③ k -means Clustering
- ④ Deriving k -means
- ⑤ k -means Clustering for Image Segmentation
- ⑥ Summary



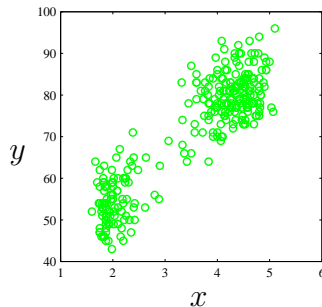
Clustering

- Clustering/segmentation assigns data records to clusters/groups
- Similar points should be in same cluster, dissimilar points in different clusters
- Hard clustering: every data point belongs to a single group; soft clustering: a data point can belong to more than one cluster



Example: Old Faithful

- Hydrothermal geyser in Yellowstone National Park, Wyoming, USA.



- x -axis duration of eruption in minutes
- y -axis time to next eruption in minutes



Outline

- 1 Unsupervised Learning
- 2 Clustering
- 3 k -means Clustering
- 4 Deriving k -means
- 5 k -means Clustering for Image Segmentation
- 6 Summary



k -means clustering

- Data set $S = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $1 \leq i \leq N$
- A priori chosen number k of groups
- Each group i is identified by a prototype/mean vector/cluster centroid $\boldsymbol{\mu}_i \in \mathbb{R}^n$
- All records assigned to group i are collected in S_i
- Similarity is measured by the Euclidean distance
- Objective function (distortion measure) to be minimized by finding optimal partitions S_i and cluster centroids $\boldsymbol{\mu}_i$ ($i = 1, \dots, k$):

$$J = \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$



k-means outline

Goal:

$$\min_{\substack{\mu_1, \dots, \mu_k \\ S_1, \dots, S_k : S = \\ S_1 \cup \dots \cup S_k}} \sum_{i=1}^k \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \mu_i\|^2$$

Iterate:

Data assignment: Assign each data point to cluster represented by the most similar prototype. This leads to a new partitioning of the data.

Centroid relocation: Recompute cluster centroids as mean of data points assigned to respective cluster.



k -means clustering algorithm

Algorithm 1: k -means clustering

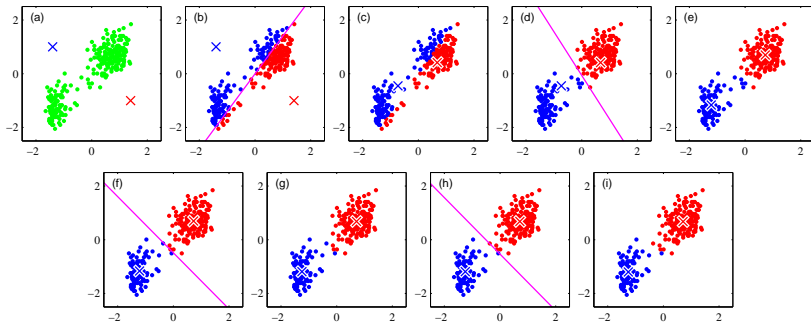
Input: $S = \{x_1, \dots, x_N\}$, number of clusters k

Output: cluster centers μ_1, \dots, μ_k , partitioning of the data
 S_1, \dots, S_k

```
1 initialize class centroids  $\mu_1, \dots, \mu_k$ 
2 repeat
3    $\forall i = 1, \dots, k : S'_i \leftarrow S_i$ 
   /* data assignment; ties are broken at random
   or by deterministic rule */
4    $\forall i = 1, \dots, k : S_i \leftarrow \{x \mid x \in S \wedge i = \operatorname{argmin}_j \|\mu_j - x\|\}$ 
   /* centroid relocation */
5    $\forall i = 1, \dots, k : \mu_i \leftarrow \frac{1}{|S_i|} \sum_{x \in S_i} x$ 
6 until  $\forall i = 1, \dots, k : S'_i = S_i$ 
Result:  $\mu_1, \dots, \mu_k; S_1, \dots, S_k$ 
```



k -means for Old Faithful



What are good initializations?

Noticed anything remarkable with the axes?



Outline

- 1 Unsupervised Learning
- 2 Clustering
- 3 k -means Clustering
- 4 Deriving k -means
- 5 k -means Clustering for Image Segmentation
- 6 Summary



Minimization of distortion measure

Data assignment: For fixed cluster centroids, \mathbf{x} should be assigned to nearest cluster i , because $\|\boldsymbol{\mu}_j - \mathbf{x}\| \geq \|\boldsymbol{\mu}_i - \mathbf{x}\|$ and thus assigning to j could only increase J .

Centroid relocation: Let μ_{ij} and x_j be the j th component of $\boldsymbol{\mu}_i$ and \mathbf{x} , respectively. Setting

$$\frac{\partial J}{\partial \mu_{ij}} = -2 \sum_{\mathbf{x} \in S_i} (x_j - \mu_{ij})$$

to zero gives

$$\boldsymbol{\mu}_i = \frac{1}{|S_i|} \sum_{\mathbf{x} \in S_i} \mathbf{x} .$$

Thus, cluster means minimize J with fixed partitioning.



Stopping criterion

For simplicity, let us assume deterministic breaking of ties.

Termination: Each partitioning uniquely defines cluster means. Each set of cluster means implies a particular partitioning. Thus, once the partitioning does not change after a relocation step the algorithm has converged.



Outline

- 1 Unsupervised Learning
- 2 Clustering
- 3 k -means Clustering
- 4 Deriving k -means
- 5 k -means Clustering for Image Segmentation
- 6 Summary



k -means for image segmentation

- Images are quite redundant.
- Many small patches are very similar.
- In the example we treat each RGB pixel as a 3D vector.
- Compression strategy:
Cluster with k -means and transmit cluster centers (code vectors) and assignments.

Original image



Image segmentation results

Original image



$K = 2$



$K = 3$



$K = 10$



Compression

- Compression for 8 bit accuracy and N pixel image
- Original image: $3 \cdot 8 \cdot N$ bits
- Cluster means (code vectors): $3 \cdot 8 \cdot k$ bits
- Assignments: $N \cdot \log_2 k$ bits
- Ratio, $k = 2, 3, 10$: 4.2%, 8.3%, 16.3%



Outline

- ① Unsupervised Learning
- ② Clustering
- ③ k -means Clustering
- ④ Deriving k -means
- ⑤ k -means Clustering for Image Segmentation
- ⑥ Summary



Summary and references

Clustering/segmentation:

- Clustering automatically groups data according to task-specific similarity measure
- There is neither a single “best” cluster algorithm nor a single “best” segmentation

k -means:

- ⊕ Simple, still gives good results
- ⊕ Just a single hyperparameter
- ⊖ k has to be chosen beforehand
- ⊖ Result heavily depends on initialization
 - Random data points are usually chosen as initial cluster means
 - Algorithm is usually run several times in practice

Pictures from C. M. Bishop. *Pattern Recognition and Machine Learning*, Springer, 2006, sections 9.1 & 9.3.2; slides inspired by Ole Winther.

