Clustering by K-Means

Context

- Recap
  - Main components of unsupervised machine learning
  - K-Means algorithm
- Exercise: https://drive.google.com/file/d/1MBDns_s4lp-Up2M7Dy52B3F4DST_gZca/view?usp=drive_link

Components of unsupervised machine learning

- Unlabeled data

  The model must discover structure directly from the input

- Feature

  Characteristics or attributes of the data used by the algorithm to learn and make predictions.

- Learning algorithm
  - **Clustering: K-means** / Hierarchical clustering /DBSCAN
- Model training (fit)

  Learning patterns or structure in the data without labels.

  (Examples: Minimizing within-cluster variance (K-means).)

- Model evaluation (validation)

  Assessing the quality of discovered structure without ground truth
  - Internal metrics: Quantitative measures that evaluate the quality of the learned structure using only the data itself, e.g. cohesion, separation
  - Indirect: Visualization

K-means algorithm

1) Randomly select K cluster centers $\mu_1, \ldots, \mu_K \in \mathbf{R}^p$
2) Assign each sample to the nearest cluster center
3) Update cluster centers to mean of samples in cluster
4) Repeat steps 2-3 until convergence

K-means convergence

