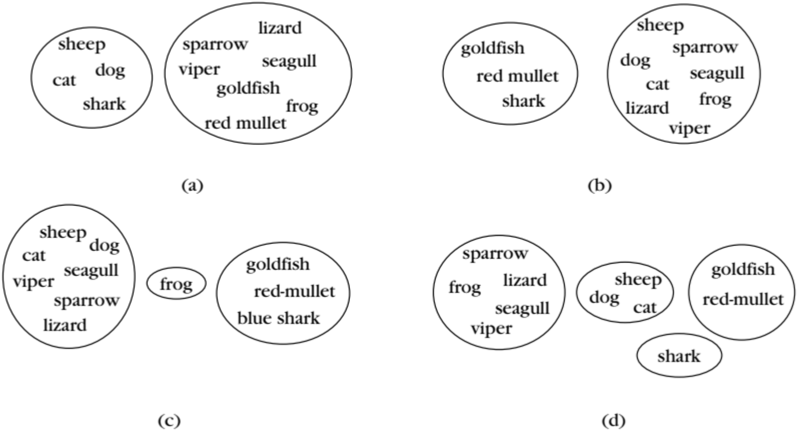
Clusterization and quantization

Marcin Kuta

Clustering


sheep
dog
cat
shark
lizard
sparrow
seagull
viper
goldfish
frog
red mullet
(a)
goldfish
red mullet
shark
sheep
sparrow
dog
seagull
cat
frog
lizard
viper
(b)
sheep
dog
cat
seagull
viper
sparrow
lizard
frog
goldfish
red-mullet
blue shark
(c)
sparrow
frog
lizard
seagull
viper
sheep
dog
cat
goldfish
red-mullet
shark
(d)

k-means

- $X = \{x_1, \ldots, x_n\}$ – set of points, $x_i \in \mathbb{R}^d$
- $C = \{c_1, \ldots, c_k\}$ – set of centroids, $c_i \in \mathbb{R}^d$
- $S = \{S_1, \ldots, S_k\}$ – set of clusters, $|S_i| = n_i$

k-means

1. Initialization
2. Assignment

$$S_i^{(t)} = \{x \in X \mid ||x - c_i|| < ||x - c_j|| \text{ for all } j \neq i, 1 \leq j \leq k\}$$

3. Update

$$c_i^{(t)} = \frac{1}{|S_i^{(t)}|} \sum_{x \in S_i^{(t)}} x$$

4. Repeat steps (2) and (3) until convergence

Cost function

Inertia = within-cluster sum squared error (SSE)

Minimization of inertia

$$C_{\text{best}} = \arg\min_{C} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - c_i||^2 \tag{1}$$

Cluster centers

- Mean
- Median
- Medoid

Initialization

- random
- k-means++
- harmonic k-means

Number of cluster

Criteria for optimal number of cluster [4]:

- Elbow method
- Silouhette score
- Akaike Information Criterion

DBSCAN

- Core points
- Border points
- Noise points

References

[1] https://github.com/pietroventurini/machine-learning-notes/blob/main/9%20-%20Cluster%20Analysis.ipynb

[2] https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.11-K-Means.ipynb

[3] https://github.com/rasbt/machine-learning-book/blob/main/ch10/ch10.ipynb

[4] https://antoinebrl.github.io/blog/kmeans