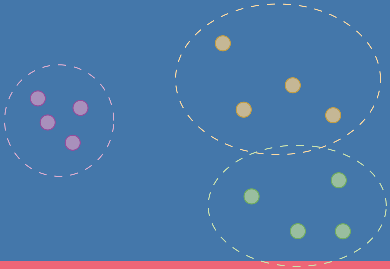
Lecture: Clustering

Martin Trapp



07.06.2023

Outline

- What is clustering?
- Example clustering algorithm
- Challenges and possible remedies
- Summary

Interactive Notebook:

Learning Goals:

- Aware of the challenges associated to clustering
- Know the k-Means algorithm and some of its pitfalls
- Know the basics of the Chinese Restaurant Process

Outline

- What is clustering?
- Example clustering algorithm
- Challenges and possible remedies
- Summary

Interactive Notebook:

Learning Goals:

- Aware of the challenges associated to clustering
- Know the k-Means algorithm and some of its pitfalls
- Know the basics of the Chinese Restaurant Process

What is clustering?

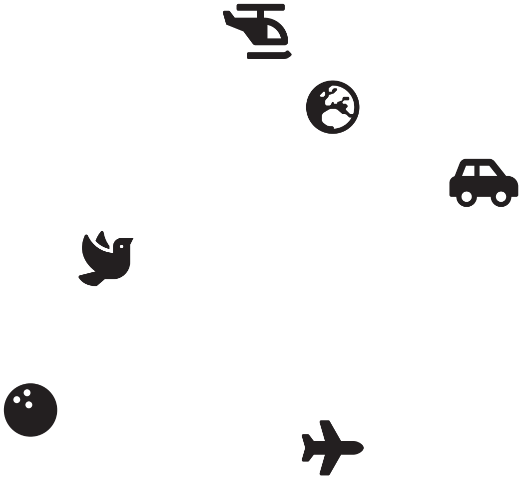
Definition (Cluster Analysis[1])

A set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual.

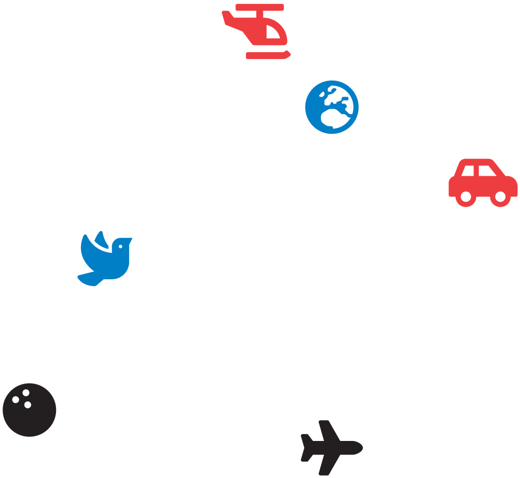
Essentially all such methods try to imitate what the eye-brain system does so well in two dimensions.

[1] B. S. Everitt and A. Skrondal (2010). *The Cambridge Dictionary of Statistics* (4th ed.) Cambridge University Press.

Clustering Example



Clustering Example



Clustering is NOT easy! 😭

- We can have many "sensible" groupings for the same data
- Clustering depends on the characteristics/features we select
- We can have a hierarchy of groupings
- It can be hard to measure the quality of the clustering

Is it hopeless?

No! ☺

Clustering is NOT easy! 😭

- We can have many “sensible” groupings for the same data
- Clustering depends on the characteristics/features we select
- We can have a hierarchy of groupings
- It can be hard to measure the quality of the clustering

Is it hopeless?

No! ☺

Clustering is NOT easy! 😭

- We can have many “sensible” groupings for the same data
- Clustering depends on the characteristics/features we select
- We can have a hierarchy of groupings
- It can be hard to measure the quality of the clustering

Is it hopeless?

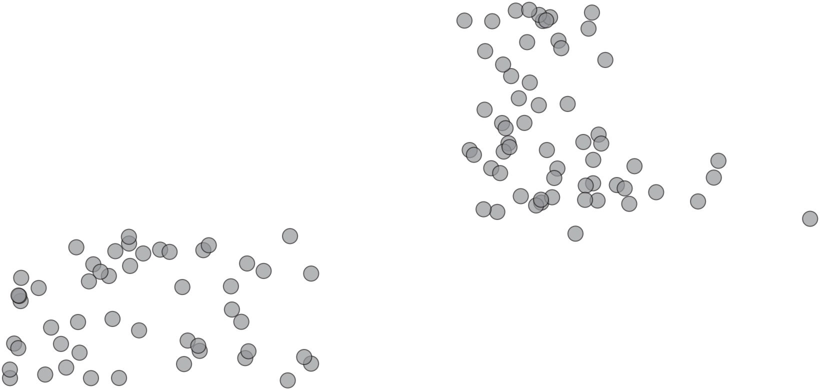
No! ☺

Clustering Methods

Selection of common clustering methods:

- Hierarchical Clustering
  - Agglomerative methods
  - Divisive methods
- Partitioning Clustering
  - K-means
  - K-medoids
- Density-based Clustering
  - DBSCAN
  - OPTICS

Let's Cluster...



K-Means

- K-Means is an iterative algorithm with the following steps:
  1. Compute distance $(x - \mu_k)^2$ for each datum x to each center μ_k
  2. For each x, find k with closest center c_k and add it to the set $\mathcal{A}_k$
  3. Recompute centers, by computing:

$$\mu_k = \frac{1}{|\mathcal{A}_k|} \sum_{i \in \mathcal{A}_k} x_i \tag{1}$$

- K-Means partitions the data set into groups/clusters by minimizing the (RSS):

$$\mathrm{RSS}_k = \sum_{i \in \mathcal{A}_k} (x_i - \mu_k)^2 \tag{2}$$

K-Means

- K-Means is an iterative algorithm with the following steps:
  1. Compute distance $(x - \mu_k)^2$ for each datum x to each center μ_k
  2. For each x, find k with closest center c_k and add it to the set $\mathcal{A}_k$
  3. Recompute centers, by computing:

$$\mu_k = \frac{1}{|\mathcal{A}_k|} \sum_{i \in \mathcal{A}_k} x_i \tag{1}$$

- K-Means partitions the data set into groups/clusters by minimizing the residual sum-of-squares (RSS):

$$\mathrm{RSS}_k = \sum_{i \in \mathcal{A}_k} (x_i - \mu_k)^2 \tag{2}$$

Problems with K-Means and Ways to Fix it

- How do we select the initial centers?
  - Try different random initialization
  - Select them in a 'clever' way, e.g., using K-means++
- How many clusters do we have?
  - Run k-means for $k = 1, 2, 3, 4, \ldots,$, e.g., and 'pick' the best one
  - Use a non-parametric (no K parameter) approach
- How can we be less sensitive to outliers?
  - Use a more robust objective
  - 'Trim' the data set by removing potential outliers

Problems with K-Means and Ways to Fix it

- How do we select the initial centers?
  - Try different random initialization
  - Select them in a 'clever' way, e.g., using K-means++
- How many clusters do we have?
  - Run k-means for $k = 1, 2, 3, 4, \ldots,$, e.g., and 'pick' the best one
  - Use a non-parametric (no K parameter) approach
- How can we be less sensitive to outliers?
  - Use a more robust objective
  - 'Trim' the data set by removing potential outliers

Problems with K-Means and Ways to Fix it

- How do we select the initial centers?
  - Try different random initialization
  - Select them in a 'clever' way, e.g., using K-means++
- How many clusters do we have?
  - Run k-means for $k = 1, 2, 3, 4, \ldots,$, e.g., and 'pick' the best one
  - Use a non-parametric (no K parameter) approach
- How can we be less sensitive to outliers?
  - Use a more robust objective
  - 'Trim' the data set by removing potential outliers

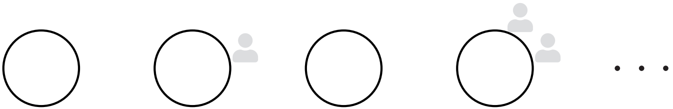
Problems with K-Means and Ways to Fix it

- How do we select the initial centers?
  - Try different random initialization
  - Select them in a ‘clever’ way, e.g., using K-means++
- How many clusters do we have?
  - Run k-means for $k = 1, 2, 3, 4, \ldots,$, e.g., and ‘pick’ the best one
  - Use a non-parametric (no K parameter) approach
- How can we be less sensitive to outliers?
  - Use a more robust objective
  - ‘Trim’ the data set by removing potential outliers

Problems with K-Means and Ways to Fix it

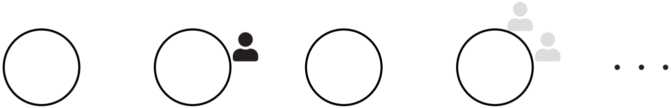
- How do we select the initial centers?
  - Try different random initialization
  - Select them in a ‘clever’ way, e.g., using K-means++
- How many clusters do we have?
  - Run k-means for $k = 1, 2, 3, 4, \ldots,$, e.g., and ‘pick’ the best one
  - Use a non-parametric (no K parameter) approach
- How can we be less sensitive to outliers?
  - Use a more robust objective
  - ‘Trim’ the data set by removing potential outliers

Chinese Restaurant Process



$$p(c=k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

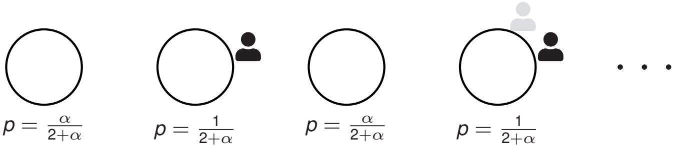
Chinese Restaurant Process



$p = \frac{\alpha}{1+\alpha}$ $\quad p = \frac{1}{1+\alpha}$ $\quad p = \frac{\alpha}{1+\alpha}$ $\quad p = \frac{\alpha}{1+\alpha}$

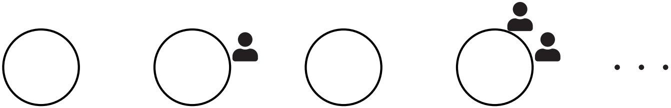
$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

Chinese Restaurant Process



$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

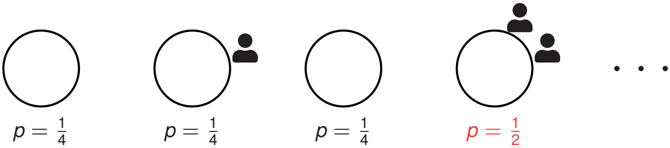
Chinese Restaurant Process



$p = \frac{\alpha}{3+\alpha}$ $\quad$ $p = \frac{1}{3+\alpha}$ $\quad$ $p = \frac{\alpha}{3+\alpha}$ $\quad$ $p = \frac{2}{3+\alpha}$

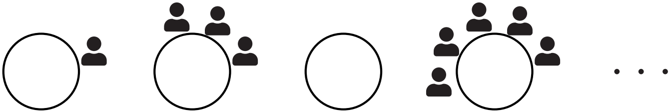
$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

Chinese Restaurant Process



$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

Chinese Restaurant Process



$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases}$$

Chinese Restaurant Process

Probability of selecting a table:

$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases} \tag{3}$$

$$\propto \begin{cases} m_k & \text{(number of data points in cluster k)} \\ \alpha \end{cases} \tag{4}$$

with $\alpha > 0$ begin the 'concentration' parameter.

Chinese Restaurant Process

Probability of selecting a table:

$$p(c = k) \propto \begin{cases} \frac{m_k}{n+\alpha} \\ \frac{\alpha}{n+\alpha} \end{cases} \tag{3}$$

$$\propto \begin{cases} m_k & \text{(number of data points in cluster k)} \\ \alpha \end{cases} \tag{4}$$

with $\alpha > 0$ begin the ‘concentration’ parameter.

Summary

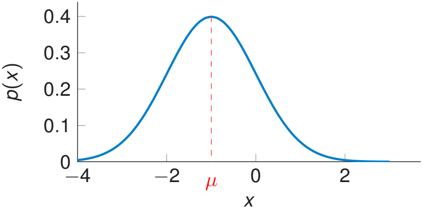
- Clustering is a challenging task
- K-Means, a simple algorithm to find a partitioning of the data
- Challenges associated to K-Means
- Non-parametric clustering with the Chinese Restaurant Process

The end of today's lecture

Thanks for listening, any questions?

Slides: `https://github.com/trappmartin/tue`

Gaussian/Normal Distributions

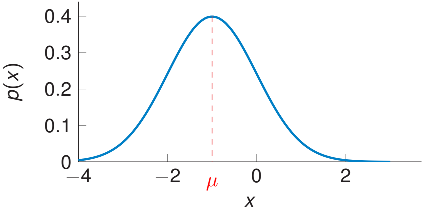


$$p(x \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{5}$$

$$\propto \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{6}$$

$$\log p(x \mid \mu, \sigma = 1) \propto -\frac{(x-\mu)^2}{2} \tag{7}$$

Gaussian/Normal Distributions

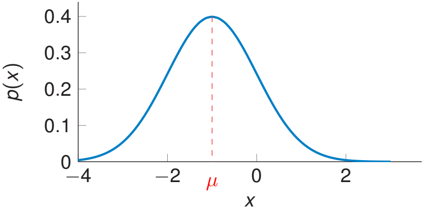


$$p(x \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{5}$$

$$\propto \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{6}$$

$$\log p(x \mid \mu, \sigma = 1) \propto -\frac{(x-\mu)^2}{2} \tag{7}$$

Gaussian/Normal Distributions

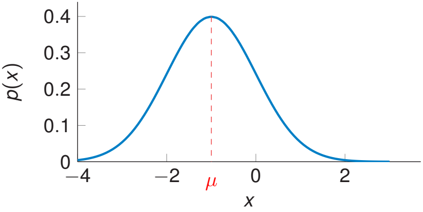


$$p(x \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{5}$$

$$\propto \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{6}$$

$$\log p(x \mid \mu, \sigma = 1) \propto -\frac{(x-\mu)^2}{2} \tag{7}$$

Gaussian/Normal Distributions

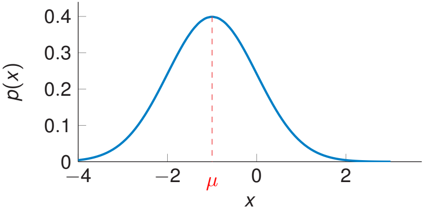


$$p(x \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{5}$$

$$\propto \exp\left(-\frac{(x-\mu)^2}{2}\right) \tag{6}$$

$$\log p(x \mid \mu, \sigma = 1) \propto -\frac{(x-\mu)^2}{2} \tag{7}$$

Gaussian/Normal Distributions



$$p(x \mid \mu, \sigma = 1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right) \quad (5)$$

$$\propto \exp\left(-\frac{(x-\mu)^2}{2}\right) \quad (6)$$

$$\log p(x \mid \mu, \sigma = 1) \propto -\frac{(x-\mu)^2}{2} = -\frac{\text{squared distance}}{2} \quad (7)$$

Chinese Restaurant Process

Log-probability of selecting a table (conditional on x):

$$\log p(c = k \mid x) \propto \log(m_k) - \frac{(x - \mu_k)^2}{2} \tag{8}$$

$$\log p(c = \text{new} \mid x) \propto \log(\alpha) - h_0 \tag{9}$$

- Clustering results is now stochastic
- Number of clusters dynamically changes

Chinese Restaurant Process

Log-probability of selecting a table (conditional on x):

$$\log p(c = k \mid x) \propto \log(m_k) - \frac{(x - \mu_k)^2}{2} \quad (8)$$

$$\log p(c = \text{new} \mid x) \propto \log(\alpha) - h_0 \quad (9)$$

- Clustering results is now stochastic
- Number of clusters dynamically changes