MATHEMATICS
FOR MACHINE LEARNING

Nagoya University, Fall 2023

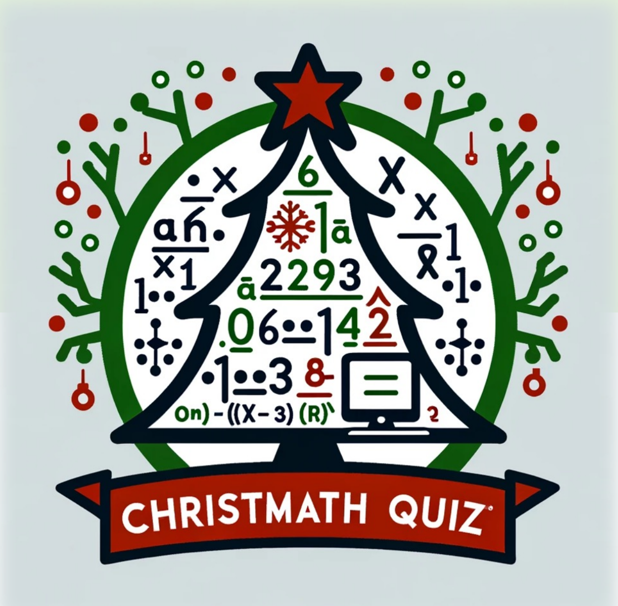
Lecture 11

Christmath Quiz,

Unsupervised learning: k-means algorithm

Join here:

www.menti.com 5454 6757

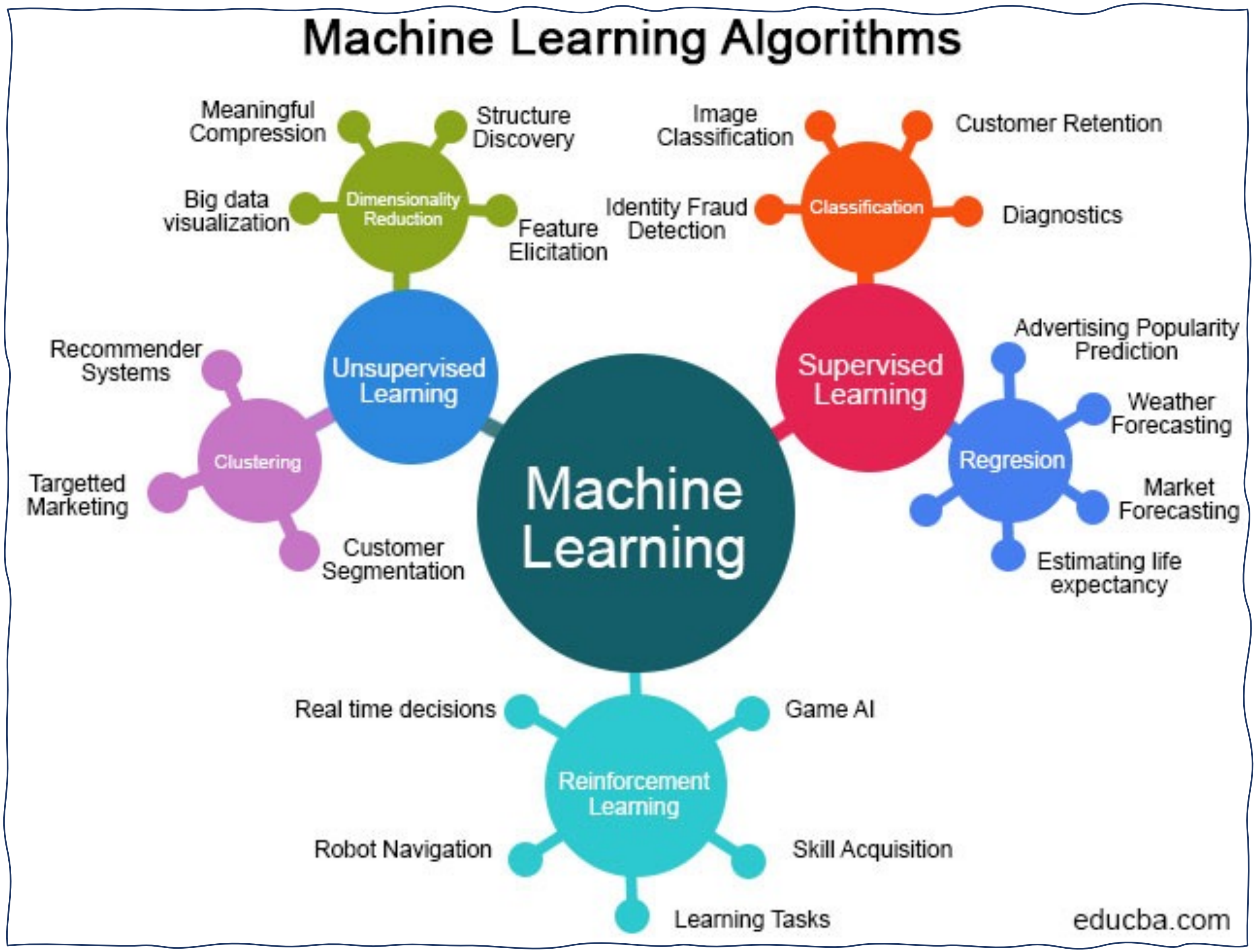


This week there is no Tutorial

The next lecture will be January 15th

https://www.henrikbachmann.com/mml2023.html

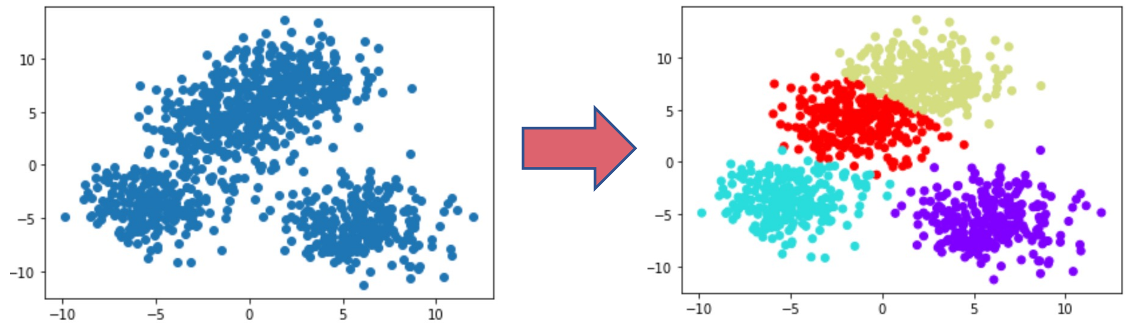
Overview

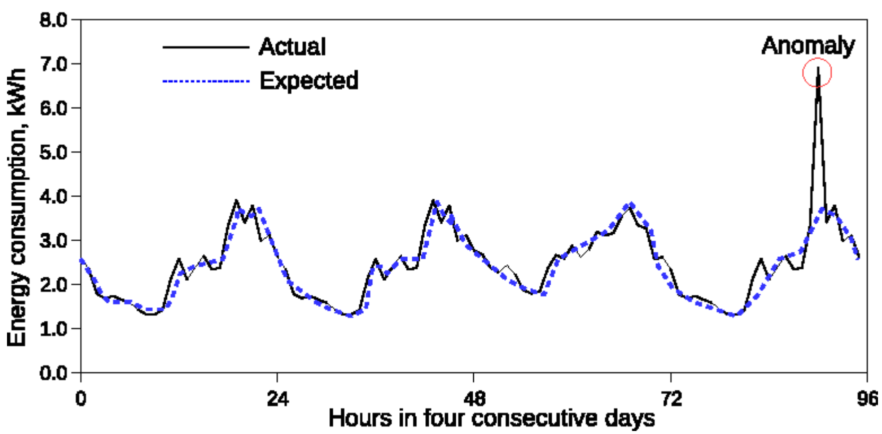


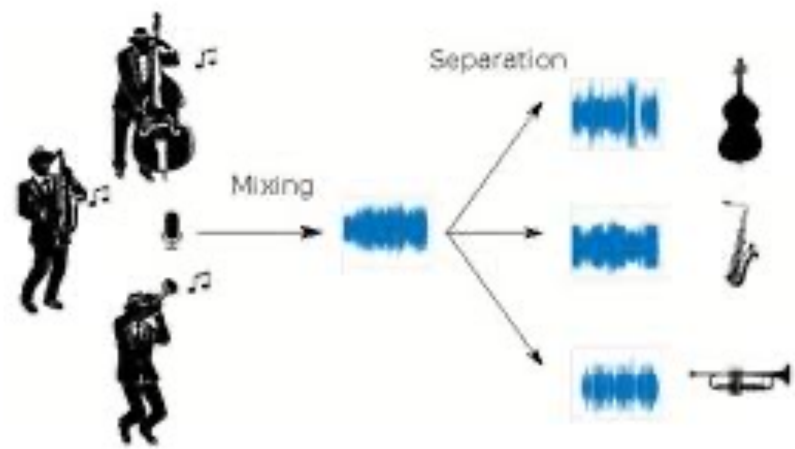
Machine Learning Algorithms
Meaningful Compression
Structure Discovery
Big data visualization
Dimensionality Reduction
Feature Elicitation
Image Classification
Customer Retention
Identity Fraud Detection
Classification
Diagnostics
Recommender Systems
Unsupervised Learning
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Clustering
Targetted Marketing
Regresion
Market Forecasting
Machine Learning
Customer Segmentation
Estimating life expectancy
Real time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks
educba.com

Unsupervised learning

- **Clustering**
- Anomaly detection
- Signal separation
- …



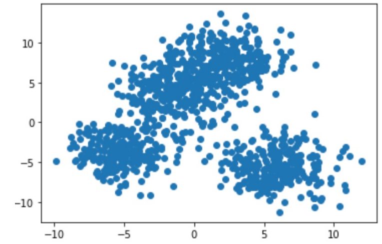




https://en.wikipedia.org/wiki/Signal_separation

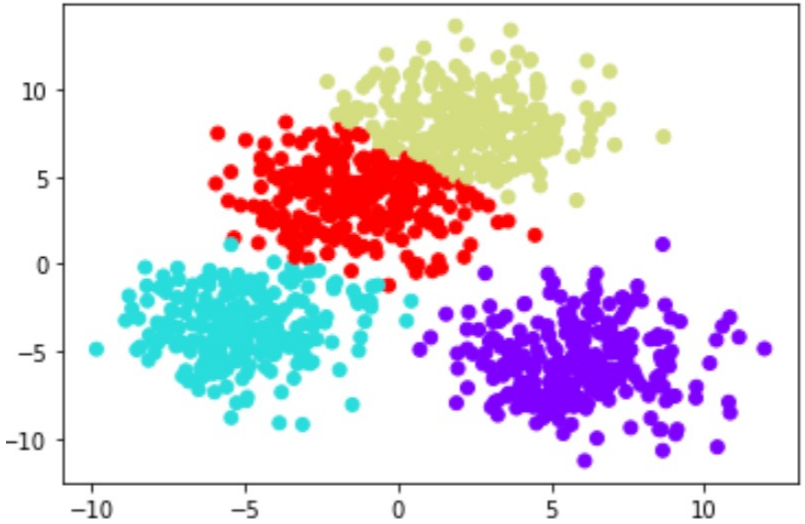
https://en.wikipedia.org/wiki/Cocktail_party_effect

Unsupervised learning: k-means clustering



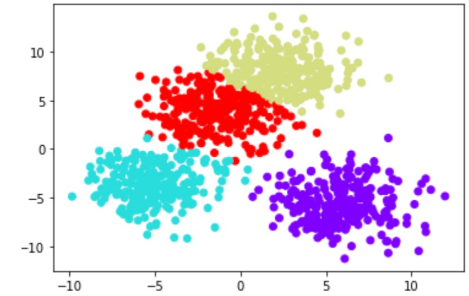
Let $P \subset \mathbb{R}^d$ be a set of points and let $k \geq 1$ be an integer.

Goal: Partition P into k clusters



Unsupervised learning: k-means clustering

Let $P \subset \mathbb{R}^d$ be a set of points and let $k \geq 1$ be an integer.



Goal: Partition P into k clusters

We want to find a map $c : P \to \{1, \ldots, k\}$ which for $i = 1, \ldots, k$ defines the **cluster**

$$C_i = \{p \in P \mid c(p) = i\} \ ,$$

such that the **within-cluster sum of squares (WCSS)**

$$S_c = \sum_{i=1}^{k} \sum_{p \in C_i} \|p - \mu_i\|^2$$

is minimal. Here $\mu_i \in \mathbb{R}^d$ denotes the **mean** of the cluster C_i given by

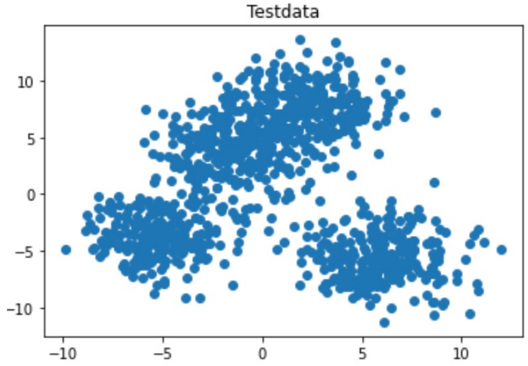
$$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p \, .$$

Unsupervised learning: k-means clustering

Let $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$ and fix a $k \geq 1$.

Goal: Construct the k-means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$. For given means one we define the cluster C_i by

$$C_i = \{p \in P \mid \|p - \mu_i\| \leq \|p - \mu_j\| \text{ for } j = 1, \ldots, k\}$$

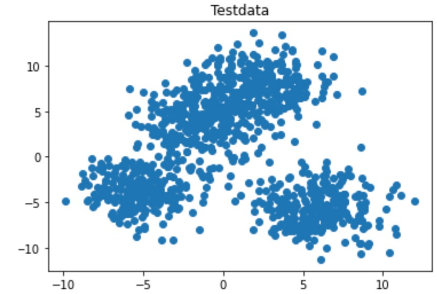


Unsupervised learning: k-means clustering

Let $P = \{p_1, \ldots, p_n\} \subset \mathbb{R}^d$ and fix a $k \geq 1$.

Goal: Construct the k-means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$. For given means one we define the cluster C_i by

$$C_i = \{p \in P \mid \|p - \mu_i\| \leq \|p - \mu_j\| \text{ for } j = 1, \ldots, k\}$$



k-means algorithm

1. Initialize the means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ with some starting value.
2. Define the clusters C_i for $i = 1, \ldots, k$ by

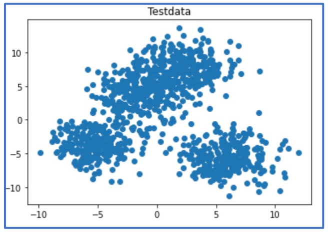
$$C_i = \{p \in P \mid \|p - \mu_i\| \leq \|p - \mu_j\| \text{ for } j = 1, \ldots, k\}$$

3. Recalculate the means μ_i for $i = 1, \ldots, k$ by

$$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p \,.$$

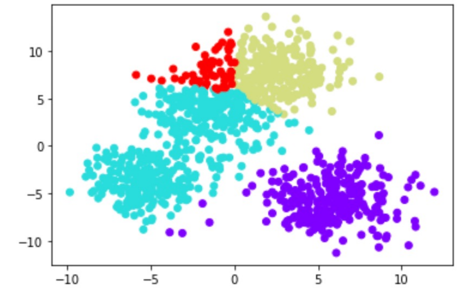
4. Repeat with step 2.

Unsupervised learning: k-means clustering

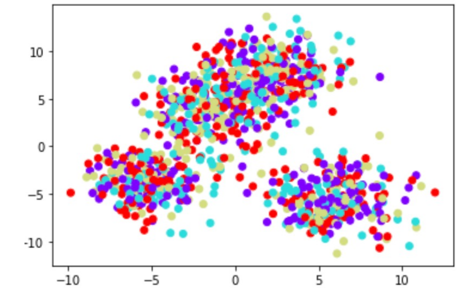


k-means algorithm

1. Initialize the means $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ with some starting value.
   (a) **Forgy method:** Choose randomly k different numbers $\{s_1, \ldots, s_k\} \subset \{1, \ldots, n\}$ and set
   $$\mu_i = p_{s_i}$$
   for $i = 1, \ldots, k$
   (b) **Random partition:** Choose $c : P \to \{1, \ldots, k\}$ randomly and set for $i = 1, \ldots, k$
   $$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p\,.$$
2. Define the clusters C_i for $i = 1, \ldots, k$ by
   $$C_i = \{p \in P \mid \|p - \mu_i\| \leq \|p - \mu_j\| \text{ for } j = 1, \ldots, k\}$$
3. Recalculate the means μ_i for $i = 1, \ldots, k$ by
   $$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p\,.$$
4. Repeat with step 2.

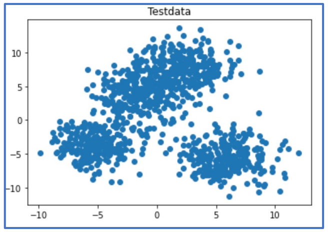


Init with the Forgy method with k = 4



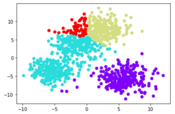
Init with Random partition with k = 4

Unsupervised learning: k-means clustering



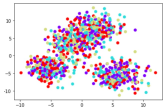
k**-means algorithm**

1. Initialize the means $\mu_1, \dots, \mu_k \in \mathbb{R}^d$ with some starting value.
   (a) **Forgy method:** Choose randomly k different numbers $\{s_1, \dots, s_k\} \subset \{1, \dots, n\}$ and set



$$\mu_i = p_{s_i}$$

   for $i = 1, \dots, k$



   (b) **Random partition:** Choose $c : P \to \{1, \dots, k\}$ randomly and set for $i = 1, \dots, k$

$$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p \,.$$

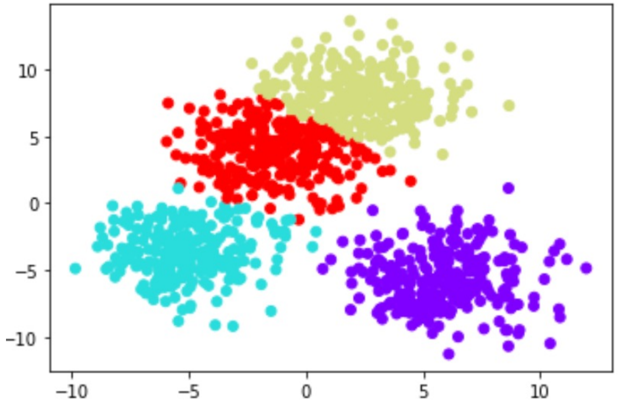
2. Define the clusters C_i for $i = 1, \dots, k$ by

$$C_i = \big\{p \in P \mid \|p - \mu_i\| \le \|p - \mu_j\| \text{ for } j = 1, \dots, k\big\}$$

3. Recalculate the means μ_i for $i = 1, \dots, k$ by

$$\mu_i = \frac{1}{|C_i|} \sum_{p \in C_i} p \,.$$

4. Repeat with step 2.



Result after 20 steps