Aprententatge Automàtic per a Xarxes (ML4Net)

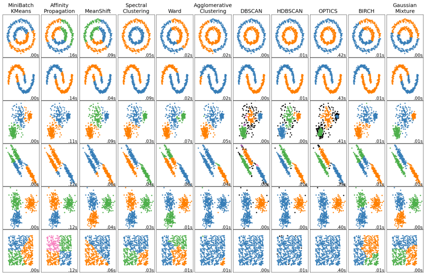
Seminar 3 - K-means

Francesc Wilhelmi & Boris Bellalta
School of Engineering
Universitat Pompeu Fabra

Clustering via unsupervised learning

Key concepts:

- Features (x): Available data points to be clustered, $X = \{x_1, x_2, ..., x_n\}$, where $x_i \in \mathbb{R}^d$.
- Labels (y): The ground truth (if any!) associated with the data points.
- Model (h): Function $f : X \to \{1, 2, ..., k\}$ that assigns each data point $x \in X$ to a cluster $c \in \{1, 2, ..., k\}$.
- Goal: Group similar data into clusters (other unsupervised learning goals include anomaly detection, dimensionality reduction, or density estimation).

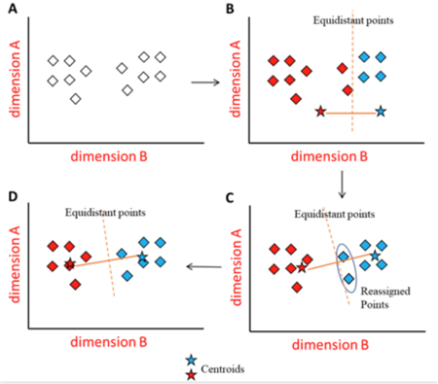


[Source: `https://scikit-learn.org/stable/modules/clustering.html`]

K-means

- We want to divide the data points $x \in X$ into C disjoint clusters.
- Each cluster c is described by a centroid, which is computed as the mean of the samples in the cluster μ_c.
- How do we find the best centroids?

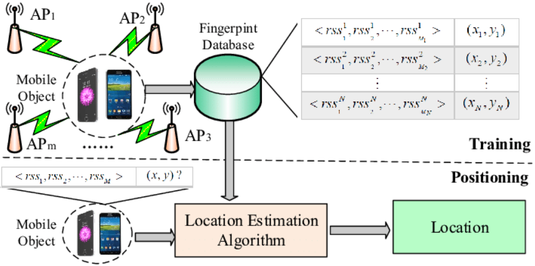
$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$



[Source: https://www.blopig.com/blog/2020/07/k-means-clustering-made-simple/]

Dataset (I)

- Wi-Fi fingerprinting is a technique used for indoor positioning
- It is based on the RSSI measured at different APs, which are divided into reference points (RPs).
- Two phases:
  - Offline phase: Acquire measurements (RSSI) from different RPs at different APs and derive a model (e.g., K-means, KNNs, NNs).
  - Online phase: New RSSI values are processed and passed to the trained model.

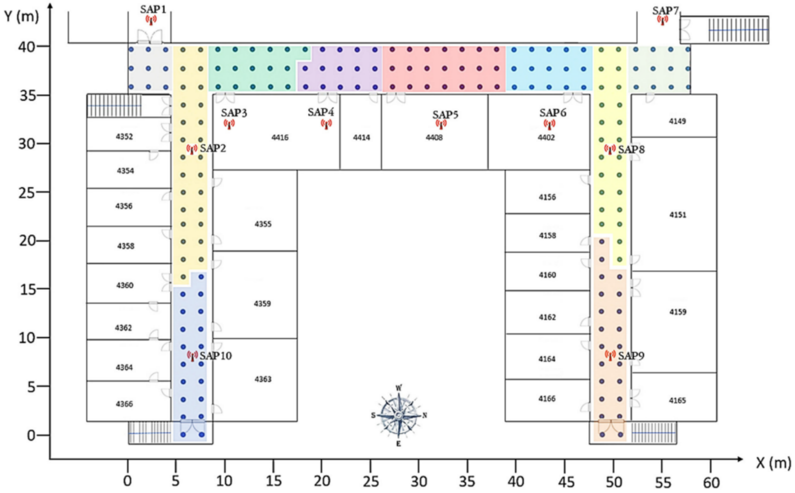


$< rss_1^1, rss_2^1, \cdots, rss_{M_1}^1 >$	(x_1, y_1)
$< rss_1^2, rss_2^2, \cdots, rss_{M_2}^2 >$	(x_2, y_2)
$\vdots$	$\vdots$
$< rss_1^N, rss_2^N, \cdots, rss_{M_N}^N >$	(x_N, y_N)

$< rss_1, rss_2, \cdots, rss_M >$	(x, y) ?

Gu, F., Hu, X., Ramezani, M., Acharya, D., Khoshelham, K., Valaee, S., & Shang, J. (2019). Indoor localization improved by spatial context—A survey. ACM Computing Surveys (CSUR), 52(3), 1-35.

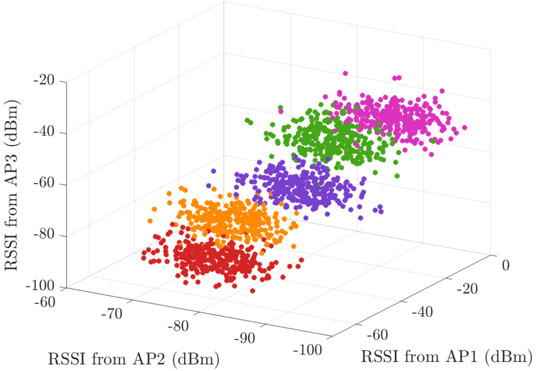
Dataset (II)



Ezhumalai, B., Song, M., & Park, K. (2021). An efficient indoor positioning method based on Wi-Fi RSS fingerprint and classification algorithm. Sensors, 21(10), 3418.

Dataset (III)

- `rssi_data.csv`: Wi-Fi RSSI measurements taken at $K = 3$ different APs (AP1, AP2, and AP3), for $P = 5$ positions in which the STA was placed.
  - In each row, there is the RSSI perceived by each of the APs for a single measurement.
  - The dataset includes $M = 300$ measurements per location $p \in P$
  - It contains $(P \times M) \times K$ values.
- `labels_data.csv`: The ground truth (i.e., the real position of the STA, p) for each measurement.
  - Contains $P \times M$ values.


RSSI from AP3 (dBm)
RSSI from AP2 (dBm)
RSSI from AP1 (dBm)