Abteilung für
Angewandte Mathematik

Prof. Dr. Sören Bartels, Tatjana Schreiber May 14th, 2025

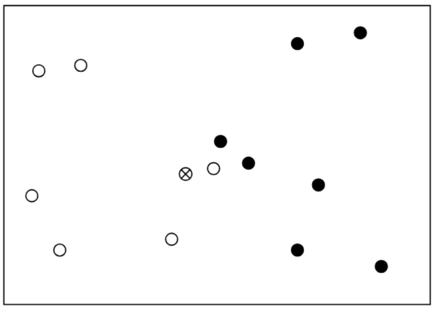
Algorithmic Aspects of Data Analytics and Machine Learning

SS 2025 — Sheet 4

`https://aam.uni-freiburg.de/agba/lehre/ss25/algml/index.html`

Due: May 23, 2025, 2 p.m.

Task 1 (4 points)



Consider the following dataset $D \subset \mathbb{R}^2 \times \{1, 2\}$, where white points have label 1 and black points have label 2. The label of the point marked with a cross is unknown.

Using visual inspection only, determine the label of the point marked with a cross according to the k-nearest neighbors (k-NN) rule with majority vote for $k = 1, \ldots, 7$.

Task 2 (4 points)

Given the following Netflix rating matrix

$$R = \left(\begin{array}{ccccc} & M_1 & M_2 & M_3 & M_4 \\ U_1 & 5 & 3 & - & 1 \\ U_2 & 4 & - & - & 1 \\ U_3 & 1 & 1 & - & 5 \\ U_4 & 1 & - & - & 4 \\ U_5 & - & 1 & 5 & 4 \end{array}\right).$$

With ratings given by users $U_1, \ldots, U_5$ for movies $M_1, \ldots, M_4$.
Complete the matrix R using the k-nearest neighbors k-NN algorithm with $k = 2$. Use cosine similarity, with respect to the users.

Task 3 (4 points)

Let $A \subset \mathbb{R}^d$ be a finite set, and let $d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denote the Euclidean metric.
Show that for every finite set $A \subset \mathbb{R}^d$, the following holds:

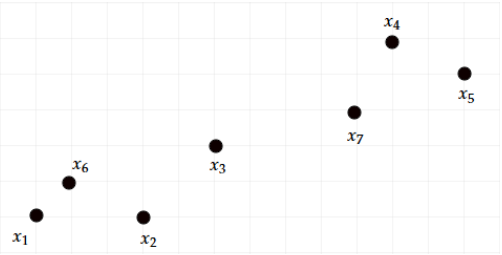
$$\frac{1}{|A|} \sum_{a \in A} a = \arg\min_{\mu \in \mathbb{R}^d} \sum_{a \in A} d(a, \mu)^2.$$

That is, the mean of the points in A uniquely minimizes the sum of squared Euclidean distances, and is thus consistent with the standard definition of the centroid in linear algebra.

Task 4 (4 points)

Apply the 2-means algorithm to the given dataset, using the following initialization:

$$\mu_1 = x_7, \quad \mu_2 = x_2$$



Assume the Euclidean metric on $\mathbb{R}^2$. For each iteration, indicate the resulting clusterings C_i and the updated centroids μ_i (a rough estimation by eye is sufficient).

Repeat the algorithm with the initialization:

$$\mu_1 = x_1, \quad \mu_2 = x_6.$$

Practical exercise

The following exercise is not mandatory; the points are bonus points that you can collect. Please submit your solutions as a MATLAB or Python file by May 23, 2 p.m., via email to tatjana.stiefken@mathematik.uni-freiburg.de. Please comment your code and your results.

Project (4* points)

Generate a 2D synthetic dataset with two classes:

- Class 0: points in $B_1(x) := \{x \mid |x| < 1\}$
- Class 1: points on $\partial B_1(x) = \{x \mid |x| = 1\}$.

Split the dataset into training (70%) and test (30%) sets. Train a k-NN classifier for various values of k (e.g., $k = 1, 3, 5, 10, 50$). For each k:

- Plot the decision boundary (use a mesh grid for visualization).
- Compute training and test classification errors.

Plot training and test error vs. k in a line chart. Comment your results in terms of underfitting and overfitting.
Implementation: Use either Python or MATLAB. In MATLAB, you can use `fitcknn` and visualize with `gscatter`. In Python, consider using `sklearn.neighbors.KNeighborsClassifier`.