Unsupervised Learning Methods
Problem Set I –
Optimization and Clustering

Due: 03.05.2021

Guidelines

- Answer all questions (PDF + Jupyter notebook).
- You must type your solution manual (handwriting is not allowed).
- Submission in pairs (use the forum if needed).
- You **may** submit the entire solution in a single ipynb file (or in PDF + ipynb files).
- You **may** (and should) use the forum if you have any questions.
- Good luck!

1 Optimization

1.1 Convexity

Convex set

Let:

$$\mathbb{R}^d_{\geq 0} = \left\{ \boldsymbol{x} \in \mathbb{R}^d \,\middle|\, \min_i x_i \geq 0 \right\}$$

where $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$.

1.1.1

Prove or disprove: $\mathbb{R}^d_{\geq 0}$ is convex.

Convex combination

Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a convex set and consider $\{\boldsymbol{x}_i \in \mathcal{C}\}_{i=1}^{N}$.

1.1.2

Prove that for any $N \in \mathbb{N}$:

$$\sum_{i=1}^{N} \alpha_i \boldsymbol{x}_i \in \mathcal{C}$$

where α_i are such that:

- $\alpha_i \geq 0$ for all i.
- $\sum_{i=1}^{N} \alpha_i = 1$.

Let $\mathcal{C} \subset \mathbb{R}^2$ and consider $\{\boldsymbol{x}_i \in \mathcal{C}\}_{i=1}^{10}$ such that $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ for all $i \neq j$.

1.1.3

Prove or disprove: Necessarily, any point $\boldsymbol{y} \in \mathcal{C}$ can be represented as a convex combination of $\{\boldsymbol{x}_i\}_{i=1}^{10}$.

Convex functions

Let $f : \mathbb{R} \to \mathbb{R}$ be such that for all $x, y \in \mathbb{R}$:

$$f(y) \geq f(x) + f'(x)(y - x)$$

1.1.4

Prove that f is a convex function.
Hint:

- Let: $z := \alpha x + (1 - \alpha) y$
- Note that $\begin{cases} f(y) \geq f(z) + f'(z)(y - z) \\ f(x) \geq f(z) + f'(z)(x - z) \end{cases}$

1.2 The Gradient

Directional derivative

Let $f : \mathbb{R}^d \to \mathbb{R}$ and let $\boldsymbol{x}_0 \in \mathbb{R}^d$.

1.2.1

Prove that:

$$\forall \boldsymbol{h} \in \mathbb{R}^d : \nabla f(\boldsymbol{x}_0)[\boldsymbol{h}] = \langle \boldsymbol{g}_0, \boldsymbol{h} \rangle \implies \boldsymbol{g}_0 = \nabla f(\boldsymbol{x}_0)$$

Definition
$f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is said to be linear if:

$$f(\alpha \boldsymbol{x} + \beta \boldsymbol{y}) = \alpha f(\boldsymbol{x}) + \beta f(\boldsymbol{y})$$

for all $\alpha, \beta \in \mathbb{R}$ and for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d_1}$.
Let $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ be a linear function.

1.2.2

Prove that:

$$\nabla f(\boldsymbol{x})[\boldsymbol{h}] = f(\boldsymbol{h})$$

for all $\boldsymbol{x}, \boldsymbol{h} \in \mathbb{R}^{d_1}$

1.2.3 Some useful exercises

Compute the directional derivative $\nabla f(\boldsymbol{x})[\boldsymbol{h}]$ and the gradient $\nabla f(\boldsymbol{x})$ for:

1. $$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$$

2. $$f(\boldsymbol{X}) = \mathrm{Tr}\left\{\boldsymbol{X}^T \boldsymbol{A} \boldsymbol{X}\right\}$$

   where $\boldsymbol{X} \in \mathbb{R}^{N \times d}$ and $\mathrm{Tr}\{\cdot\}$ is the trace operator.

3. $$f(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

4. $$f(\boldsymbol{X}) = \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2$$

   where:

   (a) $\boldsymbol{Y} \in \mathbb{R}^{D \times N}$, $\boldsymbol{A} \in \mathbb{R}^{D \times d}$ and $\boldsymbol{X} \in \mathbb{R}^{d \times N}$.

(b) $\|\cdot\|_F^2$ is the Frobenius norm, that is, $\|\boldsymbol{X}\|_F^2 = \langle \boldsymbol{X}, \boldsymbol{X} \rangle = \mathrm{Tr}\left\{\boldsymbol{X}^T \boldsymbol{X}\right\}$.

5.
$$f(\boldsymbol{X}) = \left\langle \boldsymbol{X}^T \boldsymbol{A}, \boldsymbol{Y}^T \right\rangle$$

where $\boldsymbol{Y} \in \mathbb{R}^{D\times N}$, $\boldsymbol{A} \in \mathbb{R}^{d\times D}$ and $\boldsymbol{X} \in \mathbb{R}^{d\times N}$.

6.
$$f(\boldsymbol{x}) = \boldsymbol{a}^T g(\boldsymbol{x})$$

where:

(a) $g : \mathbb{R} \to \mathbb{R}$ is scalar function (for example $g(x) = \sin(x)$) with a known derivative g'.

(b) $g(\boldsymbol{x}) := \begin{bmatrix} g(x_1) \\ \vdots \\ g(x_d) \end{bmatrix} \in \mathbb{R}^d$

7.
$$f(\boldsymbol{X}) = \langle \boldsymbol{A}, \log[\boldsymbol{X}] \rangle$$

where:

(a) $\boldsymbol{X} \in \mathbb{R}^{d\times d}$

(b) $\log[\boldsymbol{X}]$ is an element-wise log, that is:

$$\boldsymbol{M} = \log[\boldsymbol{X}] \implies \boldsymbol{M}[i,j] = \log(\boldsymbol{X}[i,j])$$

8.
$$f(\boldsymbol{X}) = \langle \boldsymbol{a}, \mathrm{diag}(\boldsymbol{X}) \rangle$$

where:

(a) $\boldsymbol{X} \in \mathbb{R}^{d\times d}$

(b) $\mathrm{diag} : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ returns the diagonal of a matrix, that is:

$$\boldsymbol{b} = \mathrm{diag}(\boldsymbol{X}) \implies \boldsymbol{b}[i] = \boldsymbol{X}[i,i]$$

1.3 Descent Methods (Gradient Descent and Momentum)

Solve this section in the attached notebook.

1.4 Constraint optimization

Minimax

Let $G(x, y) = \sin(x + y)$.

1.4.1

Show that:

1. $\underset{x}{\min}\underset{y}{\max} G(x, y) = 1$
2. $\underset{y}{\max}\underset{x}{\min} G(x, y) = -1$

Rayleigh quotient

- The Rayleigh quotient is defined by:

$$f(\boldsymbol{x}) = \frac{\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}}{\boldsymbol{x}^T \boldsymbol{x}}$$

  for some symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$.

1.4.2

1. Show that

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{cases} \min_{\boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} \\ \text{s.t.} \quad \|\boldsymbol{x}\|_2^2 = 1 \end{cases}$$

2. Write the Lagrangian of the constraint objective $\mathcal{L}(\boldsymbol{x}, \lambda)$.
3. Show that:

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \lambda) = 0 \iff \boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}$$

   in other words, the stationary points $(\boldsymbol{x}, \lambda)$ are the eigenpairs of $\boldsymbol{A}$ (eigenvectors and eigenvalues).

---

2 Clustering

2.1 K-Means

Objective The K-Means objective is given by:

$$\arg \min_{\{\mathcal{D}_k\},\{\boldsymbol{\mu}_k\}} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{D}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

2.1.1

Show that the following two objectives are equivalent to the K-Means:

1. As a sole function of the clusters:

$$\arg \min_{\{\mathcal{D}_k\}} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}_k} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$$

2. As a sole function of the centroids:

$$\arg \min_{\{\boldsymbol{\mu}_k\}} \sum_{i=1}^{N} \min_k \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

2.1.2

Prove or disprove:
The K-Means algorithm **always** converges to a global minimum.

2.1.3 K-Means +

2.1.4 Super-pixels

Solve this section in the attached notebook.

2.2 GMM

Gaussian random vector

- Let $\underline{X} \sim \mathcal{N}\left(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right)$ be a Gaussian random vector.
- Let $Y = \boldsymbol{a}^T \underline{X} + b$ be a random variable

2.2.1

Find $f_Y\left(y\right)$, the pdf of Y (as a function of $\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \boldsymbol{a}, b$).

Covariance

A matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is called Symmetric Positive Semi-Definite (SPSD) if $\boldsymbol{A}^T = \boldsymbol{A}$ and for any $\boldsymbol{v} \in \mathbb{R}^d$:

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0$$

In other words:

$$\boldsymbol{A} \succeq 0 \iff \begin{cases} \boldsymbol{A}^T = \boldsymbol{A} \\ \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0 \quad \forall \boldsymbol{v} \end{cases}$$

Let $\underline{X}$ be a random vector with covariance $\boldsymbol{\Sigma}_x$.

2.2.2

Prove that $\boldsymbol{\Sigma}_x$ is an SPSD matrix.

2.2.3 GMM +

2.2.4 Digits

Solve this section in the attached notebook.

2.3 Hierarchical Clustering

Complete-linkage

The complete-linkage distance between the two clusters $\mathcal{C}_1 = \{\boldsymbol{x}_i\}_{i=1}^{N_1}$ and $\mathcal{C}_2 = \{\boldsymbol{x}_j\}_{j=1}^{N_2}$:

$$d^2_{\text{complete-link}}(\mathcal{C}_1, \mathcal{C}_2) = \begin{cases} 0 & \mathcal{C}_1 = \mathcal{C}_2 \\ \max_{\boldsymbol{x}_i \in \mathcal{C}_1, \boldsymbol{x}_j \in \mathcal{C}_2} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 & \text{else} \end{cases}$$

2.3.1

Prove that the complete-linkage is indeed a metric.

Lance-Williams

The Lance-Williams update rule (see the full algorithm in the lecture notes):

$$D_{\widetilde{ij},k} \leftarrow \alpha_i D_{i,k} + \alpha_j D_{j,k} + \beta D_{i,j} + \gamma \left|D_{i,j} - D_{j,k}\right|$$

Consider the three clusters $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ with

$$D_{i,j} = d_{\text{single-link}}(\mathcal{C}_i, \mathcal{C}_j)$$

2.3.2

Prove that

$$D_{\widetilde{12},3} = d_{\text{single-link}}(\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{C}_3)$$

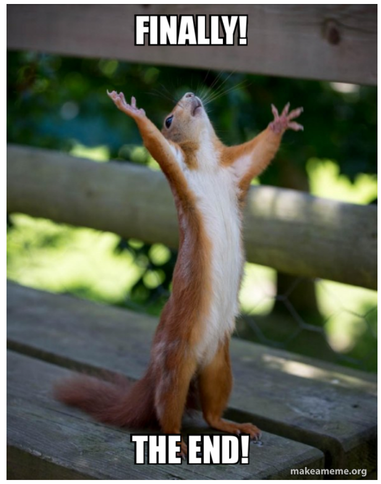
In words, show that the Lance-Williams algorithm is correct for the single-linkage dissimilarity.

2.4 DBSCAN

2.4.1 DBSCAN implementation

2.4.2 Clustering methods comparison

Solve this section in the attached notebook.


FINALLY!
THE END!
makeameme.org