Mathematics of Data Science 2017

Part 1 – Maximum Entropy

In the following, we denote $(a,b) \in \mathbb{R}_+^n \times \mathbb{R}_+^m$ with $\sum_{i=1}^n a_i = \sum_{j=1}^m b_j = 1$. We define the entropy of $P \in \mathbb{R}_+^{n\times m}$ as

$$H(P) = -\sum_{i,j} P_{i,j} \log(P_{i,j}), \tag{1}$$

with the convention $0\log(0) = 0$.

1) Show that $-H$ is a strictly convex function.

2) We denote $a \otimes b = (a_i b_j)_{i,j}$. Compute $H(a \otimes b)$.

3) Solve the following optimization problem

$$\min_{P\in\mathbb{R}_+^{n\times m}} \left\{ -H(P) \;;\; \forall i, (P\mathbb{1}_m)_i = \sum_j P_{i,j} = a_i, \forall j, (P^\top \mathbb{1}_n)_i = \sum_i P_{i,j} = b_j \right\}.$$

4) For $a \in \mathbb{R}_{+,*}^n$ (strictly positive) and $r \in \mathbb{R}_+^n$, we define the Kulback-Leibler divergence between the two vectors (the same expression holds also for matrices, where the sum is on (i,j) instead of just i) as

$$\mathrm{KL}(r|a) \stackrel{\text{def.}}{=} \sum_i \log\left(\frac{r_i}{a_i}\right) r_i - r_i + a_i. \tag{2}$$

Show that the function $\mathrm{KL}(\cdot|a)$ is convex and compute its minimizer. Deduce that KL is "distance-like", i.e. that $\mathrm{KL}(r|a) \geqslant 0$ and $\mathrm{KL}(r|a) = 0$ if and only if $r = a$. Show that, if P is such that

$$P\mathbb{1}_m = (\sum_j P_{i,j})_i = a \quad \text{and} \quad P^\top \mathbb{1}_n = (\sum_i P_{i,j})_j = b$$

then one has

$$\mathrm{KL}(P|a\otimes b) = \mathrm{KL}(P|a'\otimes b') + \mathrm{KL}(a'\otimes b'|a\otimes b).$$

Part 2 – Hadamard-Walsh Transform

We denote $G = (\mathbb{Z}/2\mathbb{Z})^p$, which is a group with $n \stackrel{\text{def.}}{=} 2^p$ elements. An element $x \in G$ is written as $x = (x_i)_{i=1}^p$ with $x_i \in \{0,1\}$ and is equivalently represented as $0 \leqslant x < n$ where the x_i is the binary writing of x in base 2. We denote $\mathbb{R}[G]$ the vector space of functions $f : G \to \mathbb{R}$, endowed with the canonical inner product $\langle f, g\rangle \stackrel{\text{def.}}{=} \sum_{x\in G} f(x)g(x)$.

1) For $\omega \in G$, we denote $\psi_\omega(x) \stackrel{\text{def.}}{=} (-1)^{\sum_{i=1}^p x_i\omega_i}$. Show that $(\psi_\omega)_{\omega\in G}$ is an orthogonal basis of $\mathbb{C}[G]$.

2) We denote $\hat{f}(\omega) \stackrel{\text{def.}}{=} \langle f, \psi_\omega\rangle$ the Hadamard-Walsh transform of f. We denote $f_0, f_1 : (\mathbb{Z}/2\mathbb{Z})^{p-1} \to \mathbb{R}$ defined as

$$f_0(x_1,\ldots,x_{p-1}) \stackrel{\text{def.}}{=} f(x_1,\ldots,x_{p-1},0) \quad \text{and} \quad f_1(x_1,\ldots,x_{p-1}) \stackrel{\text{def.}}{=} f(x_1,\ldots,x_{p-1},1).$$

Find a relation between $\hat{f}$ and $(\hat{f}_0, \hat{f}_1)$. Write in pseudo-code a fast recursive algorithm to compute $\hat{f}$ from f. What it is the number of operations of this algorithm ?

3) Detail a fast algorithm to compute the inverse Walsh transform, i.e. which computes f from $\hat{f}$. What it is the number of operations of this algorithm ?

4) Given $(f,g) \in \mathbb{R}[G]$, how would you define the convolution of f and g? Show that $\widehat{f \star g} = \hat{f} \odot \hat{g}$ where $\odot$ is the pointwise multiplication of vectors.

Part 3 – Haar Wavelets

The goal of this exercise is to detail the construction of an orthogonal basis of $\mathrm{L}^2([0,1])$, which was introduced by Alfred Haar in 1909, and is the first example of Wavelet basis.

For $j \in \mathbb{N}$, one defines the space $\mathcal{V}_j \subset \mathrm{L}^2([0,1])$ of fonctions which are constant on each interval $I_{j,k}$, defined for $k \in \{0,\dots,2^j-1\}$ as

$$\forall\, k \in \{0,\dots,2^j-1\}, \quad I_{j,k} \stackrel{\text{def.}}{=} \begin{cases} [\frac{k}{2^j}, \frac{k+1}{2^j}[& \text{si} \quad k < 2^j-1, \\ [1-\frac{1}{2^j}, 1] & \text{si} \quad k = 2^j-1. \end{cases}. \tag{3}$$

One also defines the functions

$$\forall\, x \in \mathbb{R}, \quad \varphi(x) \stackrel{\text{def.}}{=} \begin{cases} 1 & \text{si} \quad x \in [0,1[, \\ 0 & \text{sinon.} \end{cases} \quad \text{and} \quad \psi(x) \stackrel{\text{def.}}{=} \begin{cases} 1 & \text{si} \quad x \in [0,1/2[, \\ -1 & \text{si} \quad x \in [1/2,1[, \\ 0 & \text{sinon.} \end{cases}$$

Their versions dilated by $1/2^j$ and translated at position $k/2^j$ are defined, for $j \in \mathbb{N}$, as

$$\forall\, k \in \{0,\dots,2^j-1\}, \quad \psi_{j,k}(x) \stackrel{\text{def.}}{=} 2^{j/2}\psi\left(2^j x - k\right), \quad \text{and} \quad \varphi_{j,k}(x) \stackrel{\text{def.}}{=} 2^{j/2}\varphi\left(2^j x - k\right). \tag{4}$$

1) Show that there exists a space $\mathcal{W}_j \subset \mathrm{L}^2([0,1])$ such that

$$\forall j \geqslant 0, \quad \mathcal{V}_{j+1} = \mathcal{V}_j \oplus^{\perp} \mathcal{W}_j,$$

where one denotes $U = V \oplus^{\perp} W$ if $U = V \oplus W$ (the sub-spaces are in direct sums) and $V \perp W$ (the sub-spaces are orthogonal). What is the dimension of $\mathcal{W}_j$?

2) The function $(\psi_{j,k})_{j,k}$ are called "Haar wavelets".

a) Draw the graphs of the functions $\psi_{0,0}, \psi_{1,0}$ are $\psi_{2,2}$.

b) Show that

$$\mathcal{B}_j^{\varphi} \stackrel{\text{def.}}{=} \left\{\varphi_{j,k}\ ;\ k \in \{0,\dots,2^j-1\}\right\} \quad \text{et} \quad \mathcal{B}_j^{\psi} \stackrel{\text{def.}}{=} \left\{\psi_{j,k}\ ;\ k \in \{0,\dots,2^j-1\}\right\}$$

are ortho-bases of respectively $\mathcal{V}_j$ and $\mathcal{W}_j$.

c) For $j_0 \in \mathbb{N}$, show that

$$\{\varphi\} \cup \bigcup_{j<j_0} \mathcal{B}_j^{\psi}$$

is an ortho-basis of $\mathcal{V}_{j_0}$.

d) Show that the projection of $f \in \mathrm{L}^2([0,1])$ on $\mathcal{V}_j$ converges to zero as $j \to +\infty$. One can start by the case of a continuous function f.

e) Show that

$$\{\varphi\} \cup \bigcup_{j\in\mathbb{N}} \mathcal{B}_j^{\psi}$$

is an hilbertian basis of $\mathrm{L}^2([0,1])$.

3) For $f \in \mathrm{L}^2([0,1])$, and $j \in \mathbb{N}$, one denotes

$$\forall\, k \in \{0,\dots,2^j-1\}, \quad a_{j,k} \stackrel{\text{def.}}{=} \langle f, \varphi_{j,k}\rangle \quad \text{and} \quad d_{j,k} \stackrel{\text{def.}}{=} \langle f, \psi_{j,k}\rangle,$$

which defines two vectors $a_j \stackrel{\text{def.}}{=} (a_{j,k})_{k=0}^{2^j-1} \in \mathbb{R}^{2^j}$ and $d_j \stackrel{\text{def.}}{=} (d_{j,k})_{k=0}^{2^j-1} \in \mathbb{R}^{2^j}$. We suppose in this question that $f \in \mathcal{V}_{j_0}$ for $j_0 \in \mathbb{N}$ and denote $n \stackrel{\text{def.}}{=} 2^{j_0}$.

a) What is the value of d_j for $j > j_0$?

b) How to compute a_{j_0} as a function of $(f(k/n))_{k=0}^{n-1}$?

c) Write, for all $j \in \mathbb{N}$ and $k \in \{0, \ldots, 2^j - 1\}$, $(a_{j,k}, d_{j,k})$ as a function of $(a_{j+1,2k}, a_{j+1,2k+1})$. One can start by writing the function $\varphi_{j,k}$ et $\psi_{j,k}$ as linear combinations of the functions $\varphi_{j+1,2k}$ and $\varphi_{j+1,2k+1}$.

d) Describe an algorithm which computes the whole set of coefficients

$$\{d_j\}_{j=0}^{j_0-1} \cup \{a_0\}$$

starting from a_{j_0} as input data.

e) What is the number of operations (additions and multiplications) involved in this algorithms ?

f) Show that the transformation

$$\mathcal{H}_{j_0} : a_{j_0} \ \in \mathbb{R}^n \longmapsto \{d_j\}_{j=0}^{j_0-1} \cup \{a_0\}$$

defines an orthogonal linear map. It is called the discrete Haar transform.

g) What is the number of operations (additions and multiplications) necessary to compute the dot product Ha between an arbitrary matrix $H \in \mathbb{R}^{n \times n}$ and an arbitrary vector $a \in \mathbb{R}^n$? Compare this to the complexity involved in the computation of $\mathcal{H}_{j_0}(a_0)$.

4) Discussion : contrast the Haar transform with the Walsh transform described in Part 2. You can for instance describe how they behave on simple vectors, compare the structure of their algorithms, the type of group action involved, etc.

Part 4 – Group Lasso

For some matrix $A \in \mathbb{R}^{p \times n}$ and vector $y \in \mathbb{R}^p$, we recall that the lasso problem reads

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda \|x\|_1 \quad \text{where} \quad \|x\|_1 \stackrel{\text{def.}}{=} \sum_{i=1}^{n} |x_i|. \tag{5}$$

We consider a partition $\{1, \ldots, n\} = \cup_{k=1}^{K} B_k$ where $\mathcal{B} = \{B_k\}_k$ is a set of non-intersecting groups of index. The associated group-ℓ^1 norm is

$$\|x\|_{\mathcal{B}} \stackrel{\text{def.}}{=} \sum_{k=1}^{K} \|x_{B_k}\|$$

where $\|u\| = \sqrt{\sum_j u_j^2}$ is the Euclidean norm and $x_B = (x_i)_{i \in B} \in \mathbb{R}^{|B|}$ selects the entries of x indexed by B. The group Lasso problem reads

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - y\|^2 + \lambda \|x\|_{\mathcal{B}}. \tag{6}$$

1) Explain why (5) is a special case of (6). Show that (6) is a convex problem which has at least a solution.

2) We recall that the proximal operator of a convex function f is

$$\forall\, x \in \mathbb{R}^n, \quad \mathrm{Prox}_f(x) \stackrel{\text{def.}}{=} \operatorname*{argmin}_{x' \in \mathbb{R}^n} \frac{1}{2}\|x - x'\|^2 + f(x').$$

Compute, for $\tau > 0$, $\mathrm{Prox}_{\tau\|\cdot\|^2}$ and $\mathrm{Prox}_{\tau\|\cdot\|}$. Compute $\mathrm{Prox}_{\tau\|\cdot\|_{\mathcal{B}}}$.

3) Recall the iterative thresholding algorithm to minimize (5). Adapt it to minimize (6).

4) Explain intuitively why the solution of (6) is expected to have a "group sparse" property (you should give a precise meaning to this naming).

Part 5 – Unbalanced Optimal Transport

Given two input vectors $(a, b) \in \mathbb{R}^n_{+,*} \times \mathbb{R}^m_{+,*}$ (we impose here that these vector are strictly positive). We recall that the discrete entropy-regularized optimal transport problem is defined as

$$\min_{P \in \mathbb{R}^{n \times m}_+} \left\{ \langle C,\, P \rangle - \varepsilon H(P) \;;\; P \mathbb{1}_m = a, P^\top \mathbb{1}_n = b \right\}, \tag{7}$$

for some cost matrix $C \in \mathbb{R}^{n \times m}$, where the entropy H is defined in (1).

We define an "unbalanced" optimal transport problem, for $\rho \geqslant 0$, as

$$\mathcal{P}_\rho = \min_{P \in \mathbb{R}^{n \times m}_+} \langle C,\, P \rangle - \varepsilon H(P) + \rho \,\mathrm{KL}(P \mathbb{1}_m | a) + \rho \,\mathrm{KL}(P^\top \mathbb{1}_n | b) \tag{8}$$

where $\mathrm{KL}(r|a)$ is the Kulback-Leibler divergence defined in (2).

1) What happens in (7) when $\sum_i a_i \neq \sum_j b_j$? What happens in (8) ?

2) Shows that (8) has a unique minimizer P_ρ, which converges, as $\rho \to +\infty$ to the unique minimizer of (7).

3) Compute the Legendre-Fenchel transform $\mathrm{KL}^*(\cdot|a)$ of the function $\mathrm{KL}(\cdot|a)$, defined as

$$\mathrm{KL}^*(u|a) = \max_{r \in \mathbb{R}^n} \langle u,\, r \rangle - \mathrm{KL}(r|a).$$

In the following, we admit that $(\mathrm{KL}^*)^* = \mathrm{KL}$.

4) Using the previous computation, by exchanging a min and a max, show that one has the following dual problem

$$\mathcal{P}_\rho = \max_{(f,g) \in \mathbb{R}^n \times \mathbb{R}^m} -\Phi(f) - \Psi(g) - \varepsilon \sum_{i,j} \exp\left(\frac{-C_{i,j} + f_i + g_j}{\varepsilon} \right), \tag{9}$$

for some convex functions Φ, Ψ to be determined. How does one compute the solution P_ρ of (8) from the optimal (f, g) ?

5) For a fixed g (resp. f), compute the solution of (9) when performing the minimization only with respect to f (resp. g). Deduce an iterative minimization algorithm which performs a minimization with respect to f and g alternatively. Explain why, when $\rho \to +\infty$, one retrieves Sinkhorn's algorithm.