Machine learning methods to help classify liver tumors from radiomics data

Alexandre SELVESTREL[1]

[1]Laboratoire des systèmes, Orsay France
[2]Centrale-Supélec, Gif-sur-Yvette France

Synthèse

L'objectif de mon stage était de réaliser une classification automatique (via du machine learning) de tumeurs du foie basée sur des radios et sur quelques données cliniques (âge, sexe du patient ...). Cette classification a permis de tester des modèles tensoriels [1,2] et de vérifier si ceux-ci donnaient de meilleures performances que les autres modèle. Ce stage était effectué au laboratoire des systèmes (l2S) en partenariat avec l'assistance publique des hôpitaux de Paris (AP-HP). Sur le versan médical, nous avons pu bénéficier de l'aide de Sébastien Mulé, Maître de conférence à la faculté de santé, Université Paris-Est Créteil (UPEC) et Radiologie, chef du département imagerie de l'hôpital Henri Mondor.

Enjeux: Ces stage s'inscrit dans la cadre de la collaboration entre le l2S et l'AP-HP et vient plus généralement renforcer les liens entre le laboratoire et le monde médical. Du point de vue du l2S, il s'agit de mettre à l'épreuve une méthode de machine learning particulière, basée sur des modèles tensoriels et qui semble spécifiquement adaptée aux données étudiées. Or, cette méthode n'a pas encore été beaucoup utilisée, en particulier dans le domaine de la santé. Dans le cas où des avancées théoriques seraient proposées sur cette méthode, celles-ci pourraient faire l'objet d'une publication scientifique future. Par ailleurs, en me formant au machine learning appliqué au domaine médical, le laboratoire s'assure dès le début du stage qu'en poursuivant en doctorat, je disposerai des compétences nécessaires pour être immédiatement opérationnel.

Pour l'AP-HP, l'enjeu est de faire progresser la recherche sur le cancer du foie. En effet, la détermination de la nature de la tumeur du foie d'un patient est un problème complexe auquel il n'existe pas de solution complètement satisfaisante à l'heure actuelle. Or, les médecins disposant des radios des patients malades, il serait dommage de ne pas les utiliser pour tenter de proposer un outil de diagnostic automatique. Même dans le cas où cet outil serait moins performant que ce qui existe déjà, il pourraît être utile aux médecins pour déterminer de nouveaux indices qui caractérisent la classe d'une tumeur.

Solutions et résultats: Nous avons commencé par implémenter des modèles statistiques basics (régression logistique lasso et random forest) sur les données médicales. Cela nous a permis de prendre connaissance des données, d'établir un protocole pour comparer les différents modèles (en se basant sur l'area under curve: AUC) et d'établir une valeur de référence pour la performance de la classification (AUC = 0.68). Nous avons ensuite cherché à améliorer ce score en prorammant une régression logistique tensorielle (voir la section "Méthodes"). Mais malgré une tentative d'amélioration du modèle en séparant les variables en plusieurs blocs, aucun gain de perfrmance n'était observé.

Afin de vérifier que notre nouveau modèle était pertinent nous avons ensuite cherché à tester son efficacité sur des données simulées. Dans ce cas, notre modèle tensoriel a montré des performances bien meilleures (+ 0.1 d'AUC en moyenne) que les modèles non tensoriels. Cela nous a permis de conclure que notre modèle était pertinent et que le manque de performance observé sur les données médicales était probablement dû à un mauvais pré-traitement des données.

Enfin, nous avons approfondi les méthodes d'extraction des features à partir des radios et identifié des pistes d'amélioration pour les scripts pyradiomics utilisés. En changeant l'extraction des features nous sommes passé à une (AUC = 0.85) sur données médicales pour un modèle de régression logistique lasso. Notre modèle tensoriel quant à lui a permis d'obtenir une AUC de ...

Contents

1 Introduction

Il existe deux grands types de tumeur du foie: les carcinome hépatocellulaire (CHC) et les cholangiocarcinomes (CCK). Certaines tumeurs présentent même des caractéristiques CCK et CHC selon l'endroit du foie observé et sont alors dites mixtes. Pour les médecins, déterminer la classe d'une tumeur du foie est très important car le traitement à choisir dépend de cette classe. Pour l'instant, deux grandes approches sont à leur disposition: la microscopie et la radiographie avec injection de produit contrastant.

La microscopie est la méthode la plus fiable car elle permet de directement analyser les cellules tumorales. Toutefois, puisqu'elle nécessite de prélever un petit morceau de foie cancéreux, elle demande une opération et peut entraîner des complications chez le patient. De plus, étant donné qu'on a seulement accès à un fragment du foie, on peut généraliser à tort la nature de tumeur détectée à l'intégralité du foie. Cela peut être gênant dans le cas des tumeurs mixtes où certaines zones cancéreuses sont CHC et d'autres sont CCK.

La radiographie avec injection de produit contrastant est au contraire non invasive. L'idée est d'observer les différents réhaussements lumineux du tissus tumoral en fonction du temps après injection du produit contrastant. En plus d'être non invasive, elle a l'avantage de donner accès à la tumeur en 3D dans son intégralité. C'est pour cela que les médecins (et l'AP-HP) cherchent à la développer davantage. Cependant, les images obtenues par radio ne permettent pas de déterminer avec certitude la nature de la tumeur. En effet, les caractéristiques des tumeurs CHC et CCK sont souvent très proches et les experts ne sont pas toujours d'accord entre eux lorsqu'ils analysent les images. Par ailleurs, de manière générale, il n'est pour l'instant pas possible de poser un diagnostique fiable sur une tumeur du foie à partir de méthodes non invasives seules [3].

Cependant, les limitations que nous avons présentées pour la radiographie ne concernent que les analyses d'images effectuées à l'oeuil nu. On peut donc espérer s'en affranchir en utilisant du machine learning. Il s'agit précisément de l'enjeu de ce stage. L'objectif est d'élaborer un modèle capable de classer le mieux possible les tumeurs à partir de radios et d'informations cliniques sur les patients (âge et sexe). On souhaite que le modèle soit facilement interprétable. En effet, nous devons aider les médecins à déterminer quels critères sont clefs dans la classification, ce qu'un modèle "boîte noire" ne permettrait pas. On préfère donc utiliser du machine learning classique par rapport au deep learning.

Cependant, avant même de parler de modèle de Machine Learning, il faut choisir une méthode d'extraction des données. En effet, les modèles de machine learning classique sont généralement assez mauvais en traitement d'images brutes et il faut donc choisir une manière d'extraire les features intéressantes des radios. Ce choix n'a rien d'évident et il n'y a à priori pas de "bonne manière de faire". Par exemple, on peut extraire des grandeurs "globales", qui décrivent la tumeur dans son intégralité, en 3D, (volume, luminosité globale etc...) ou au contraire extraire des informations coupe par coupe dans la tumeur (surface de tumeur dans la coupe, luminosité de la tumeur dans la coupe etc...). Cette deuxième méthode permet de récupérer davantage de données par individu mais celles-ci risquent d'être redondantes d'une coupe à l'autre pour une même tumeur. On peut aussi ajouter qu'il y a un nombre non négligeable de données manquantes, par exemple liées au fait que le patient a pu bouger durant une radio -la rendant inutilisable- et qu'il faut s'y adapter. Pour réaliser l'extraction des features, des scripts Python utilisant la bibliothèque pyradiomics [4] avaient déjà été écrits par mes encadrants avant mon stage mais ils n'étaient pas optimaux. Ainsi, une partie importante de mon travail a consisté à les améliorer.

Concernant le modèle en lui-même, nous avons décidé qu'il devrait tenir compte du fait qu'à chaque radio, pour un même patient, ce sont les mêmes features qui sont observées à des temps différents. Dans le cas d'une extraction coupe par coupe, on veut aussi qu'il tienne compte du fait que ce sont les mêmes grandeurs qui sont évaluées à différents niveaux de "profondeur" d'une même tumeur. Ce type de modèles est appelé modèle tensoriel [1, 2]. Cependant, afin de coller au mieux à la structure des données, nous introduisons

une extension d'un de ces modèles, en regroupant les variables par blocs. Cette extension sera développée dans la partie "Méthode" de ce rapport. Le but de ce stage est donc double: d'une part il s'agit de chercher à améliorer la classification des tumeurs du foie à partir de radios et d'informations cliniques, et d'autre part il s'agit de vérifier si l'extension du modèle tensoriel que nous proposons peut réellement apporter une plus-value par rapport aux modèles déjà existants.

2 Methodology

2.1 Notations and introdution to tensorial data

We will denote as a tensor any multidimensionnal array, i.e. $\underline{\mathbf{X}} = (x_{i_1 i_2 \ldots i_m})_{i_1 \in [\![1,I_1]\!], i_2 \in [\![1,I_2]\!], \ldots i_M \in [\![1,I_M]\!]}$. It is the extension of the matrix to any finite dimension. To avoid confusion with the notion of dimension of a vector space, and in line with the conventions promoted in Kolda and Bader [5], these dimensons will be referred to as "mode" in the following and their number will be referred to as the "order" of the tensor.

We designate as tensorial data any data where the explanatory variables are structured along several dimensions. We will also call these dimensions modes in the following. For example, if like in our real data, we measure the same quantities at several fixed times and depths, we say that time and depth are modes in our data. Then, instead of having a matrix of explanatory variables $\mathbf{X} = (x_{ij})_{i \in [\![1,n]\!], j \in [\![1,J]\!]}$ (where i is the individual and j is the quantity of interest), we get a tensor of explanatory variables $\underline{\mathbf{X}} = (x_{ijk_1k_2 \ldots k_M})_{i \in [\![1,n]\!], j \in [\![1,J]\!], k_1 \in [\![1,K_1]\!] \ldots k_m \in [\![1,K_M]\!]}$ (where i is the individual, j is the quantity of interest and where for $m \in [\![1,M]\!]$, k_m is the k_m-th modality of the m-th mode of the data). In accordance with the definitions given above, we can also say that the number of the individual and the number of the quantity of interest are modes of the tensor $\underline{\mathbf{X}}$.

In terms of notations, we use those of Kolda and Bader [5], especially concerning matricization (see section 2.4 of [5]). However, as some details need to be precised, we do this here:

- The concatenation of two matrices $\mathbf{A}$ and $\mathbf{B}$ by juxtaposing their columns side by side is denoted $[\mathbf{A}\ \mathbf{B}]$.
- To avoid overuse of the symbol ^{T}, we also define a notation to designate the juxtaposition of two matrices one below the other. Thus, the matrix defined by block with $\mathbf{A}$ above $\mathbf{B}$ is denoted $[\mathbf{A}; \mathbf{B}]$. It can also be written $[\mathbf{A}^T\ \mathbf{B}^T]^T$ but this multiplies the ^{T} symbols, which impairs legibility.
- Since vectors are column matrices, using the same notation, we write the concatenation of two vectors $\mathbf{u}$ and $\mathbf{v}$ as follows: $[\mathbf{u}; \mathbf{v}]$.
- The vector (column) whose elements are $(u_i)_{i \in [\![1,I]\!]}$ is also denoted $(u_1, u_2, \ \cdots\ u_I)$.
- If $\mathbf{X}$ is a matrix of explanatory variables, $\mathbf{x}_i$ is the vector (column) composed of the i-th row of $\mathbf{X}$.
- The vector of length I filled with 1 is denoted by $\mathbb{1}_I$.
- We denote Diag($\mathbf{u}$) the diagonal matrix whose diagonal is the vector $\mathbf{u}$.

2.2 Features extraction

2.2.1 Presentation of real data

2.2.2 feature extraction in 3D

2.2.3 Feature extraction in 2D

2.3 Simulated data

2.4 Machine learning models

In this section, we describe all the machine learning methodes that we used and compared in order to get our results. We start briefly by non tensorial methods and then we describe in details the tensorial methods that we used. For the sake of simplicity, we only describe the situation where $\underline{\mathbf{X}}$ is a tensor of order 3. However, all the methods described here can be generalized to any order of tensor.

2.4.1 Non tensorial methods

For these methods, we start by unfolding the tensorial data $\underline{\mathbf{X}}$ into the matrix $\mathbf{X}_{(1)} = [\mathbf{X}_{::1} \ \dots \ \mathbf{X}_{::K}]$. We then complete this matrix by concatenating (along the columns) the matrix of non tensorial data $\mathbf{X}_{\text{tab}}$ (where "tab" stands for "tabular"). By doing so we obtain $\mathbf{X}_{\text{tot}} = [\mathbf{X}_{(1)} \ \mathbf{X}_{\text{tab}}]$.

We first train a penalized logistic regression lasso on $\mathbf{X}_{\text{tot}}$. Then, still based on the matrix $\mathbf{X}_{\text{tot}}$, we train a group lasso [6]. The idea of this second model is to group together all the explanatory variables corresponding to a same slice of $\underline{\mathbf{X}}$. We do this first for slices at given j and then for slices at given k. The goal is to give the model a vague description of the tensorial structure of the data. Finally, because we also identified blocs of variables in our tensorial data, we try a last group lasso by grouping the variables along these blocs.

2.4.2 Multiway logistic regression with lasso

We now turn to tensor approaches. We start by studying a multiway logistic regression penalized by lasso. This model is described for rank 1 in Le Brusquet et al. [1] and in Girka et al. [2] for its extension to rank $R \in \mathbb{N}^*$. In this report, we directly describe the generalization to rank $R \in \mathbb{N}^*$, rank 1 being a special case of this model.

The fundamental idea of the model is to decompose the parameter $\boldsymbol{\beta}_{\text{tens}} \in \mathbb{R}^{JK}$ associated with the tensor explanatory variables of the logistic regression as:

$$\boldsymbol{\beta}_{\text{tens}} = \sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \tag{1}$$

with for all $r \in [\![1, R]\!]$, $\boldsymbol{\beta}_r^J \in \mathbb{R}^J$ and $\boldsymbol{\beta}_r^K \in \mathbb{R}^K$. To take account of the M clinical variables, which are not tensorial, we associate them with a coefficient $\boldsymbol{\beta}_{\text{tab}} \in \mathbb{R}^M$. In this way, the parameter $\boldsymbol{\beta}$ of the logistic regression is written: $[\boldsymbol{\beta}_{\text{tens}}; \boldsymbol{\beta}_{\text{tab}}]$.
As usual with logistic regressions, we consider that each realization of the explained variable y_i ($i \in [\![1, n]\!]$)

follows an independent Bernoulli law conditionally on $\mathbf{x_i}$. This law is defined by:

$$\mathbb{P}(y_i = 1 \,|\, \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta} - \beta_0)} \tag{2}$$

where $\beta_0 \in \mathbb{R}$ is the intercept

We set $\boldsymbol{\beta}^J = \left[\boldsymbol{\beta}_1^J;\ \cdots\ ;\ \boldsymbol{\beta}_R^J\right]$ and $\boldsymbol{\beta}^K = \left[\boldsymbol{\beta}_1^K;\ \cdots\ ;\ \boldsymbol{\beta}_R^K\right]$. In order to simplify the calculations, while ensuring that the penalty continues to promote sparse models, we adapt the definition of the lasso penalty. The new penalty defines the following optimization problem:

$$\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}} = \underset{\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}}}{\operatorname{argmin}} \left(-\sum_{i=1}^{N} \log(\mathbb{P}(y_i = 1 \,|\mathbf{x}_i)) + \sum_{r=1}^{R} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 \right) \tag{3}$$

Optimization is performed by alternating directions between $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{uni}}\right]$ and $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{uni}}\right]$. The stopping criterion is defined by the relative difference between the value of the objective function after optimization in the first direction and the value of the same function after optimization in the second direction. We note that optimizing the loss function in each of these directions is tantamount to performing a simple logistic regression with a lasso penalty. Indeed, if we denote C the loss function of classical logistic regression penalized by lasso (for any $K_0 \in \mathbb{N}^*$):

$$C : \begin{cases} \mathbb{R} \times \mathbb{R}^{K_0} \times \mathbb{R}^{N \times K_0} \times \mathbb{R}^N \times \mathbb{R} & \longrightarrow & \mathbb{R} \\ (\beta_0, \boldsymbol{\beta}, \mathbf{X}, \mathbf{y}, \lambda) & \longmapsto & -\sum_{i=1}^{N} [y_i(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}) - \log(1 + \exp(\beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}))] + \lambda \|\boldsymbol{\beta}\|_1 \end{cases} \tag{4}$$

optimizing the overall loss function with respect to $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{uni}}\right]$ amounts to solve

$$\underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{JR+M}}{\operatorname{argmin}} \; C(\beta_0, (\mathbf{Q}^J)^{-1} \boldsymbol{\beta}, \mathbf{Z}^J \mathbf{Q}^J, \mathbf{y}, \lambda) \tag{5}$$

Where $\mathbf{Q}^J$ and $\mathbf{Z}^J$ are defined as follows:

$$\mathbf{Z}^J = [\mathbf{Z}_1^J \ \cdots \ \mathbf{Z}_R^J \ \ \mathbf{X}_{\text{tab}}] \tag{6}$$

$$\text{where } \forall r \in [\![1, R]\!], \qquad \mathbf{Z_r^J} = \sum_{k=1}^{K} \mathbf{X}_{::k} (\beta_r^K)_k \qquad (\mathbf{Z}_r^J \in \mathbb{R}^{N \times J}) \tag{7}$$

$$\mathbf{Q}^J = \text{Diag}([\|\boldsymbol{\beta}_1^K\|_1^{-1} \mathbb{1}_J;\ \cdots\ ;\ \|\boldsymbol{\beta}_R^K\|_1^{-1} \mathbb{1}_J;\ \mathbb{1}_M]) \tag{8}$$

Girka et al. [2] demonstrate this result by noting that for $i \in [\![1, n]\!]$,

$$\mathbf{x}_{(1)i}^T \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \right) = \sum_{r=1}^{R} \left[(\mathbf{x}_{(1)i}^T \left(\boldsymbol{\beta}_r^K \otimes \mathbf{I}_J \right) \right] \boldsymbol{\beta}_r^J \tag{9}$$

$$= \sum_{r=1}^{R} (\mathbf{z}_r^J)_i^T \boldsymbol{\beta}_r^J \tag{10}$$

and that

$$\sum_{r=1}^{R} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 = \|\mathbf{R}_{\text{tens}}^J \boldsymbol{\beta}^J\|_1 \tag{11}$$

$$\text{with} \quad \mathbf{R}_{\text{tens}}^J = \text{Diag}([\|\boldsymbol{\beta}_1^K\|_1 \mathbb{1}_J; \ \cdots \ ; \|\boldsymbol{\beta}_R^K\|_1 \mathbb{1}_J]) \tag{12}$$

Thus,

$$(\mathbf{x}_{\text{tot}})_i^T \boldsymbol{\beta} = (\mathbf{z}_i^J)^T [\boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}] \tag{13}$$

$$\text{and} \quad \sum_{i=1}^{N} \|\boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 = \|(\mathbf{Q}^J)^{-1} \boldsymbol{\beta}\|_1 \tag{14}$$

This justifies the previous results

For optimization with respect to $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$, the method is analogous. The only difference concerns the definition of $\mathbf{Z}^K$. It is:

$$\mathbf{Z}^K = [\mathbf{Z}_1^K \ \cdots \ \mathbf{Z}_R^K \ \ \mathbf{X}_{\text{tab}}] \tag{15}$$

$$\text{with} \forall r \in [\![1, R]\!] \quad \mathbf{Z}_r^K = \sum_{j=1}^{J} (\beta_r^J)_j \mathbf{X}_{:j:} \tag{16}$$

This is justified by:

$$\mathbf{x}_{(1)_i}^T \left(\sum_{r=1}^{R} \boldsymbol{\beta}_r^K \otimes \boldsymbol{\beta}_r^J \right) = \sum_{r=1}^{R} \left[(\mathbf{x}_{(1)_i}^T \left(I_K \otimes \boldsymbol{\beta}_r^J \right) \right] \boldsymbol{\beta}_r^K \tag{17}$$

$$= \sum_{r=1}^{R} (\mathbf{z}_r^K)_i^T \boldsymbol{\beta}_r^K \tag{18}$$

2.4.3 Multiway and multibloc logistic regression with lasso

We now present the lasso-penalized multiway and multiblock logistic regression. This model draws heavily on the multiway logistic regression we have just presented, while also taking into account a block structure of tensor data. More precisely, each of these blocs will have its own independent coefficient $\boldsymbol{\beta}_l$, which was not the case in the previous model. As tabular quantities are not measured at different times, they are not placed in any particular block. They will be included in the model in the same way as in the multiway case. Mathematically, we define the model as follows:

Let $L \in \mathbb{N}^*$ denote the number of blocks of variables. For any $l \in [\![1, L]\!]$, let d_l be the number of tensorial quantities of interest in block l. Thus we have :

$$\sum_{l=1}^{L} d_l = J$$

We reorganize $\underline{\mathbf{X}}$ by grouping together slices $\mathbf{X}_{:j:}$ associated with quantities from the same block. More precisely, for all $l \in [\![1, L]\!]$, we call $\underline{\mathbf{X}}^l$ the tensor constituted by the slices $\mathbf{X}_{:j:}$ associated with the l-th bloc.

We then concatenate these tensors along their second mode to obtain the new tensor of explanatory variables: $\underline{\mathbf{X}}'$.

The new $\boldsymbol{\beta}$ structure is defined by blocks. It is:

$$\boldsymbol{\beta} = \left[\sum_{r=1}^{R} \boldsymbol{\beta}^K_{(1,r)} \otimes \boldsymbol{\beta}^J_{(1,r)};\ \cdots\ ; \sum_{r=1}^{R} \boldsymbol{\beta}^K_{(L,r)} \otimes \boldsymbol{\beta}^J_{(L,r)}\ ; \boldsymbol{\beta}_{\text{tab}} \right] \tag{19}$$

With for all $(l,r) \in [\![1,L]\!] \times [\![1,R]\!]$, $\boldsymbol{\beta}^J_{(l,r)} \in \mathbb{R}^{d_l}$ and $\boldsymbol{\beta}^K_{(l,r)} \in \mathbb{R}^K$

We call $\boldsymbol{\beta}^J$ and $\boldsymbol{\beta}^K$ the vectors

$$\boldsymbol{\beta}^J = \left[\boldsymbol{\beta}^J_{(1,1)};\ \cdots\ ; \boldsymbol{\beta}^J_{(1,R)};\ \cdots\ \cdots\ ; \boldsymbol{\beta}^J_{(L,1)}\ \cdots\ ; \boldsymbol{\beta}^J_{(L,R)} \right] \tag{20}$$

$$\boldsymbol{\beta}^K = \left[\boldsymbol{\beta}^K_{(1,1)};\ \cdots\ ; \boldsymbol{\beta}^K_{(1,R)};\ \cdots\ \cdots\ ; \boldsymbol{\beta}^K_{(L,1)}\ \cdots\ ; \boldsymbol{\beta}^K_{(L,R)} \right] \tag{21}$$

In a similar way to what is done in the multiway model, we adapt the lasso penalty, so that the new optimization problem becomes:

$$\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}} = \underset{\beta_0, \boldsymbol{\beta}^J, \boldsymbol{\beta}^K, \boldsymbol{\beta}_{\text{tab}}}{\operatorname{argmin}} \left(-\sum_{i=1}^{N} \log(\mathbb{P}(y_i = 1 \,|\mathbf{x}_i)) + \sum_{l=1}^{L} \sum_{r=1}^{R} \|\boldsymbol{\beta}^K_{(l,r)} \otimes \boldsymbol{\beta}^J_{(l,r)}\|_1 + \|\boldsymbol{\beta}_{\text{tab}}\|_1 \right) \tag{22}$$

Once again, this problem is solved by alternating optimization directions $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}\right]$ and $\left[\beta_0; \boldsymbol{\beta}^K; \boldsymbol{\beta}_{\text{tab}}\right]$. Each of these two problems can be reduced to a lasso-penalized classical logistic regression.
Indeed, optimizing according to $\left[\beta_0; \boldsymbol{\beta}^J; \boldsymbol{\beta}_{\text{tab}}\right]$ is equivalent to searching

$$\underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{RJ+M}}{\operatorname{argmin}} \quad C(\beta_0, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda) \tag{23}$$

Where $\mathbf{Q}^J$ and $\mathbf{Z}^J$ are defined as follows:

$$\mathbf{Z}^J = [\mathbf{Z}^J_{(1,1)}\ \cdots\ \mathbf{Z}^J_{(1,R)}\ \cdots\ \cdots\ \mathbf{Z}^J_{(L,1)}\ \cdots\ \mathbf{Z}^J_{(L,R)}\ \ \mathbf{X}_{\text{tab}}] \tag{24}$$

$$\text{where } \forall r \in [\![1,R]\!], \qquad \mathbf{Z}^J_{(l,r)} = \sum_{k=1}^{K} \mathbf{X}^l_{::k} \left(\beta^K_{(l,r)}\right)_k \qquad \left(\mathbf{Z}^J_{(l,r)} \in \mathbb{R}^{n \times d_l}\right) \tag{25}$$

$$\mathbf{Q}^J = \text{Diag}([\|\boldsymbol{\beta}^K_{(1,1)}\|_1^{-1}\mathbb{1}_{d_1};\ \cdots\ ; \|\boldsymbol{\beta}^K_{(1,R)}\|_1^{-1}\mathbb{1}_{d_1};\ \cdots\ \cdots\ ; \|\boldsymbol{\beta}^K_{(L,1)}\|_1^{-1}\mathbb{1}_{d_L};\ \cdots\ ; \|\boldsymbol{\beta}^K_{(L,R)}\|_1^{-1}\mathbb{1}_{d_L};\ \mathbb{1}_M]) \tag{26}$$

The demonstration of this result is similar to that of the multiway case. Indeed, we note that

$$(\mathbf{x}'_{(1)})_i^T \left[\sum_{r=1}^{R} \boldsymbol{\beta}^K_{(1,r)} \otimes \boldsymbol{\beta}^J_{(1,r)};\ \cdots\ ; \sum_{r=1}^{R} \boldsymbol{\beta}^K_{(L,r)} \otimes \boldsymbol{\beta}^J_{(L,r)} \right] = \sum_{l=1}^{L} \sum_{r=1}^{R} \left(\mathbf{x}^l_{(1)}\right)_i^T \left(\boldsymbol{\beta}^K_{(l,r)} \otimes \boldsymbol{\beta}^J_{(l,r)}\right) \tag{27}$$

$$= \sum_{l=1}^{L} \sum_{r=1}^{R} \left[\left(\mathbf{x}^l_{(1)}\right)_i^T \left(\boldsymbol{\beta}^K_{(l,r)} \otimes I_{d_l}\right) \right] \boldsymbol{\beta}^J_{(l,r)} \tag{28}$$

$$= \sum_{l=1}^{L} \sum_{r=1}^{R} \left(\mathbf{z}^J_{(l,r)}\right)_i^T \boldsymbol{\beta}^J_{(l,r)} \tag{29}$$

An that

$$\sum_{l=1}^{L}\sum_{r=1}^{R}\|\boldsymbol{\beta}^K_{(l,r)}\otimes\boldsymbol{\beta}^J_{(l,r)}\|_1 = \|\mathbf{R}^J_{\text{tens}}\boldsymbol{\beta}^J\|_1 \tag{30}$$

$$\text{with}\quad \mathbf{R}^J_{\text{tens}} = \text{Diag}([\|\boldsymbol{\beta}^K_{(1,1)}\|_1\mathbb{1}_{d_1};\ \cdots;\|\boldsymbol{\beta}^K_{(1,R)}\|_1\mathbb{1}_{d_1};\ \cdots\ \cdots;\|\boldsymbol{\beta}^K_{(L,1)}\|_1\mathbb{1}_{d_L};\ \cdots;\|\boldsymbol{\beta}^K_{(L,R)}\|_1\mathbb{1}_{d_L};\mathbb{1}_M]) \tag{31}$$

We deduce that

$$[\mathbf{x}'_{(1)\,i};\mathbf{x}_{\text{tab}_i}]\boldsymbol{\beta} = (\mathbf{z}^J_i)^T[\boldsymbol{\beta}^J;\boldsymbol{\beta}_{\text{tab}}] \tag{32}$$

$$\text{and}\quad \sum_{l=1}^{L}\sum_{r=1}^{R}\|\boldsymbol{\beta}^K_{(l,r)}\otimes\boldsymbol{\beta}^J_{(l,r)}\|_1 + \|\boldsymbol{\beta}_{\text{uni}}\|_1 = \|(\mathbf{Q}^J)^{-1}\boldsymbol{\beta}\|_1 \tag{33}$$

Wich justifies the previous results.

For optimization with respect to $\left[\beta_0;\boldsymbol{\beta}^K;\boldsymbol{\beta}_{\text{tab}}\right]$, the method is analogous. The only difference concerns the form of $\mathbf{Z}^K$. It is written as:

$$\mathbf{Z}^K = [\mathbf{Z}^K_{(1,1)}\ \cdots\ \mathbf{Z}^K_{(1,R)}\ \cdots\ \cdots\ \mathbf{Z}^K_{(L,1)}\ \cdots\ \mathbf{Z}^K_{(L,R)}\ \mathbf{X}_{\text{tab}}] \tag{34}$$

$$\text{where } \forall r\in[\![1,R]\!],\quad \mathbf{Z}^K_{(l,r)} = \sum_{j=1}^{d_l}\mathbf{X}^l_{:j:}\left(\beta^J_{(l,r)}\right)_j \qquad \left(\mathbf{Z}^K_{(l,r)}\in\mathbb{R}^{N\times K}\right) \tag{35}$$

The justification of that last result is analogous to the one used in the multiway case.

Notes:

- In the multiway and multibloc logistic regression with lasso, it is possible to allow for the choice of a specific rank for each block. We have not yet studied this model extension, but deriving its equations is straightforward, based on the equations provided in this report.
- We decided to optimize the loss function completely in one direction before turning to the other one instead of alternating one step in each direction because the first procedure was more stable and could be implemented efficiently using the glmnet package in R [7].

Pseudo-code:

In order to clarify the algorithm that we use, we give here the pseudo-code of our implementation. In order to be more readable, we keep the notations that were used during the presentation of the model.

Inputs

- $\epsilon > 0$, $\lambda > 0$, $R \in \mathbb{N}^*$
- $\boldsymbol{\beta}^{K(0)} \in \mathbb{R}^{LRK}$

Treatment

- $q \leftarrow 0$

Repeat

- Construct $\mathbf{Z}^J$ according to eqs. (24) and (25)
- Construct $\mathbf{Q}^J$ according to eq. (26)
- $(\beta_0^{(q)}, \boldsymbol{\beta}^{J(q)}) \longleftarrow \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{RJ+M}}{\operatorname{argmin}} \left(C(\beta_0, (\mathbf{Q}^J)^{-1}\boldsymbol{\beta}, \mathbf{Z}^J\mathbf{Q}^J, \mathbf{y}, \lambda) \right)$
- Construct $\mathbf{Z}^K$ according to eqs. (34) and (35)
- Construct $\mathbf{Q}^K$ by adapting eq. (26)
- $(\beta_0^{(q)}, \boldsymbol{\beta}^{K(q)}) \longleftarrow \underset{(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^{LRK+M}}{\operatorname{argmin}} \left(C(\beta_0, (\mathbf{Q}^K)^{-1}\boldsymbol{\beta}, \mathbf{Z}^K\mathbf{Q}^K, \mathbf{y}, \lambda) \right)$
- $q \leftarrow q + 1$

until $|C^K - C^J| < \epsilon |C^J|$

Return $(\beta_0^{(q)}, \boldsymbol{\beta}^{K(q)}, \boldsymbol{\beta}^{J(q)})$

3 results

Conclusions

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

References

[1] L. Le Brusquet, G. Lechuga, and A. Tenenhaus, "Régression Logistique Multivoie," in *JdS 2014*, (Rennes, France), p. 6 pages, June 2014.

[2] F. Girka, P. Chevaillier, A. Gloaguen, G. Gennari, G. Dehaene-Lambertz, L. Le Brusquet, and A. Tenenhaus, "Rank-R Multiway Logistic Regression," in *52èmes Journées de Statistique*, (Nice, France), 2021. les 52èmes journées de Statistique 2020 sont reportées ! Elles auront lieu du 7 au 11 Juin 2021.

[3] M. Jacquemin, "Performance of imaging in the diagnosis of hepatocellular carcinoma: a single-centre, retrospective series of 167 patients," *HAL Open Science*, 2021.

[4] J. J. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. Aerts, "Computational Radiomics System to Decode the Radiographic Phenotype," *Cancer Research*, vol. 77, pp. e104–e107, 10 2017.

[5] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[6] L. Meier, S. Van De Geer, and P. Bühlmann, "The Group Lasso for Logistic Regression," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 70, pp. 53–71, 01 2008.

[7] R. Tibshirani, T. Hastie, and J. Friedman, "Regularized paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, 02 2010.