Sujets de Projet

Sujet 1

Mining binary data

Nonnegative matrix factorizations (NMF)([1]) have been extensively used to find hidden patterns in nonnegative data. In NMF, J nonnegative features measured for i individuals are stored in the elemets of a matrix $\mathbf{Y}$. Different rows of $\mathbf{Y}$ correspond to data from different individuals, while different columns to different features. The data for each individual (row of $\mathbf{Y}$) are the approximated as weighted combination of R hidden patterns encoded with nonnegative vectors. this approximation can be written as

$$(1) \qquad \mathbf{Y} \approx \mathbf{A}\mathbf{B}^T$$

where the columns of $\mathbf{B}$ contain the hidden patterns and (A) contains the combining weights. The columns of $\mathbf{B}$ can be interpreted as nonnegative clusters hidden in the data and the rows of $\mathbf{A}$ as a measurement of membership of each individual to one of the clusters.

When $\mathbf{Y}$ has binary elements (0/1), for example, answers to "yes or no?" questions, a natural data model is to constrain the clusters and the weights to be binary leading to binary matrix factorizations ([2]). One difficulty with binary factors in (1) is that the corresponding approximation of $\mathbf{Y}$ may not be binary. Solutions for this issue proposed in the literature correspond to replace the sum operations within (1) by logical operations, for example by the logical "OR". The corresponding model is known as Boolean matrix factorization (BMF)([3]). The objective of this project is to propose new algorithms for fitting BMF to data by representing the nonlinear logical "OR" function as a polynomial function. The studentswill first writte the underlying data fitting problem as a penalized nonlinear least-squares problem, then they will develop algorithms based on the gradient descent and/or projected block-coordinate descent to fit the model. Once the algorithms have been tested with simulated data, they will be applied to a dataset of voting records ([4]) to jointly cluster congress representatives and laws.

The algorithms developed in this project will be coded and tested either in scilab or in python language.

Keywords: data mining, matrix factorizations, binary data

References:

[1] Cihocki, A., Zdunek, R., Phan, A. H., & Amari, S. I. (2009) *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Willey & Sons

[2] Zhang, Z., Li, T., Ding, C., & Zhang, X. (2007) *Binary matrix factorization with applications.* IEEE ICDM (pp. 391-400)

[3] Miron, S., Diop, M., Larue, A., Robin, E., & Brie, D. (2021) *Boolean decomposition of binary matrices using a post-nonlinear mixture approach.* Signal Processing. 178, 107809

[4] https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

Sujet 2

Acceleration of Deep Neural Networks using Low Rank Approximations

Deep neural networks (DNN) have achieved state of the art performance on a multitude of tasks such as classification, detection, and image segmentation. However, the growing complexity of such networks, requiring thousands to millions of operations each time the network is used for inference, makes their real-time use in embedded applications very difficult. in such applications, computational resources are limited, as a consequence, respecting real time constraints on latency (waiting time for the computations to finish) becomes a central issue. for this reason, intensive research has been carried out on different methods to reduce the computational complexity of these networks.

On e approach to reduce complexity ([1,2]) relies on the fact that the bulk of operations required for doing inference with a DNN are linear transformations

$$(1) \qquad y_i = W_i x_i$$

where index $i \in \{1, 2, \ldots, N\}$ indicates that these quantities are related to the i-th layer of the network, x_i is a vector of size J_i corresponding to the input of the layer, W_i is a matrix of parameters characterizing the layer and it has size $I_i x J_i$, and y_i is a vector of size I_i required to evaluate the output of the layer. If W_i can be well approximated by its best lower rank-matrix $\hat{W}_i$, with $Rank(\hat{W}_i) = R_i$ where $R_i << I_i,\ J_i$, then the product (1) can be approximated as follows

$$(2) \qquad y_i \approx \hat{W}_i x_i = U_i . \Sigma_i V_i^T x_i$$

where (U_i, Σ_i, V_i) are the truncated factors of the singular value decomposition (SVD) of W_i leading to $\hat{W}_i$. Since the sizes of the truncated factors are respectively $I_i x R_i$, $R_i x R_i$ and $J_i x R_i$, the number of arithmetic operations to evaluate (2) is of order $(I_i + J_i + R_i) x R_i$. for an approximation with small R_i, this number can be much smaller than the number of operations for evaluating (1) which is of order $I_i x J_i$. Clearly, when such an approximation is applied to all the layers of the network, this will lead to an overall reduction of latency when using DNN.

In this project, the students will test this approach in a simple DNN (LeNet5 ([3])) designed for an image classification task (MNIST classification dataset ([4])). The students will first study thr DNN architecture, then they will code it in python using a specialized DNN library, tensorflow2 for example, and finally they will modify it, so its linear transformations are carried out with approximation (2). The students will test different combinations of R_i to check experimentally their effects both on the latency and on the loss of classification performance incurred by the errors introduced by (2).

Keywords: deep learning, low rank approximations, singular value decomposition

References:

[1] Zhang, X., Zou, J., Ming, X., He, K., & Sun, J. (2015). *Efficient and accurate approximations of nonlinear convolutional networks.* In IEEE CVPR (pp. 1984-1992)

[2] Yu, X., Liu, T., Wang, X., & Tao, D. (2017). *On compressing deep models by low rank and sparse decomposition.* In IEEE CVPR (pp. 7370-7379)

[3] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). *Gradient-based learning applied to document recognition.* Proceedings of the IEEE, 86(11), 2278-2324

[4] https://yan.lecun.com/exdb/mnit/

Sujet 3

Détection de tumeurs avec un réseau de neurones en imagerie médicale

Contexte:
Dans le domaine de l'imagerie médicale, le prétraitement des images radiographiques est essentiel afin de détecter efficacement des tumeurs cancéreuses. Il s'agit également de localiser ces tumeurs et d'estimer leurs tailles. De cette manière, les médecins peuvent exploiter ces informations pour diagnostiquer la maladie mais aussi suivre l'/'evolution du traitement.

Dans ce projet, nous chercherons à détecter des tumeurs cancéreuses dans le foie. Nous utiliserons des données en accès libre ([1]) qui ont été annontées par des experts (radiologistes et oncologues) : les tumeurs sont donc connues et localisées. La figure ci-après donne une idée des images qui seront utilisées : nous pouvons voir l'image originale et l'image "mask" qui représente les pixels associés au foie. Cette image "mask" représente l'annotation fournie par les spécialistes. Il convient de construire un algorithme qui permet de retrouver ce masque de manière automatique dans l'image originale. Quelques codes sont déjà disponibles sur le site Kaggle ([1]) pour aider les étudiants à manipuler les images avec Python.

Dans le domaine du traitement d'images, les méthodes les plus efficaces actuellement s'appuient sur des réseaux de neurones profonds ([2]). Le but de ce projet est donc de développer un réseau de neurones profond pour détecter les tumeurs dans les images du foie. De nombreux cours sont disponibles en ligne pour aider les étudiants à prendre en main les réseaux de neurones ([3]). Nous utiliserons Python, et notamment l'environnement Pytorch ([4]) dédié aux réseaux de neurones profonds, pour réaliser ce projet.

Comme les données proviennent d'un chalenge Kaggle en accès libre ([1]), les étudiants pourront éventuellement déposer leur code sur le site Kaggle lorsque le projet sera terminé.

Objectifs:
Le projet est composé de trois tâches principales :

- La première tâche consiste à se familiariser avec les réseaux de neurones profonds et Pytorch. De nombreux documents et codes sont disponibles en ligne ([4])
- la seconde tâche consiste à concevoir et programmer un réseau de neurones profond pour détecter et localiser les tumeurs cancéreuses ([1]). Il s'agira notamment d'identifier quels sont les réseaux de neurones actuellement disponibles qui pourraient efficacement résoudre le problème de détection abordé dans le projet.
- Enfin, les étudants doivent évaleur attentivement les performances de l'algortihme dvélopp/'e pour conclure sur sa pertinence.

Les développements informatique et l'analyse mathématique seont présentés dans un fichier Jupyter Notebook et un rapport technique.

References:

[1] https://www.kaggle.com/andrewmvd/liver-tumor-segmentation

[2] https://neuralnetworksanddeeplearning.com/

[3] https://fleuret.org/dlc/

[4] https://pytorch.org

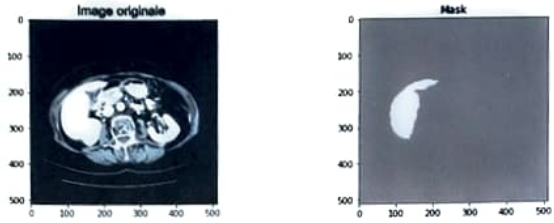

Figure 1 : Exemple de figure à traiter dans le projet.

Sujet 4

Introduction de nouveaux opérateurs rotationnels dans CONVIV

Dans un système moléculaire, les mouvements de vibrations des noyaux atomiques sont responsables des spectres infra-rouges et Raman notamment, utilisés pour la détection et la caractérisation de molécules dans de nombreux domaines fondamentaux (astrophysique, biochimie) et appliqués (suivi de polluants, détection d'incendie, détection à distance de substances toxiques, analyse d'oeuvre d'art anciennes, ...). Les mouvements de rotation moléculaires donnent eux des raies spectrales allant du domaine micro-onde à celui des ondes sub-millimétriques, qui sont de véritables carte d'identité pour les molécules. Ces raies sont notamment observées par les radioastronomes dans de nombreux milieux interstellaires et permettent de remonter à la composition chimique de ces milieux ([1]).

Nous avons développé au laboratoire une méthode pour approcher les solutions de l'équation de Schrodinger stationaire pour l'hamiltonien ro-vibrationnel d'un molÃ©cule, solutions qui permettent de calculer *ab initi* de tesl spectres. Le projet consiste à introduire dans le code CONVIV ([2]), qui implmente cette méthode, le calcul des éléments de matrice de nouveaux opérateurs rotationnels dans une base de fonctions propre de toupie symétrique. Il s'agir d'opérateurs élémentaires sur lesquels on peut décomposer l'hamiltonien ro-vibrationnel moléculaire dans un système de coordonnées curvilignes généralement plus aptes à décrire la réalit/'e physiquue que les coordonnées cartésiennes coupl'ees aux angles d'Euler, utilisées actuellement.

Concrètement, il s'agira dans un premier temps de trouver les formules des éléments de matrice grâce par exemple au code Mathematica, puis dans un second temps de programmer ces formules en Fortran dans CONVIV.

References:

[1] P. Cassam-Chenai, "Séparation et contraction de variables en spectroscopie moléculaire", TI NÂ°AF110, (Techniques de l'Ingénieur, Paris, 10/04/2014), ebook

[2] CONVIV

Sujet 5

Détection d'anomalies dans les séries temporelles

Contexte:
La détection d'anomalies dans des séries temporelles est un sujet de recherche et de développements industriels très intenses. Plusieurs approches sont possibles, des approches statiques, de sapproches algébriques, des approches basées machine learning, etc. Dans le cadre de ce projet, nous étudierons en particulier les approches par apprentissage profond (Deep Learning) et les impacts des différentes étapes d'analyse : préparation des données, extraction de caractéristiques, classification, comparaison.

Objectifs:

1. Comprendre et formaliser les étapes clefs de la détection d'anomalies dans les séries temporelles en reproduisant quelques méthodes de référence
2. Evaleur le rôle et l'impact de chaque étape de l'analyse de données
3. Comparer les méthodes entre elles sur des benchmarks de référence
4. Evaluer les métriques (RMSE? Hamming distance,..) de comparaison des méthodes

Ce projet pourra doner lieu à un stage dans les entreprises Ezako ou NXP et pourrait ensuite donner lieu à une alternance dans l'une de ces entreprises.

Outils informatiques:
Pytorch, Tensorflow, Maven

Mots-clef:
Détection d'anomalies, deep learning, séries temporelles

Références:

[1] Ranjan, C., Mustonen, M., Paynabar, K., & Pourak, K. (2018) *Dataset; Rare Event Classification in Multivariate Time Series.* arXiv preprint arXiv:1809.10717

2] Sabata, T., & Holena, M. (2020) *Active Learning for LSTM-autoencoder-based Anomaly Detection in Electrocardiagram Readings*

[3] https://towardsdatascience.com/extreme-rare-event-classification-using-autoencoders-in-keras-a565b386f098

[4] Bulusu, S., Kaikhura, B., Li, B., Varshney, P. K., & Song, D. (2020) *Anomalous instance detection in deep learning: A survey* arXiv preprint arXiv:2003.06979

Sujet 6

Modeling of protein cavities with 3D point clouds

Context:
3D Point Clouds (or 3D points sets) are a widespread representation in many domains well known by large audience: robotics, 3D reconstruction, games, autonomous navigation, and so on. A little-known fact is that 3D representation can be also relevant in cheminformatics. To find shape similarities between molecules for instance:in that case, molecules are represented by sets of 3D points distributed onto their eternal surface, and then aligned to know if they are similar as illustrated by Figure 1. Promising results have been already reached ([1]) in our team to align molecular structures, by using 3D point clouds of molecualr surfaces (Figure 1). Code is available on Github : https://github.com/SENSAAS/sensaas.

Goal of this project:
To continue exploring this way, we now aim during this student project to

- develop a short algorithm (in Python) to generate the point-based surface of a concave surface such as a protein cavity (Figure 2), by using local 3D descriptors such as FPFH ([2]), and/or (depending on efficiency of the students, and the number of student on the project ...)
- study Caviar ([3]), an open source tool for protein cavity identification and rationalization, written un Python : https://github.com/jr-marchand/caviar

Keywords:
3D point clouds, modeling, 3D descriptors/protein cavity identification

References:

[1] SensAAS: Shape-based Alignment by Registration of Colored Point-based Surfaces, Dominique Douguet, Frédéric Payan Molecular Informatics (Wiley), June, 2020

[2] Radu Bogdan Rusu, Nico Blodov, and Michael Beetz. 2009. Fast pooint feature histograms (FPFH) for 3D registration. in Proceedings of the 2009 IEEE international conference on Robotics and Automation (ICRA'09)

[3] CAVIAR: a method for automatic cavity detection, description znd decomposition into subcavities, Jean-Rémy Marchand, Bernanrd Pirard, Peter Ertl, Finton Sirockin, 10.26434/chemrxiv.12806819.v3

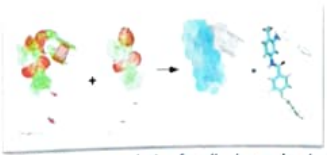

Figure 1: Our solution for aligning molecules

Figure 2 : protein cavity

Sujet 7

Le modèle de régression linéaire simple et multiple: outils de prévision

Le modèle de régression linéaire est une modélisation qui est encore aujourd'hui une des méthodes statistiques les plus utilisées afin d'Á©tablir un lein fonctionnel entre une variable réponse et des variables potentiellement explicatives. C'est la raison pour laquelle il convient de connaître cette méthode avec ses propriÁ©tés, ses interprŕations et ses difficultés. Par ailleurs, si il y a une cinquantaine d'années la question n'était pas d'actualité en raison principalement de capacités informatiques plus limitées, actuellement une question centrale liée à la régression linéaire est la problématique de la sélection de variables qui vise à identifier le plus petit groupe de variables qui permettront une modélistion tout aussi correcte, mais avec un modèle beaucoup plus simple à utiliser en pratique en raison d'un complexité plus faible.

L'objectif de ce projet va donc consister dans un premier tempsà revenir en détail sur la régression linéire et notamment les méthodes d'estimation que sont les moindres carrés et la méthode du maximum de vraisemblance. Dans un second temps, l'idée sera d'aborder la problématique du test de modèles gaussiens emboîtés afin de pouvoir aborder par la suite la question de la sélection de variables via une stratégie forward ou backward. Pour finir, nous regarderons une autre méthode de sélection de variables, la méthode Lasso, qui peut se voir comme un problème d'optimisation sous contraintes du critère des moindres carrés. Enfin, selon le temps, nous aborderons la question de la comparaison de modèles et l'importance d'un ćhanitillon de validation ou des méthodes de validation croisée.

Ainsi, au cours de ce projet, un lein sera fait entre les probabilités, les statistiques et l'optimisation notamment afin d'insister sur la connexion entre différents domaines des mathématiques.
Par ailleurs, le tout sera appliqué sur des donn^{f}es simulées et des données réelles afin d'insister sur l'interprétation des résultats. Le logiciel utilisé sera le logiciel R.

Mots-Clef:
Régression linéire, séction de variables, pénalisation, vecteurs gaussiens

Sujet 8

Le processus de Poisson comme outils dans la compréhension de la connectivité fonctionnelle au sein du cerveau

Le cerveau humain se compose de milliers de neurones qui sont tous biologiquement connectés. Cependant, cette connectivité biologique n'explique pas l'aspect fonctionnel de ce dernier. Or, la connaissance de la connectivité fonctionnelle pourrait permettre à terme, à l'image de la thérapie génique, d'apporter des solutions à des maladies neurologiques. C'est dans cette optique que différentes méthodes ont été développées afin d'identifer les neurones qui coopèrent en réponse à une stimulation donnée. Une des premières méthodologie proposée est celle des Unitary Events. Cependant, des travaux ont montré que cette méthode souffrait d'un certain nombres d'approximation dans la méthodologie. en réponse, une autre méthode à été proposés, méthode basée d'une part sur une approximation gaussienne et d'autre part sur le recours aux processus de Poisson homogène.

L'objectif de ce projet consistera dans un premier temps à comprendre l'objectif qui se cache derrière cette méthode appelée MTGAUE (Multiple Tests based on a Gaussian Approximation of Unitary events). Pour cela, vous aurez l'article originel et une version plus courte soumise à une conférence de statistiques. Il vous sera alors demander d'expliquer les grandes Á©tapes de la méthodes et les outils mis en oeuvre. Dans un second temps, l'objectif sera de comprendre les processus de Poisson, d'apprendre à les modéliser et de programmer différentes notions mises en oeuvre dans l'article MTGAUE comme la notion de coincidences entre deux processus de Poisson et suivant le temps, la mise en oeuvre de la procédure de test qui apparait dans la méthodologie proposée.

Le logiciel utilisé sera le logiciel R.