Projets d'apprentissage statistique

- Feuille de route -

Objectifs du travail

Chaque élève étudie en détail un article (au moins) de ce thème. Vous devez montrer au travers de votre rendu par écrit et de la restitution de slides (voir ci-dessous) que vous avez compris la problématique scientifique et l'apport de l'article dans ce domaine. Il vous est demandé d'expliquer et d'illustrer numériquement les aspects principaux de l'article, ce qui suppose d'implémenter les algorithmes proposés.

Au vue des difficultés rencontrées, le travail peut se focaliser sur un seul aspect de l'article, mais une vue globale et des comparaisons entre plusieurs méthodes seront récompensés.

Travail écrit

Date de rendu : **10/11/2020, 23h59**

- fournir un dépôt `git` (github, gitlab, etc.)
- un fichier au format **.pdf** de moins de 10 pages présentant le travail effectué (LaTeX obligatoire). Il s'agit d'expliquer la problématique, les solutions possibles et celles que vous avez choisies en présentant les forces et faiblesses. Une attention particulière doit être portée à l'explication des algorithmes considérés. Si besoin, on pourra utiliser certaines des données proposées sur le site http://archive.ics.uci.edu/ml/) afin d'illustrer vos travaux.
- un fichier source contenant le code Python commenté convenablement.
- un fichier (script) faisant office de démonstration sur un exemple simple, qui peut tourner rapidement (moins de 10mn : si plus donner un ordre de grandeur du temps de calcul).

Les projets donnent lieu à la préparation d'un document de soutenance sous beamer, *cf.* https://github.com/josephsalmon/OrganizationFiles) avec moins de 20 slides.

Attention : Les projets ne respectant pas l'une de ces règles recevront une pénalité de 20 %.

- Sujets proposés -

Systèmes de recommandation

Comment prédire de manière automatique les notes qu'un utilisateur donnerai à un film ou à une musique qu'il ne connaît pas encore ? Cette question est devenu primordiale avec a attiré beaucoup de recherche avec le développement d'entreprises comme Netflix (et son fameux concours à un million de dollars), Amazon, etc. Pour une introduction à ce thème on pourra consulter la présentation de Haas *et al.* "Large-Scale Matrix Factorization" http://www3.in.tum.de/scalableanalytics/presentations/gemulla.pdf.

Une base de donnée à utiliser pour ce projet est Movielens http://grouplens.org/datasets/movielens/. La taille de la base choisie sera en fonction de l'ambition de chacun (100k, 1M ou 10M).

- Recht et Ré [9] : "Parallel Stochastic Gradient Algorithms for Large-Scale Matrix Completion"
  http://www.eecs.berkeley.edu/~brecht/papers/11.Rec.Re.IPGM.pdf
  (HERMAN FANCHON)

Machine learning pour traiter des images : débruitage, super-résolution, etc.

On propose dans cette partie des applications du *machine learning* à des problématiques d'imagerie plus classique. Un ingrédient essentiel est la notion de "patchs" (petite sous-images extraites d'une image originale), en lien avec l'apprentissage d'une représentation, d'un dictionnaire. On validera les méthodes par validation croisée sur une petite base d'images classiques ou personnelles.

Concernant les problématiques de reconnaissance de visages on utilisera par exemple la base de donnée http://www.facedetection.com/facedetection/datasets.htm.

— Burger *et al.* [1, 2] : "Image denoising with multi-layer perceptrons"
http://arxiv.org/abs/1211.1544
http://arxiv.org/abs/1211.1552
(WANG RUOYU)

Réduction de la dimension

La réduction de la dimension est un outil puissant permettant aux praticiens de l'apprentissage machine de visualiser et de comprendre de grands ensembles de données à haute dimension. L'une des techniques de visualisation les plus répandues est la t-SNE, mais ses performances souffrent de la taille des données et son utilisation correcte peut être difficile.

L'UMAP [6] est une nouvelle technique qui offre un certain nombre d'avantages par rapport à la t-SNE, notamment une vitesse accrue et une meilleure préservation de la structure globale des données. Il s'agira, d'examiner la théorie qui sous-tend l'UMAP afin de mieux comprendre comment l'algorithme fonctionne, comment l'utiliser efficacement et comment ses performances se comparent à celles de la t-SNE.

Site d'intérêt : https://pair-code.github.io/understanding-umap/

— t-SNE : Laurens van der Maaten and Geoffrey Hinton "Visualizing Datausingt-SNE" [10] https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

— UMAP : L. McInnes and J. Healy : "UMAP : Uniform Manifold Approximation and Projection for Dimension Reduction" [6] https://arxiv.org/abs/1802.03426
(SAUTON Loïc)

Méthodes probabilistes pour la grande dimension

Face à la taille grandissante des données à traiter, les techniques usuelles d'algèbre linéaire ont besoin d'être re-visitées pour pouvoir être envisagées en grande dimension. Pour cela on recourt de plus en plus fréquemment à des méthodes probabilistes et statistiques : sous-échantillonnage, projections aléatoires, etc. Les enjeux les plus importants reposent sur l'amélioration des techniques de type SVD (*Singular Value Decomposition*) et k-plus proches voisins notamment. En introduction on peut regarder l'exposé donné par E. Candès à l'IHP en janvier 2014 : http://horizons.sfds.asso.fr/?page_id=649.

— Ali Rahimi and Ben Recht : "Random Features for Large-Scale Kernel Machines"[8]
https://people.eecs.berkeley.edu/~brecht/papers/07.rah.rec.nips.pdf
(LEPERCQUE CASSANDRE)

Le Lasso en grande dimension : limites et compétiteurs

On s'intéresse ici à des méthodes populaires de sélection de variable et de régression en grande dimension. On testera les limites du Lasso à la fois d'un point de vue théorique, et d'un point de vue pratique. On considérera des comparaisons avec des méthodes gloutonnes sur des simulations, type OMP.

— Zhang [11] : "Adaptive forward-backward greedy algorithm for learning sparse representations"
(code http://cran.at.r-project.org/web/packages/foba/)
http://stat.rutgers.edu/home/tzhang/papers/it11-foba.pdf **(COIFFIER OPHELIE)**

— Chen et Caramanis [3] : "Noisy and missing data regression : distribution-oblivious support recovery"
http://jmlr.org/proceedings/papers/v28/chen13d.html
(ROEMISCH RUDOLFF)

Arbres et forêts

Les forêts aléatoires (*Random Forests*) introduites par L. Breiman dans le contexte de la classification/régression est considéré comme l'une des méthodes d'apprentissage les plus efficaces. Il s'agira ici de s'approprier les concepts sur lesquels reposent la procédure (ré-échantillonnage, agrégation et randomisation), les résultats théoriques et empiriques reflétant ses propriétés/limitations et de le mettre en oeuvre sur des données simulées et réelles.

— Lakshminarayanan *et al.* 2014 : "Mondrian Forests : Efficient Online Random Forests" [4] https://arxiv.org/pdf/1406.2673.pdf **(KANDOUCI WALID)**

Peer-grading

Lorsqu'on utilise un tel système de notation par les paires (*peer grading*), la partie la plus difficile est de réduire le biais de l'étudiant vers le haut. C'est de loin la partie la plus faible des systèmes de notation par les pairs, l'une des exigences les plus évidentes pour s'améliorer. On s'intéressera à l'analyse de tels systèmes dans cette partie.

— Piech *et al.* 2013 : "Tuned Models of Peer Assessment in MOOCs"[7]
https://web.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf
voir aussi :
https://www.cfa.harvard.edu/sed/staff/Sadler/articles/Sadler%20and%20Good%20EA.pdf et
https://arxiv.org/pdf/1506.00852.pdf
(LEFORT TANGUY)

Apprentissage de dictionnaire

Le "sparse coding", c'est-à-dire la modélisation de vecteurs de données sous forme de combinaisons linéaires éparses d'éléments de base, est largement utilisé dans l'apprentissage machine, les neurosciences, le traitement du signal et les statistiques. On se penchera sur l'apprentissage de l'ensemble des éléments de base, également appelé dictionnaire, pour l'adapter à des données spécifiques, une approche qui s'est révélée très efficace pour la reconstruction et la classification des signaux dans les domaines de l'audio et du traitement des images.

— Mairal et al. "Online Dictionary Learning for Sparse Coding" [5] https://www.di.ens.fr/~fbach/mairal_icml09.pdf la librairie SPAMs pourra être utilisée : http://spams-devel.gforge.inria.fr/ **(LAKEHAL RYMA)**

Références

[1] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising with multi-layer perceptrons, part 1 : comparison with existing algorithms and with bounds. *arXiv preprint arXiv :1211.1544*, 2012. 2

[2] H. C. Burger, C. J. Schuler, and S. Harmeling. Image denoising with multi-layer perceptrons, part 2 : training trade-offs and analysis of their mechanisms. *arXiv preprint arXiv :1211.1552*, 2012. 2

[3] Y. Chen and C. Caramanis. Noisy and missing data regression : distribution-oblivious support recovery. In *ICML*, pages 383–391, 2013. 2

[4] B. Lakshminarayanan, D. M. Roy, and Y. W. Teh. Mondrian forests : Efficient online random forests. In *NIPS*, pages 3140–3148, 2014. 3

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, pages 19–60, 2010. 3

[6] L. McInnes and J. Healy. Umap : Uniform manifold approximation and projection for dimension reduction. *ArXiv*, abs/1802.03426, 2018. 2

[7] C. Piech, J. Huang, Z. Chen, C. B. Do, A. Y. Ng, and D. Koller. Tuned models of peer assessment in moocs. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 153–160, 2013. 3

[8] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *NIPS*, pages 1177–1184. Curran Associates, Inc., 2008. 2

[9] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Math. Program. Comput.*, 5(2) :201–226, 2013. 1

[10] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(Nov) :2579–2605, 2008. 2

[11] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theory*, 57(7) :4689–4708, 2011. 2