Projets d'apprentissage statistique

- Feuille de route -

Les projets sont à rendre dernier délai pour le 7 mai à 23h59 et peuvent être fait par groupe d'un ou deux élèves (un troisième passager clandestin n'est pas envisageable). Ils seront notés sur 13 points (le challenge étant lui sur 7 points) : comprenant une partie rendu et une partie soutenance. Les soutenances auront lieux la semaine du 12 mai.

Objectifs du travail

Chaque groupe choisit un thème et étudie en détail un article (au moins) de ce thème. Vous devez montrer au travers de votre rendu par écrit et de la restitution orale (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 et, le cas échéant, d'exposer les grandes lignes des démonstrations clés.

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. Pour les projets théoriques on peut axer le travail principalement sur une démonstration, mais il ne faut pas négliger totalement la partie algorithmique.

Travail écrit

Pour la partie rendu, on créera un fichier **nom1prenom1-nom2prenom2.zip** à soumettre sous ÉOLE contenant trois fichiers suivants :

- un fichier au format **.pdf** de moins de 12 pages présentant le travail effectué (LaTeX est conseillé mais non 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 commenté convenablement. Le langage de programmation est laissé à votre convenance : Python, R, C++, Java, Matlab, etc.
- un fichier (script) faisant office de démonstration sur un exemple simple, qui peut tourner rapidement en vue de montrer son exécution lors de la soutenance.

Attention : Les élèves peuvent poster leurs documents sur ÉOLE, mais ils n'ont aucune autre action possible, pas d'ouverture ou de modification de document, pas de suppression.

Travail oral

Les projets donnent lieu à une soutenances d'une durée de 20 mn comprenant

- 15 mn de présentation (sous la forme de votre choix),
- 5 mn de questions

Les niveaux de difficultés des projets sont donnés à titre indicatif de ⋆ à ⋆⋆⋆ (et pour les projets théoriques de † à † † †).

Attention : Les projets ne respectant pas l'une des règles ci-dessus recevront une pénalité de 4 points.

- Ressources en lignes utiles -

Vidéo de cours en lignes

- Sam Roweis : http://videolectures.net/mlss06tw_roweis_mlpgm/
- Andrew Ng : http://www.youtube.com/watch?v=UzxYlbK2c7E

Cours en lignes – pdf

- Francis Bach et Sylvain Arlot : http://www.di.ens.fr/~arlot/2012orsay.htm
- Sham Kakade : http://ttic.uchicago.edu/~gregory/courses/LargeScaleLearning/
- Shai Shalev-Shwartz : "Online Learning and Online Convex Optimization" http://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf [29]

- 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 que l'on peut utiliser pour les projets est Movielens http://grouplens.org/datasets/movielens/. Le taille de la base choisie sera en fonction de l'ambition des groupes (100k, 1M ou 10M).

- ⋆ Koren *et al.* [21] : "Matrix factorization techniques for recommender systems" http://www2.research.att.com/~volinsky/papers/ieeecomputer.pdf
- ⋆⋆⋆ Recht et Ré [27] : "Parallel Stochastic Gradient Algorithms for Large-Scale Matrix Completion" http://www.eecs.berkeley.edu/~brecht/papers/11.Rec.Re.IPGM.pdf
- ⋆⋆⋆ Niu *et al.* [25] : "Hogwild ! : A lock-free approach to parallelizing stochastic gradient descent" http://www.eecs.berkeley.edu/~brecht/papers/hogwildTR.pdf

Contact : Joseph Salmon, Alexandre Gramfort

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.

- ⋆ Aharon *et al.* [1] : "K-SVD : An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation" (code http://www.cs.technion.ac.il/~ronrubin/software.html) http://people.csail.mit.edu/danielzoran/
- ⋆ Zoran et Weiss [35] : "From Learning Models of Natural Image Patches to Whole Image Restoration" (code http://people.csail.mit.edu/danielzoran/) http://www.cs.huji.ac.il/~daniez/EPLLICCVCameraReady.pdf
- ⋆⋆ Levit et Nadler [23] : "Natural Image Denoising : Optimality and Inherent Bounds" http://www.wisdom.weizmann.ac.il/~levina/papers/LbDenoise-LevinNadlerCVPR11.pdf
- ⋆⋆ Guillaumin *et al.* [16] : "Is that you ? Metric Learning Approaches for Face Identification" http://www.vision.ee.ethz.ch/~mguillau/publications/Guillaumin2009iccv2.pdf
- ⋆⋆⋆ Burger *et al.* [9, 10] : "Image denoising with multi-layer perceptrons" http://arxiv.org/abs/1211.1544 http://arxiv.org/abs/1211.1552

Contact : Joseph Salmon, Alexandre Gramfort

Machine learning pour le traitement du son

Comment reconnaitre automatiquement l'instrumentation, le genre ou l'humeur dans un morceau de musique ? Comment extraire automatiquement la partition musicale ? On attaquera ces problèmes en considérant :

- des techniques avancées de régression logistique (régularisée et à noyau) dans les cas supervisés (par exemple pour la reconnaissance des instruments de musique) ;
- des techniques de factorisation en matrices non-négatives exploitant des fonctionnelles de coût adaptées (notamment Itakura Saito) pour les cas non-supervisés (transcription automatique).

⋆ Zhu et Hastie [34] : "Kernel Logistic Regression and the Import Vector Machine" http://dept.stat.lsa.umich.edu/~jizhu/pubs/Zhu-JCGS05.pdf

⋆⋆ Févotte *et al.* [15] : "Nonnegative Matrix Factorization with the Itakura-Saito Divergence : With Application to Music Analysis" http://www.unice.fr/cfevotte/publications/journals/neco09_is-nmf.pdf

Contact : Slim Essid

Machine learning pour le traitement du texte

Classification automatique de caractère et utilisation d'algorithme de partitionnement de données (*clustering*).

⋆⋆ Joachims [19] : "Text Categorization with Support Vector Machines : Learning with Many Relevant Features" http://www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

⋆⋆⋆ Hamerly et Elkan [18] : "Alternatives to the k-means algorithm that find better clusterings" http://charlotte.ucsd.edu/users/elkan/cikm02.pdf

⋆⋆⋆ Banerjee *et al.* [3] : "Clustering with Bregman Divergences" http://machinelearning.wustl.edu/mlpapers/paper_files/BanerjeeMDG05.pdf

Contact : Chloé Clavel, Alexandre Gramfort, Laurence Likforman, Joseph Salmon

Apprentissage profond, réseau de neurones, etc.

Des avancées récentes en terme d'apprentissage de représentations ont ouvert de nouvelles voies aux méthode classique de type réseaux de neurones. On s'intéresse ici à des façons d'apprendre de manière automatique des représentations, en utilisant des Machines de Boltzmann ou de l'apprentissage approfondi (*deep learning*).

⋆⋆⋆ Salakhutdinov et Hinton [28] :"An Efficient Learning Procedure for Deep Boltzmann Machines" http://www.utstat.toronto.edu/~rsalakhu/papers/neco_DBM.pdf

⋆⋆⋆ Bengio *et al.* [4] "Representation Learning : A Review and New Perspectives" http://arxiv.org/pdf/1206.5538v2.pdf

Contact : Alexandre Gramfort, Joseph Salmon

Méthodes probabilistes pour la grande dimension

Fâce à la taille grandissante des données à traiter, les techniques usuelles d'algèbre linéaire ont besoin d'être re-visitées pour pouvoir être envvisagées en grande dimension. Pour cela on recourt de plus en plus fréquement à des méthodes probabilistes et statistiques : sous-échantillonage, 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.

⋆ Halko *et al.* : "Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions" [17] http://arxiv.org/abs/0909.4061

⋆ Bingham et Mannila [7] : "Random projection in dimensionality reduction : Applications to image and text data" http://www.cs.montana.edu/~gradl/randproj_kdd.pdf

† † † Shalev-Schartz et Zhang [31] : "Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization" http://www.optimization-online.org/DB_FILE/2013/09/4038.pdf

Contact : Anne Sabourin, Alexandre Gramfort, Joseph Salmon

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. Côté simulation on s'intéressera aux cas dans lesquel les corrélations des variables explicatives (ou *features*) aident/pénalisent la qualité de la sélection/prédiction [13]. On considérera des comparaisons avec des méthodes gloutonnes sur des simulations (*e.g.,* [22])

† † † Dalalyan *et al.* [13] : "On the Prediction Performance of the Lasso"
http://arxiv.org/abs/1402.1700

† † † Zhang [33] : "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

† † Xu *et al.* [32] : " Robust Regression and Lasso"
http://arxiv.org/pdf/0811.1790v1.pdf

⋆ ⋆ ⋆ Chen et Caramanis [11] : "Noisy and missing data regression : distribution-oblivious support recovery"
http://jmlr.org/proceedings/papers/v28/chen13d.html

Contact : Anne Sabourin, Alexandre Gramfort, Joseph Salmon

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.

⋆ Breiman [8] : "Random forests"
http://oz.berkeley.edu/~breiman/randomforest2001.pdf

† † Biau *et al.* [5] : "Consistency of Random Forests and Other Averaging Classifiers "
machinelearning.wustl.edu/mlpapers/paper_files/biau08a.pdf

Contact : Anne Sabourin, Alexandre Gramfort, Joseph Salmon

SVM

On étudie dans cette partie le traitement de SVM ou de machine à noyau, et différentes manières de résoudre les problèmes d'otpimisation associés. Comme solutions potentielles on regardera des méthodes basées sur la descente de (sous-)gradient stochastique, ainsi que sur des plans de coupes par exemple.

⋆ Joachims *et al.* [20] : "Cutting-plane training of structural SVMs" http://www.cs.cornell.edu/people/tj/publications/joachims_etal_09a.pdf

⋆ ⋆ ⋆ Shalev-Shwartz *et al.* [30] : "Pegasos : Primal Estimated sub-GrAdient SOlver for SVM"
http://www.cs.huji.ac.il/~shais/papers/ShalevSiSrCo10.pdf

Contact : Alexandre Gramfort, Joseph Salmon

Méthodes d'agrégation et méthodes gloutonnes

Le Lasso est l'une des méthodes les plus populaires utilisées pour la prédiction et la sélection de variables. Pour autant cette technique bien que comprise d'un point de vue pratique et théorique n'est pas toujours optimale. On considérera ici comme alternatives des méthodes de types gloutonnes (*greedy*) ou à poids exponentielle, que l'on comparera avec la méthode Lasso [6].

† † † Dai *et al.* [12] : "Deviation optimal learning using greedy Q-aggregation"
https://www.princeton.edu/~rigollet/PDFs/DaiRigZha12.pdf

⋆ Dalalyan et Tysbavkov [14] : "Sparse Regression Learning by Aggregation and Langevin Monte-Carlo"
http://arxiv.org/pdf/0903.1223v3.pdf

† † Audibert [2] : "Progressive mixture rules are deviation suboptimal"
http://certis.enpc.fr/~audibert/Mes%20articles/NIPS07a.pdf

Contact : Anne Sabourin, Joseph Salmon

Calibration de probabilités en classification binaire

Une méthode de classification binaire fournit une prédiction 0 ou 1 pour une nouvelle observation. Or en pratique il est souvent pertinent de connaître la probabilité que la réponse soit 1. Cela permet par exemple d'avoir un degré de confiance sur la prédiction. Deux méthodes classiques existent : la méthode de Platt qui fournit une sortie probabiliste au SVM binaire, ainsi que la méthode non-paramétrique basée sur une régression isotonique. Vous reproduirez certaines figures des papiers et évaluerez les performances pour différents classifieurs binaires (SVM, Logistique, Naive Bayes, etc.) dans un contexte réel et de simulation.

†† Niculescu *et al.* [24] : "Predicting Good Probabilities With Supervised Learning"
http://machinelearning.wustl.edu/mlpapers/paper_files/icml2005_Niculescu-MizilC05.pdf

⋆⋆ Platt [26] : "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods"
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.41.1639&rep=rep1&type=pdf

Contact : Alexandre Gramfort, Joseph Salmon

Références

[1] M. Aharon, M. Elad, and A. Bruckstein. K-SVD : An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11) :4311–4322, 2006. 2

[2] J-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *NIPS*, pages 41–48, 2007. 4

[3] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *J. Mach. Learn. Res.*, 6 :1705–1749, 2005. 3

[4] Y. Bengio, A. Courville, and P. Vincent. Representation learning : A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8) :1798–1828, Aug 2013. 3

[5] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9 :2015–2033, 2008. 4

[6] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4) :1705–1732, 2009. 4

[7] E. Bingham and H. Mannila. Random projection in dimensionality reduction : applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001. 3

[8] L. Breiman. Random Forests. *Mach. Learn.*, 45(1) :5–32, 2001. 4

[9] 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. http://arxiv.org/abs/1211.1544. 2

[10] 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. http://arxiv.org/abs/1211.1552. 2

[11] Y. Chen and C. Caramanis. Noisy and missing data regression : Distribution-oblivious support recovery. In *ICML*, pages 383–391, 2013. 4

[12] D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Statist.*, 40(3) :1878–1905, 2012. http://arxiv.org/abs/1203.2507. 4

[13] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *arXiv preprint arXiv :1402.1700*, 2014. 4

[14] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5) :1423–1443, 2012. 4

[15] C. Févotte, N. Bertin, and J-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis. *Neural Computation*, 21(3), Mar 2009. 3

[16] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you ? metric learning approaches for face identification. In *ICCV*, pages 498–505. IEEE, 2009. 2

[17] N. Halko, P. Martinsson, and J. A. Tropp. Finding structure with randomness : Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 :217, 2011. 3

[18] G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In *ICIKM*, pages 600–607. ACM, 2002. 3

[19] T. Joachims. Text categorization with support vector machines : Learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Machine Learning : ECML-98*, volume 1398 of *Lecture Notes in Computer Science*, pages 137–142. Springer Berlin Heidelberg, 1998. 3

[20] T. Joachims, T. Finley, and C.-N. J. Yu. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1) :27–59, 2009. 4

[21] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8) :30–37, Aug 2009. 2

[22] J. Lederer. Trust, but verify : benefits and pitfalls of least-squares refitting in high dimensions. *arXiv preprint arXiv :1306.0113*, 2013. 4

[23] A. Levin and B. Nadler. Natural image denoising : Optimality and inherent bounds. In *CVPR*, 2011. http://www.wisdom.weizmann.ac.il/~levina/papers/LbDenoise-LevinNadlerCVPR11.pdf. 2

[24] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *ICML*, pages 625–632, New York, NY, USA, 2005. ACM. 5

[25] F. Niu, B. Recht, C. Ré, and S. J. Wright. Hogwild ! : a lock-free approach to parallelizing stochastic gradient descent. *NIPS*, 24 :693–701, 2011. 2

[26] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*, pages 61–74. MIT Press, 1999. 5

[27] B. Recht and C. Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2) :201–226, 2013. 2

[28] R. Salakhutdinov and G. Hinton. An efficient learning procedure for deep boltzmann machines. *Neural computation*, 24(8) :1967–2006, 2012. 3

[29] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2) :107–194, 2011. http://www.cs.huji.ac.il/~shais/papers/OLsurvey.pdf. 2

[30] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos : Primal Estimated sub-GrAdient SOlver for SVM. *Mathematical programming*, 127(1) :3–30, 2011. 4

[31] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *arXiv preprint arXiv :1309.2375*, 2013. 3

[32] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. *IEEE Trans. Inf. Theory*, 56(7) :3561–3574, 2010. 4

[33] T. Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theory*, 57(7) :4689–4708, 2011. 4

[34] J. Zhu and T. Hastie. Support vector machines, kernel logistic regression and boosting. *Multiple Classifier Systems*, 2364 :16–26, 2002. 3

[35] D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, pages 479–486. IEEE, 2011. 2