Sujets de TER

Gilles Cohen
gilles.cohen@univ-lyon1.fr

1 Réduction de dimension

La réduction de la dimension est un outil puissant qui permet la visualisation, l'analyse et la compréhension de grands ensembles de données à dimension très élevées. Il consiste à prendre des données dans un espace de grande dimension, et à les remplacer par d'autres dans un espace de dimension inférieure, mais qui contiennent encore la plupart des informations contenues dans le grand ensemble. Autrement dit, on cherche à construire moins de variables tout en conservant le maximum d'informations possible.

En Machine Learning, ce processus de traitement de données est crucial dans certains cas, parce que les jeux de données plus petits sont plus faciles à explorer, exploiter et à visualiser, et rendent l'analyse des données beaucoup plus facile et plus rapide.

Cette étape est importante aussi dans les cas du sur-apprentissage et des données très éparses (fléau de la dimensionnalité), qui nécessitent beaucoup de temps et de puissance de calcul pour les étudier.

En utilisant un espace de plus petite dimension, on obtient des algorithmes plus efficaces, ainsi qu'un ensemble de solutions plus réduit.

Parmi les différentes techniques de réduction de dimensions on s'interressera aux techniques non linéaire à apprentissage de variétés (manifold). L'une de ces techniques de réduction de dimension les plus répandues est t-SNE [1], mais ses performances souffrent de la taille des données et son utilisation correcte peut être difficile. L'UMAP [2] est une nouvelle technique qui offre un certain nombre d'avantages par rapport à la t-SNE, en particulier une vitesse améliorée et une meilleure conservation de la structure globale des données.

Il s'agira, d'examiner la théorie qui sous-tend l'UMAP afin de mieux comprendre le fonctionnement de l'algorithme, comment l'utiliser efficacement et comment ses performances se comparent à celles de t-SNE. Une comparaison pratique sur des données réelles de grandes dimension est demandée.

References

[1] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: `http://www.jmlr.org/papers/v9/vandermaaten08a.html`.

[2] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: (2018). URL: http://arxiv.org/abs/1802.03426.

2 Interprétabilité/Explicabilité en apprentissage automatique

La dernière décennie a vu une augmentation importante de l'utilisation de systèmes de décision construits sur des techniques modernes d'IA souvent opaques, comme l'apprentissage profond. Ces systèmes de type "boîte noire", basés sur de grandes quantités de données, constituent un outil clé pour la prise de décisions importantes tant pour les individus que pour les entreprises et la société en général.

Il est primordial que les utilisateurs de ces systèmes puissent évaluer ces décisions et leur faire confiance. Cet aspect de l'apprentissage automatique, encore peu étudié, mais important pour l'acceptabilité, est celui de l'interprétabilité [1]. Le domaine de l'intelligence artificielle explicable (XAI) aborde cette question, afin de permettre aux utilisateurs de mieux comprendre le comportement des systèmes IA complexes. Ces systèmes doivent être interprétables, en ce sens que l'on puisse comprendre sur quels éléments la décision s'appuie et donc pouvoir rendre des comptes. L'interprétabilité est un élément clé pour éviter que les systèmes artificiellement intelligents prennent des décisions non-transparentes, voire injustifiables, et ce quelque soient leurs performances.

L'objet du présent travail est de réaliser une évaluation comparative entre les trois méthodes XAI les plus populaires, à savoir LIME [3], SHAP [2] et CAM (Grad-CAM) [4] , où l'on mesure l'impact des régions sur la prédiction du modèle. Une étude théorique et pratique de ces trois méthodes est à réaliser en soulignant leurs différences. La partie pratique considérera un problème de classification traité à l'aide d'une approche à réseaux de neurones profonds (CNN, LSTM. ...).

References

[1] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. “Explainable AI: A Review of Machine Learning Interpretability Methods”. In: *Entropy* 23.1 (2021). URL: https://www.mdpi.com/1099-4300/23/1/18.

[2] Scott Lundberg and Su-In Lee. “A Unified Approach to Interpreting Model Predictions”. In: (2017). URL: https://arxiv.org/abs/1705.07874.

[3] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “”Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[4] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” In: *ICCV*. IEEE Computer Society, 2017, pp. 618–626. ISBN: 978-1-5386-1032-9. URL: http://dblp.uni-trier.de/db/conf/iccv/iccv2017.html#SelvarajuCDVPB17.

3 Inférence sélective

L'inférence sélective est un paradigme général permettant d'aborder les problèmes qui se posent lorsque les hypothèses statistiques ne sont pas spécifiées avant la collecte des données, mais définies au cours du processus d'analyse des données. Dans ce paradigme les analystes de données sélectionnent les modèles après avoir vu les données, et non l'inverse.

Une stratégie courante en analyse de données consiste à supposer qu'une classe de modèles M spécifiée à priori contient un modèle particulier $M \in \mathcal{M}$ qui est bien adapté aux données. Malheureusement, lorsque les données ont été utilisées pour choisir un modèle particulier $\hat{M}$, l'inférence sur le modèle sélectionné est généralement invalide, les méthodes d'inférence classiques ne peuvent pas nous fournir leurs garanties habituelles. Par exemple, dans le cadre de la régression avec de nombreux prédicteurs, si les "meilleurs" prédicteurs sont choisis par lasso ou forward stepwise, alors les tests classiques t, χ^2 , ou F pour leurs coefficients seront anti-conservateurs. L'inférence sélective vise à développer de nouvelles méthodes d'inférence qui sont compatibles avec ce nouveau paradigme.

La solution la plus répandue est le fractionnement des données, où un sous-ensemble des données est utilisé uniquement pour la sélection du modèle et un autre sous-ensemble est utilisé uniquement pour l'inférence. Cela entraîne une perte de précision pour la sélection du modèle et de puissance pour l'inférence. Récemment, Lockhart et al. [3] ; Lee et al. [1] ont développé des méthodes d'ajustement des tests de signification classiques pour tenir compte de la sélection de modèles.

Dans ce travail, on vous demandera d'étudier principalement les tests d'inférence sélective et les intervalles de confiance proposés par [4]. D'évaluer les performances en termes de puissance des tests ainsi que la probabilité de couverture des intervalles de confiance associés. De plus, [2] soulignent que les procédures susmentionnées présentent un problème de "sur-conditionnement" qui conduit souvent à des intervalles de confiance très larges ; par conséquent, l'impact du conditionnement sera également à étudier par une étude de simulation et l'importance des hypothèses de normalité qui sous-tendent le test sera également évaluée.

Vous devrez faire également une revue de la littérature sur la régression linéaire, les méthodes de sélection de modèles, notamment Forward Stepwise Selection, LASSO et LARS, ainsi que l'algorithme d'inférence sélective proposé par Tibshirani et al [4].

References

[1] Jason D. Lee et al. "Exact post-selection inference, with application to the lasso". In: *The Annals of Statistics* 44.3 (June 2016). DOI: 10.1214/15-aos1371. URL: https://doi.org/10.1214%2F15-aos1371.

[2] Keli Liu, Jelena Markovic, and Robert Tibshirani. "More powerful post-selection inference, with application to the Lasso". In: (2018). URL: https://arxiv.org/abs/1801.09037.

[3] Richard Lockhart et al. "A significance test for the lasso". In: *The Annals of Statistics* 42.2 (Apr. 2014). DOI: 10.1214/13-aos1175. URL: https://doi.org/10.1214%2F13-aos1175.

[4] Ryan J. Tibshirani et al. "Exact Post-Selection Inference for Sequential Regression Procedures". In: *Journal of the American Statistical Association* 111.514 (2016), pp. 600–620. URL: https://doi.org/10.1080/01621459.2015.1108848.

4 Extreme Gradient Boosting (XGBoost) et LightGBM

L'arbre de décision par boosting de gradient (GBDT) [2, 4] est un algorithme d'apprentissage automatique populaire, et il existe un certain nombre d'implémentations efficaces dont XGBoost [1]. Bien que de nombreuses optimisations techniques aient été adoptées dans ces implémentations, l'efficacité et l'évolutivité restent insatisfaisantes lorsque la dimension des caractéristiques est élevée et que la taille des données est importante. L'une des principales raisons est que, pour chaque caractéristique, il est nécessaire d'analyser toutes les instances de données pour estimer le gain d'information de tous les points de séparation possibles, ce qui prend beaucoup de temps. Pour résoudre ce problème, *LightGBM* [3] a été proposé. Cet algorithme incorpore deux nouvelles techniques : *Gradient-based One-Side Sampling* (GOSS) et *Exclusive Feature Bundling* (EFB). Avec la première GOSS, une proportion significative d'instances de données avec de petits gradients est exclue, et le reste est utilisé pour estimer le gain d'information. Puisque les instances de données avec de plus grands gradients jouent un rôle plus important dans le calcul du gain d'information, GOSS peut obtenir une estimation assez précise du gain d'information avec une taille de données beaucoup plus petite. Avec la seconde technique EFB, les caractéristiques mutuellement exclusives (c'est-à-dire qu'elles prennent rarement des valeurs non nulles simultanément) sont regroupées, afin de réduire le nombre de caractéristiques. Un algorithme glouton peut atteindre un assez bon taux d'approximation (et peut donc réduire efficacement le nombre de caractéristiques sans nuire beaucoup à la précision de la détermination du point de séparation). Il est montré que LightGBM accélère le processus d'apprentissage du GBDT conventionnel jusqu'à plus de 20 fois tout en obtenant presque la même précision.

Dans ce travail on demande une analyse pratique du fonctionnement de cette nouvelle technique en termes de vitesse d'apprentissage, de performance de généralisation et de configuration des paramètres. De plus, une comparaison exhaustive entre LightGBM et XGBoost est demandée en utilisant des modèles soigneusement réglés ainsi que les paramètres par défaut. Enfin, une analyse approfondie du processus de réglage des paramètres des deux algorithmes devra être effectuée.

References

[1] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System." In: *KDD*. Ed. by Balaji Krishnapuram et al. ACM, 2016, pp. 785–794. URL: http://dblp.uni-trier.de/db/conf/kdd/kdd2016.html#ChenG16.

[2] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[3] Guolin Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems 30 (NIP 2017)*. Dec. 2017. URL: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/.

[4] Si Si et al. "Gradient Boosted Decision Trees for High Dimensional Sparse Output". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, June 2017, pp. 3182–3190. URL: https://proceedings.mlr.press/v70/si17a.html.

5 Clustering par apprentissage profond

Le clustering non supervisé a été largement étudié dans la communauté de l'exploration de données et de l'apprentissage automatique. Les algorithmes traditionnels de clustering, tels que k-means [4], Spectral Clustering (SC) [5] et Gaussian Mixture Model (GMM) [1] partitionnent les données en groupes distincts avec des caractéristiques construites à la main selon la similarité intrinsèque. Cependant, ces caractéristiques sont conçues pour un usage général et ne sont donc pas tout à fait adaptées à une tâche spécifique. Avec le développement de l'apprentissage profond, les réseaux neuronaux profonds (DNN) peuvent apprendre de bonnes caractéristiques pour la classification. Récemment, les algorithmes de clustering profond [8, 6, 2, 3, 9, 7] adoptent les DNN pour effectuer le clustering, montrant une amélioration spectaculaire des performances de clustering. L'idée de base est que de bonnes caractéristiques aident à produire de bons résultats de clustering, et que ce dernier guide à son tour le DNN pour apprendre de meilleures caractéristiques. Les deux processus sont exécutés

de manière itérative pour obtenir des performances supérieures. Au final, les caractéristiques apprises sont spécifiques à la tâche, ce qui est beaucoup plus approprié pour le clustering.

Le travail demandé consiste à en premier lieu à établir un état de l'art du clustering profond puis de faire une étude comparative théorique et pratique entre les méthodes suivantes [8, 3, 2] en soulignant les avantages/inconvénients de chacune des méthodes. Une seconde étape sera, après avoir retenu la meilleure approche, selon vous de clustering profond parmi les trois méthodes étudiées la comparer à une méthode traditionnelle comme par exemple k-mean.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] Kamran Ghasedi Dizaji et al. "Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization". In: (2017). URL: https://arxiv.org/abs/1704.06327.

[3] Xifeng Guo et al. "Improved Deep Embedded Clustering with Local Structure Preservation". In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. IJCAI'17. Melbourne, Australia: AAAI Press, 2017, pp. 1753–1759.

[4] J MacQueen. "Classification and analysis of multivariate observations". In: *5th Berkeley Symp. Math. Statist. Probability*. 1967, pp. 281–297.

[5] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. "On Spectral Clustering: Analysis and an algorithm". In: *ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*. MIT Press, 2001, pp. 849–856. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.8100.

[6] Xi Peng et al. "Cascade Subspace Clustering". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. AAAI'17. San Francisco, California, USA: AAAI Press, 2017, pp. 2478–2484.

[7] Yazhou Ren et al. "Deep Clustering: A Comprehensive Survey". In: (2022). URL: https://arxiv.org/abs/2210.04142.

[8] Junyuan Xie, Ross Girshick, and Ali Farhadi. "Unsupervised Deep Embedding for Clustering Analysis". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 478–487.

[9] Bo Yang et al. "Towards K-means-friendly Spaces: Simultaneous Deep Learning and Clustering". In: (2016). URL: https://arxiv.org/abs/1610.04794.