Un GAN différentiellement confidentiel pour générer des données démographiques du Maryland

Pascal Jutras-Dubé

1 Introduction

Durant les dernières décennies, avec le progrès continu de l'informatique mobile et la popularité croissante des médias sociaux, une industrie vaste et opaque accumule des quantités massives de données personnelles riches de sémantique. Ceci arrive tous les jours, souvent sans que l'utilisateur dont les données sont collectées ne le saches [Hitlin and Rainie, 2019]. Bien que l'analyse et la compréhension de ces données aient une valeur commerciale considérable (par exemple, les publicités ciblées et les recommandations personnalisées), la confidentialité devrait être prise en compte lors du partage de ces données. Dans le domaine médical, les données individuelles issues d'études cliniques peuvent contenir des informations sensibles; un individu pourrait être affecté si ses données médicales étaient partagées [Beaulieu-Jones et al., 2019]. Par exemple, un assureur qui accède à une banque de données médicales pourrait ajuster les primes d'assurance de ses clients qui y figurent. La nature sensible des données oblige souvent les scientifiques à signer des accords d'utilisation ou à établir des collaborations formelles. Ces exigences ralentissent, voire empêchent le partage des données entre les chercheurs. Les risques des données associés à la vie privée et à la sensibilité posent un défi important, celui de concevoir des modèles d'analyses de haute-qualité, mais qui satisfont aux besoins de confidentialité.

On peut se poser la question suivante: est-ce que partager la distribution de probabilité des données à la place des données sensibles originales garantit la confidentialité? Idéalement, avec la distribution générative en main, il serait possible d'échantillonner une population synthétique qui respecte les propriétés statistiques de la population véritable. Le cas échéant, il suffirait d'étudier la population synthétique de substitution pour apprendre des informations utiles sur la population originale sans rien apprendre sur les individus en isolation. Les réseaux antagonistes génératifs (GANs) [Goodfellow et al., 2014] et ses variantes ont démontré des performances impressionnantes dans la modélisation de la distribution des données sous-jacentes en combinant la complexité des réseaux de neurones profonds et la théorie des jeux pour générer des échantillons synthétiques qui sont difficilement différentiables des échantillons réels. Toutefois, en raison de la capacité élevée de mémorisation des réseaux de neurones profonds, la densité apprise peut se concentrer sur les données d'apprentissage, ce qui signifie que les GANs peuvent se souvenir des données d'entraînement. Il y a une chance considérable de retrouver les échantillons d'apprentissage en échantillonnant à répétition [Arjovsky et al., 2017].

La confidentialité différentielle (DP) [Dwork and Roth, 2014] adresse le paradoxe d'apprendre de l'information utile sur une population sans rien apprendre sur un individu. C'est une définition mathématique qui garantit que toute séquence de résultats (réponses aux requêtes faites à une base de donnée) est essentiellement également susceptible de se produire, indépendamment de la présence ou l'absence de tout individu. Xu et al. [2019] et Xie et al. [2018] présentent des architectures de WGAN [Arjovsky et al., 2017] qui respectent la confidentialité différentielle en ajoutant du bruit au gradient durant l'entraînement.

2 Expériences

En guise de projet, j'aimerais répliquer le modèle proposé par Xie et al. [2018], mais l'évaluer sur des données démographiques du Maryland provenant du recensement américain.

2.1 Données

Les *pumas (Public Use Microdata Area)*, sont des unités géographiques utilisées par le recensement américain pour fournir des informations statistiques et démographiques. Dans la base de données avec laquelle je vais travailler, chaque exemple contient 9 champs caractéristiques: *NP: Number of household people, HHT: Household or family type, HINCP: Household income, HUPAC: Household presence and age of children, WIF: Workers in family during the last 12 months, AGEP: Age of the person, SEX: Gender of the person, ESR: Employment status of the person* et *RAC1P: Recorded detailed race* ainsi que deux champs supplémentaires de géolocalisation: *PUMA: le numéro du puma* et *ST: le numéro de l'état*. Chacune de ces caractéristiques prend des valeurs numériques entières. Je travaillerai plus particulièrement avec 4 pumas de la région d'Anne Arundel. Ces pumas ont entre 3979 et 5778 exemples chaque. Puisque les données sont déjà anonymisées, l'exercice n'est qu'académique, mais la pertinence demeure puisqu'on peut quand même étudier la crédibilité des données générées par le GAN différentiellement confidentiel. Remarquons aussi que la taille de l'ensemble des données est petite comme c'est souvent le cas des bases de données médicales. Dans de telles situations, le partage de données synthétiques permet non seulement de préserver la confidentialité, mais aussi d'atténuer le problème de rareté des données.

2.2 Évaluation

Le premier défi est celui de parvenir à générer des données synthétiques avec un WGAN entraîné sans confidentialité différentielle. Ce premier générateur pourra être considéré comme une borne supérieure sur la qualité des échantillons qu'on peut espérée avoir lorsqu'on entraîne avec confidentialité différentielle. Ensuite, j'aimerais ajouter la randomisation qui garantit la confidentialité et étudier le compromis entre l'utilité et la confidentialité.

Un autre défi considérable est celui de l'évaluation de la performance des générateurs et en particulier de quantifier la relation entre le niveau de confidentialité différentielle et l'utilité des échantillons. Comme les données ne sont pas des images, on peut difficilement voir qualitativement si les échantillons sont vraisemblables. Aussi, les données ne sont pas labellées de sorte qu'on ne peut pas vraiment utiliser le *inception score*. Toutefois, pour évaluer dans quelle mesure les modèles génératifs récupèrent la relation entre les dimensions des données, on peut utiliser DWPre (dimension-wise prediction) [Choi et al., 2017]. L'idée est d'entraîner un classifieur à prédire une dimension des données à partir des autres. Si les données synthétiques sont réalistes, on s'attend à ce que la performance d'un classifieur entraîné sur les données synthétiques soit comparable à lorsqu'il est entraîné sur les vraies données.

En somme, l'objectif de l'exercice est de voir si les résultats de Xu et al. [2019] et Xie et al. [2018] sont transférables à un autre dataset et d'étudier comment évaluer le rapport entre la confidentialité et la qualité de la population synthétique pour des données qui ne sont pas des images.

Pour entraîner les modèles, j'aurai accès à Google Colab, voire peut-être Google Colab Pro.

References

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 214–223, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL `http://proceedings.mlr.press/v70/arjovsky17a.html`.

K. B. Beaulieu-Jones, Z. Steven Wu, C. Williams, R. Lee, S. P. Bhavnani, J. Brian Byrd, and C. S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *ahajournals*, 2019.

E. Choi, S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1703.06490, Mar. 2017.

C. Dwork and A. Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. 2014.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1406.2661, June 2014.

P. Hitlin and L. Rainie. Facebook algorithms and personal data. *Pew Research Center*, 2019.

L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially Private Generative Adversarial Network. *arXiv e-prints*, art. arXiv:1802.06739, Feb. 2018.

C. Xu, J. Ren, D. Zhang, Y. Zhang, Z. Qin, and K. Ren. Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security*, PP:1–1, 02 2019. doi: 10.1109/TIFS.2019.2897874.