یا ذالامن و الامان

آزمایشگاه امنیت داده و شبکه

# حریم خصوصی در یادگیری ماشین


ارائه‌دهنده

مجید ذوالفقاری


فروردین ۱۳۹۹

# فهرست مطالب

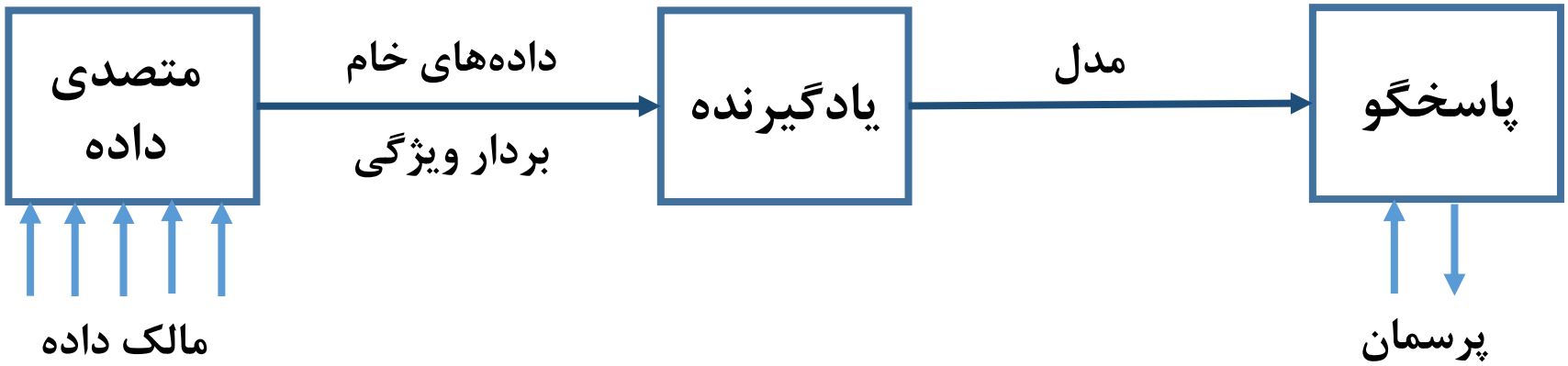
# مقدمه

- امروزه ارزشمندترین منبع دنیا، داده است. [1]
- از این‌رو، حفاظت از این منابع ارزشمند، به یکی از مهم‌ترین دغدغه‌ها تبدیل شده و قوانین و استانداردهای مختلفی به منظور حفظ حریم خصوصی داده‌ها وضع شده است.
- بسیاری از داده‌ها به نحوی با الگوریتم‌های یادگیری ماشین سروکار دارند.
- به عنوان مثال:
  - استفاده از داده‌های مکانی به منظور یادگیری مسیرهای بهینه
  - استفاده از داده‌های ژنتیکی افراد در بیوانفورماتیک
  - استفاده از داده‌های شبکه‌های اجتماعی و تجارت الکترونیکی در توصیه‌گرها
  - استفاده از داده‌های سلامتی در سامانه‌های تشخیص بیماری
  - استفاده از داده‌های بیولوژیکی برای احراز اصالت
- این مساله، اهمیت توجه به حریم خصوصی در یادگیری ماشین را نشان می‌دهد.

[1] https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

# حریم خصوصی در یادگیری ماشین

- در هر عملیات یادگیری ماشین، سه نقش متصور است:
  - متصدی داده
  - یادگیرنده (انجام پردازش بر روی داده‌ها و تولید مدل)
  - پاسخگو (پاسخگویی به پرسمان‌ها)
- در صورت توزیع این نقش‌ها بین چند طرف، نیاز به اعمال فناوری‌هایی به منظور حفظ حریم خصوصی



# انواع حملات

- **مهم‌ترین انواع حملات مرتبط با حریم خصوصی یادگیری ماشین:**
  - حملات بازسازی داده‌ها
    - به دست آوردن داده‌های خام با داشتن بردار ویژگی‌ها
  - حملات معکوس‌سازی مدل
    - به دست آوردن بردار ویژگی‌ها با داشتن مدل
  - حملات استنتاج عضویت
    - استنتاج اینکه آیا داده مورد نظر جزو داده‌های یادگیری مدل بوده یا خیر.

# حملات بازسازی داده‌ها

- هدف مهاجم، بازسازی داده‌های خام با استفاده از بردار ویژگی‌هاست.
- لزوم دسترسی مهاجم به بردار ویژگی‌ها
  - کنجکاوی کارپذیر پردازش
  - دسترسی جعبه سفید مهاجم به مدل
    - ذخیره‌سازی بردار ویژگی‌ها در خود مدل در برخی الگوریتم‌ها مانند SVM یا kNN.
- به عنوان مثال:
  - بازسازی تصویر اثرانگشت با استفاده از الگوی نقاط ویژه اثرانگشت (فِنگ و جِین ۲۰۱۱)
  - بازسازی مجموعه‌داده یک بخش‌کننده مبتنی بر درخت تصمیم با استفاده از ویژگی‌های داده‌ها (گمبز و همکاران ۲۰۱۲)
  - بازسازی نحوه لمس موبایل توسط کاربر با استفاده از ویژگی‌های لمس (مانند سرعت و جهت) (الرباعی و چنگ ۲۰۱۶)

# حملات معکوس‌سازی مدل

- هدف مهاجم، یافتن داده‌ها یا بردار ویژگی‌هایی است که مدل با استفاده از آن‌ها ایجاد شده است.
- استفاده از میزان اطمینان مدل به پاسخ
- نحوه دسترسی مهاجم:
  - جعبه سفید (دسترسی به پارامترهای مدل)
  - جعبه سیاه (امکان ارسال پرسمان به مدل و گرفتن پاسخ)
- به عنوان مثال:
  - بازیابی تصاویر قابل تشخیص از چهره افراد با داشتن نام آن‌ها و دسترسی جعبه سیاه به مدل (فردریکسون و همکاران ۲۰۱۴)
  - بازیابی داده‌های تصویری آموزش در شبکه عصبی عمیق با استفاده از GAN (ژانگ و همکاران ۲۰۱۹)
  - بازیابی داده‌های تصویری آموزش با داشتن مدل در یادگیری عمیق همکارانه (هه و همکاران ۲۰۱۹)

# حملات استنتاج عضویت

- مهاجم با داشتن یک مدل و یک نمونه، تعیین می‌کند که آیا آن نمونه عضوی از مجموعه‌داده آموزش استفاده شده برای ساخت آن مدل بوده است یا خیر.
- تبدیل مساله استنتاج به مساله بخش‌بندی
  - استفاده از تفاوت پاسخ مدل برای نمونه‌هایی که داخل مجموعه‌داده آموزش هستند و نیستند.
- نحوه دسترسی مهاجم:
  - جعبه سفید (دسترسی به پارامترهای مدل)
  - جعبه سیاه (امکان ارسال پرسمان به مدل و گرفتن پاسخ)
- به عنوان مثال:
  - اجرای حمله استنتاج عضویت بر روی شبکه عصبی (شکری و همکاران ۲۰۱۷)
  - اجرای حمله استنتاج عضویت بر روی مدل یادگیری عمیق (رحمان و همکاران ۲۰۱۸)
  - بررسی میزان تاثیر ویژگی های داده و مدل بر روی میزان آسیب پذیری نسبت به حملات استنتاج عضویت (مه‌جبین و همکاران ۲۰۲۰)

# راهکارهای حفظ حریم خصوصی در یادگیری ماشین

- به طور کلی، راهکارهای حفظ حریم خصوصی در یادگیری ماشین را می‌توان به دو دسته تقسیم کرد:

1) راهکارهای مبتنی بر رمزنگاری
   - استفاده از رمزنگاری همریخت
     - چالش اصلی: بالا بودن درجه توابع فعال‌سازی رایج
     - برخی از راهکارهای ارائه‌شده برای ارائه توابع جایگزین:
       - شی و همکاران (۲۰۱۴)، داولین و همکاران (۲۰۱۶)، حسامی‌فرد و همکاران (۲۰۱۷)، شو و همکاران (۲۰۱۹)

2) راهکارهای مبتنی بر ایجاد آشفتگی
   - استفاده از حریم خصوصی تفاضلی
     - چالش‌های اصلی: دقت و کارآیی

# حریم خصوصی تفاضلی

- ارائه شده توسط دیورک و همکاران (۲۰۰۶)
- تعریف حریم خصوصی تفاضلی نسبی:
  - یک سازوکار تصادفی M را حافظ حریم خصوصی تفاضلی با پارامترهای (ε,σ) می‌گوییم، اگر به ازای هر مجموعه خروجی Ω و مجموعه‌داده‌های D و D′ رابطه زیر برقرار باشد:

$$Pr[\mathcal{M}(D) \in \Omega] \leq \exp(\epsilon) \cdot Pr[\mathcal{M}(D') \in \Omega] + \delta.$$

- تعریف حریم خصوصی تفاضلی محض:
  - در رابطه بالا اگر مقدار σ برابر صفر باشد، سازوکار تصادفی M را حافظ حریم خصوصی تفاضلی با پارامتر ε می‌نامیم.

# حریم خصوصی تفاضلی - ادامه

- تعریف حساسیت:
  - حساسیت یک تابع نشان‌دهنده بیشترین تفاوت بین خروجی های آن تابع بر روی مجموعه‌داده‌های همسایه است.
- حساسیت تابع f به صورت زیر تعریف می‌شود:

$$\Delta f = \max_{D,D'} ||f(D) - f(D')||_1$$

- رایج‌ترین سازوکار برای حفظ حریم خصوصی تفاضلی، سازوکار لاپلاس (ارائه شده توسط دیورک و همکاران ۲۰۰۶) است:

$$\mathcal{M}(D) = f(D) + Lap\left(\frac{\Delta f}{\epsilon}\right)$$

- برای توابع پیوسته نیز می‌توان از سازوکار نمایی (ارائه‌شده توسط مک‌شری و همکاران ۲۰۰۷) استفاده کرد.

# استفاده از حریم خصوصی تفاضلی در یادگیری ماشین

- به طور کلی، راهکارهای حفظ حریم خصوصی تفاضلی در یادگیری ماشین را می‌توان به دو دسته تقسیم کرد:

1) راهکارهای مبتنی بر افزودن نویز به داده‌ها
   - یادگیری باناظر
   - یادگیری بدون ناظر
2) راهکارهای مبتنی بر لحاظ کردن حریم خصوصی تفاضلی در توابع هدف

# یادگیری باناظر حافظ حریم خصوصی تفاضلی

- یکی از بیشترین الگوریتم‌هایی که در این بخش بر روی آن کار شده، درخت تصمیم است.
- برخی راهکارهای ارائه شده:
  - افزودن نویز به بهره اطلاعاتی هر ویژگی (بلوم و همکاران ۲۰۰۶)
  - افزودن نویز نمایی و استفاده از سازوکار احتمالاتی در گام انتخاب ویژگی (فرایدمن و همکاران ۲۰۱۰)
  - افزودن نویز لاپلاس و انتخاب تصادفی ویژگی‌ها (جاگاناتام و همکاران ۲۰۱۲)
- مزایای این دسته از روش‌ها پیاده‌سازی آسان آن‌هاست.
- معایب این دسته از روش‌ها استفاده چندباره از بودجه حریم خصوصی و در نتیجه پایین بودن دقت و کارآیی است.

# یادگیری بدون‌ناظر حافظ حریم خصوصی تفاضلی

- یکی از معروف‌ترین الگوریتم‌های این بخش، خوشه بندی است
- هدف خوشه بندی حافظ حریم خصوصی، افزودن عدم قطعیت به مراکز خوشه‌ها و تعداد داده‌های هر خوشه است.
- افزودن نویز به مراکز خوشه‌ها به علت حساسیت زیاد آن‌ها، غیرعملی است.
- برخی راهکارهای ارائه شده:
  - محاسبه حساسیت محلی مراکز خوشه‌ها به هر یک از نقاط و افزودن نویز به تعدادی از نقاط هر خوشه، به گونه‌ای که مرکز خوشه ثابت بماند. (نیسیم و همکاران ۲۰۰۷) و (ونگ و همکاران ۲۰۱۵)
  - افزودن نویز با حفظ فاصله بین نقاط (ونگ و همکاران ۲۰۱۶)

# لحاظ کردن حریم خصوصی تفاضلی در توابع هدف

- طراحی یک یادگیرنده حافظ حریم خصوصی که خروجی آن مدلی با دقت مناسب و حافظ حریم خصوصی داده‌های یادگیری باشد.
- برخی راهکارهای ارائه شده:
  - طراحی یک مدل یادگیری عمیق توزیع‌شده (شکری و همکاران ۲۰۱۵)
  - تعریف یک تابع زیان حافظ حریم خصوصی با افزودن نویز به هرگام SGD (آبادی و همکاران ۲۰۱۶)
  - طراحی تابع هدف حافظ حریم خصوصی برای خودکدگذار (فان و همکاران ۲۰۱۶)
  - شبکه های عصبی بازگشتی حافظ حریم خصوصی (مکماهان و همکاران ۲۰۱۸)
  - روشی برای یادگیری متحدانه حافظ حریم خصوصی (تروئکس و همکاران ۲۰۱۹)

# کاربردهای دیگر حریم خصوصی تفاضلی

- استفاده از حریم خصوصی تفاضلی در انتشار داده (پایو و همکاران، ۲۰۱۹) (ژانگ و همکاران ۲۰۱۹)
- حریم خصوصی تفاضلی محلی (یه و همکاران، ۲۰۱۹) (چودوری، ۲۰۱۹)
- حریم خصوصی تفاضلی در زنجیره بلوکی (گای و همکاران ۲۰۱۹)
- حفظ حریم خصوصی تفاضلی در شبکه‌های اجتماعی (هوانگ و همکاران، ۲۰۲۰)

# جمع بندی و نتیجه‌گیری

- ارزشمند بودن داده‌ها و لزوم حفاظت از آن‌ها
- لزوم توجه به حریم خصوصی در یادگیری ماشین
- بررسی انواع حملات به حریم خصوصی در یادگیری ماشین
- معرفی حریم خصوصی تفاضلی
- بررسی راهکارهای مبتنی بر حریم خصوصی تفاضلی در یادگیری ماشین
- برخی از مسائل باز این زمینه:
  - حریم خصوصی تفاضلی محلی
  - حریم خصوصی تفاضلی برای داده‌های مرتبط
  - طراحی سازوکار حافظ حریم خصوصی تفاضلی

# منابع و مراجع


1. Feng, J., & Jain, A. K. (2010). Fingerprint reconstruction: from minutiae to phase. IEEE transactions on pattern analysis and machine intelligence, 33(2), 209-223.
2. Al-Rubaie, M., & Chang, J. M. (2016). Reconstruction attacks against mobile-based continuous authentication systems in the cloud. IEEE Transactions on Information Forensics and Security, 11(12), 2648-2663.
3. Gambs, S., Gmati, A., & Hurfin, M. (2012, July). Reconstruction attack through classifier analysis. In IFIP Annual Conference on Data and Applications Security and Privacy (pp. 274-281). Springer, Berlin, Heidelberg.
4. Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).
5. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., & Song, D. (2019). The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks. arXiv preprint arXiv:1911.07135.
6. He, Z., Zhang, T., & Lee, R. B. (2019, December). Model inversion attacks against collaborative inference. In Proceedings of the 35th Annual Computer Security Applications Conference (pp. 148-162).
7. Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In 2017 IEEE S&P (pp. 3-18). IEEE.
8. Rahman, M. A., Rahman, T., Laganière, R., Mohammed, N., & Wang, Y. (2018). Membership Inference Attack against Differentially Private Deep Learning Model. Transactions on Data Privacy, 11(1), 61-79.

# منابع و مراجع - ادامه


9. Mahjabin Tonni, S., Farokhi, F., Vatsalan, D., & Kaafar, D. (2020). Data and Model Dependencies of Membership Inference Attack. arXiv, arXiv-2002.

10. Xie, P., Bilenko, M., Finley, T., Gilad-Bachrach, R., Lauter, K., & Naehrig, M. (2014). Crypto-nets: Neural networks over encrypted data. arXiv preprint arXiv:1412.6181.

11. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016, June). Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In International Conference on Machine Learning (pp. 201-210).

12. Xu, R., Joshi, J. B., & Li, C. (2019, July). CryptoNN: Training Neural Networks over Encrypted Data. In 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS) (pp. 1199-1209). IEEE.

13. Hesamifard, E., Takabi, H., & Ghasemi, M. (2017). Cryptodl: Deep neural networks over encrypted data. arXiv preprint arXiv:1711.05189.

14. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006, March). Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference (pp. 265-284). Springer, Berlin, Heidelberg.

15. Dwork, C. (2011). A firm foundation for private data analysis. Communications of the ACM, 54(1), 86-95.

16. McSherry, F., & Talwar, K. (2007, October). Mechanism design via differential privacy. In 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07) (pp. 94-103). IEEE.

# منابع و مراجع - ادامه


17. Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005, June). Practical privacy: the SuLQ framework. In Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 128-138).

18. Friedman, A., & Schuster, A. (2010, July). Data mining with differential privacy. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 493-502).

19. Jagannathan, G., Pillaipakkamnatt, K., & Wright, R. N. (2009, December). A practical differentially private random decision tree classifier. In 2009 IEEE International Conference on Data Mining Workshops (pp. 114-121). IEEE.

20. Nissim, K., Raskhodnikova, S., & Smith, A. (2007, June). Smooth sensitivity and sampling in private data analysis. In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing (pp. 75-84).

21. Wang, Y., Wang, Y. X., & Singh, A. (2015). Differentially private subspace clustering. In Advances in Neural Information Processing Systems (pp. 1000-1008).

22. Wang, Y., Wang, Y. X., & Singh, A. (2018). A theoretical analysis of noisy sparse subspace clustering on dimensionality-reduced data. IEEE Transactions on Information Theory, 65(2), 685-706.

23. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).

# منابع و مراجع - ادامه


24. Shokri, R., & Shmatikov, V. (2015, October). Privacy-preserving deep learning. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1310-1321).

25. Phan, N., Wang, Y., Wu, X., & Dou, D. (2016, February). Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In Thirtieth AAAI Conference on Artificial Intelligence.

26. McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963.

27. Truex, S., Baracaldo, N., Anwar, A., Steinke, T., Ludwig, H., Zhang, R., & Zhou, Y. (2019, November). A hybrid approach to privacy-preserving federated learning. In Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (pp. 1-11).

28. Gai, Keke, et al. "Differential Privacy-based Blockchain for Industrial Internet of Things." IEEE Transactions on Industrial Informatics (2019).

29. Ye, Qingqing, et al. "PrivKV: Key-value data collection with local differential privacy." 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019.

30. Piao, Chunhui, et al. "Privacy-preserving governmental data publishing: A fog-computing-based differential privacy approach." Future Generation Computer Systems 90 (2019): 158-174.

31. Zhang, Sen, and Weiwei Ni. "Graph Embedding Matrix Sharing With Differential Privacy." IEEE Access 7 (2019): 89390-89399.


با تشکر از توجه شما