

NIVERSIDADE FEDERAL FLUMINENSE
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENADORIA DE PÓS-GRADUAÇÃO

CADASTRAMENTO DE DISCIPLINAS - *Stricto Sensu*

Nome do Curso ou Programa: Programa de Pós-Graduação em Computação

Nome da Disciplina:

Desaprendizado de Máquinas

Ministrada : ☐ ME ☐ DO ☒ Ambos

Carga Horária/Créditos

Teóricos		Téorico-Práticos		Trabalho Orientado / Est. Superv.		Total	
Carga Horária	Nº de Créditos	Carga Horária	Nº de Créditos	Carga Horária	Nº de Créditos	Carga Horária	Nº de Créditos
		60h	4			60h	4

Ementa da Disciplina:

A crescente adoção de sistemas baseados em Inteligência Artificial (IA) trouxe à tona desafios significativos relacionados à privacidade, conformidade regulatória e segurança da informação. A dependência de modelos de Machine Learning (ML), frequentemente operando como "caixas-pretas", revelou vulnerabilidades críticas que afetam não apenas sua eficácia, mas também a segurança, a privacidade e a confiança dos usuários. Diante desse cenário, a análise de segurança dos modelos de IA e o desenvolvimento de soluções alinhadas aos desafios tecnológicos atuais tornam-se imperativos.

O Desaprendizado de Máquina (MU) surge como uma necessidade crítica para permitir que modelos treinados possam "esquecer" informações específicas de forma eficiente, sem a necessidade de retreinamento completo. Essa capacidade é uma resposta direta ao "Direito ao Esquecimento" (Right to be Forgotten), um imperativo legal e ético previsto em regulamentações como o General Data Protection Regulation (GDPR) e a Lei Geral de Proteção de Dados (LGPD). A simples exclusão de dados brutos é insuficiente, pois os modelos de ML retêm a influência desses dados em seus parâmetros, podendo levar a violações de privacidade. Além da conformidade regulatória, o MU contribui para a melhoria da justiça, equidade e confiabilidade dos modelos de IA, permitindo a remoção de dados enviesados ou desatualizados. A complexidade de implementar mecanismos de esquecimento em ML é um desafio, especialmente para modelos de larga escala, onde o retreinamento completo é computacionalmente proibitivo. O objetivo do MU é desenvolver algoritmos que removam eficientemente a influência de um subconjunto de dados de treinamento, resultando em um modelo indistinguível de um treinado desde o início sem esses dados.

Paralelamente, a segurança dos algoritmos de ML tornou-se uma preocupação central. Modelos de última geração são surpreendentemente frágeis a ataques adversariais, que manipulam deliberadamente a entrada com perturbações sutis para induzir previsões incorretas. As vulnerabilidades levantam questões sobre a segurança e confiabilidade da IA. A interrelação entre MU e segurança é complexa: o desaprendizado pode expor informações sobre dados removidos a ataques de inferência de pertencimento, mas também pode ser uma defesa contra envenenamento de dados. A transição para sistemas multi-agente (MAS) adiciona complexidades significativas ao MU, exigindo coordenação, consenso e privacidade em um ambiente distribuído. Portanto, é fundamental que os métodos de desaprendizado sejam projetados não apenas para eficiência na remoção de dados, mas também para resistência contra agentes mal-intencionados.

Esta disciplina visa explorar esses focos, investigando vulnerabilidades, ataques e desenvolvendo técnicas mais seguras para modelos de MU, tanto em ambientes centralizados quanto em sistemas multi-agente.

Ao final desta disciplina, os alunos serão capazes de: (i) Compreender os fundamentos de MU, incluindo os conceitos teóricos e a motivação por trás deles; (ii) Analisar técnicas de MU, avaliando as diferentes abordagens, distinguindo entre métodos exatos e aproximados, suas limitações e aplicações em modelos de larga escala; e, (iii) Identificar vulnerabilidades em modelos de MU, incluindo o reconhecimento dos tipos de ataques (evasão, envenenamento de dados, inferência de pertencimento) e suas implicações para a segurança e confiabilidade de sistemas de IA.

A curso terá um formato misto, entre aulas expositivas e seminários. Ao final, o aluno deverá apresentar um projeto de curso sobre um tema a ser proposto.

Bibliografia Básica:

Desaprendizado de Máquinas e a LGPD: da privacidade ao direito ao esquecimento

Daniel O. C. Cota, Daniel Carlos S. de Jesus, Antonio A. de A. Rocha

DOI: <https://doi.org/10.5753/sbc.15408.7.4>

Bibliografia Complementar:

- [1] Juliussen, B. A. (2023). Algorithms that forget: Machine unlearning and the right to erasure. *Computer Law & Security Review*, 49, 105832.
<https://www.sciencedirect.com/science/article/pii/S026736492300095X>
- [2] Zhao, Z. (2022). The Application of the Right to be Forgotten in the Machine Learning Context. *Journal of Law and Technology*, 31(1), 5. <https://scholarship.law.edu/jlt/vol31/iss1/5/>
- [3] Bourtole, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., ... & Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 141-159). IEEE.
- [4] Sekhari, A., Acharya, J., Kamath, G., & Suresh, A. T. (2021). Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34, 17805-17816.
- [5] Xu, H., Zhang, H., Wang, Y., & Li, Y. (2023). Machine Unlearning: A Survey. *arXiv preprint arXiv:2306.03558*. <https://arxiv.org/abs/2306.03558>
- [6] Nguyen, T. T., Hu, J., Oria, V., & Lupu, M. (2022). A Survey of Machine Unlearning. *arXiv preprint arXiv:2209.02299*. <https://arxiv.org/abs/2209.02299>
- [7] Qu, Y., Liu, Y., Li, B., & Wang, L. (2023). Learn to Unlearn: A Survey on Machine Unlearning. *arXiv preprint arXiv:2305.07512*. <http://ui.adsabs.harvard.edu/abs/2023arXiv230507512Q>
- [8] Goldblum, M., Fowl, L., & Goldstein, T. (2020). Data poisoning attacks against federated learning. In *International Conference on Machine Learning* (pp. 3623-3632). PMLR.
- [9] Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., & Zhang, Y. (2021). When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (pp. 896-911).
- [10] Pawelczyk, M., Di, J. Z., Lu, Y. G., Sekhari, A., Kamath, G., & Goldstein, T. (2024). Machine Unlearning Fails to Remove Data Poisoning Attacks. *arXiv preprint arXiv:2406.17216*. <https://arxiv.org/abs/2406.17216>

A SER PREENCHIDO
PELA PROPP

Código da Disciplina:

SIGLA

S

Nº DE CRÉD.

SEQ. POR ÓRGÃO