Prof. Alfio Ferrara

Introduction to Data Science for the Humanities

PhD in Philosophy and Human Sciences

Data analytics and machine learning

Lecture 1: Introduction to Data Science

`May 25, room Martinetti, 10:30 – 12:30`

A brief history of Artificial Intelligence and data science. The Data revolution. Models of machine learning. Deep learning. Ethical and social issues.

Lecture 2: The algorithmic tools of machine learning

`May 30, room Martinetti, 10:30 – 12:30`

Statistical learning. From data to their mathematical representation. Probabilistic models vs the vector space. Examples of image and textual encoding. Introduction to linear transformation and neural networks.

Lecture 3: Unsupervised learning

`June 6, room Martinetti, 10:30 – 12:30`

Principles of unsupervised learning. KMeans, an example of a clustering algorithm. Case study on image clustering.

Lecture 4: Supervised learning

`June 9, room Martinetti, 10:30 – 12:30`

Principles of supervised learning. Working with textual data. Intuition of language modeling and Recurrent Neural Networks. A case study on author prediction and text generation.

Lecture 5: Reinforcement learning and evolving neural networks

`June 13, room Martinetti, 10:30 – 12:30`

Intuition of learning by reinforcement. Differences between reinforcement learning and evolution. A case study on simulating natural and artificial selection.

Lecture 6: Deep Learning and introduction to Language Models

`June 16, room Martinetti, 10:30 – 12:30`

Introduction to Language Models. Statistical vs Neural language models. Main tasks that can be addressed using LMs.

Social and Ethical Issues in NLP

Lecture 7: Introduction to Large Language Models (LLMs)

`June 20, room Martinetti, 10:30 – 12:30`

Neural Language Models. Sequence to Sequence learning, Recurrent Neural Networks (RNN), Encoder-Decoder architectures, Attention and Transformers.

Lecture 8: Black Box models and Explainable AI

`June 23, room Martinetti, 10:30 – 12:30`

Introduction to the problem of explainable AI. Explainability and Causality. The role of Attention in the explanation of LLMs.

Lecture 9: Introduction to Social and Ethical Bias

`June 27, room Martinetti, 10:30 – 12:30`

Algorithmic bias. Hate-speech, toxicity. Privacy violation and profiling. Misinformation, fake-news, information and opinion manipulation. Technological divide.

Lecture 10: Automatic Detection of Implicit Bias and Stereotypes

`June 30, room Martinetti, 10:30 - 12:30`

Overview of the literature on the main approaches to automatic detection of bias and stereotypes.

References

Barrett, M., Kementchedjhieva, Y., Elazar, Y., Elliott, D., & Søgaard, A. (2019, November). Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6330-6335).

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Kalai, A. T. (2019, January). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120-128).

Díaz, M., Johnson, I., Lazar, A., Piper, A. M., & Gergle, D. (2018, April). Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems* (pp. 1-14).

Dixon, L., Li, J., Sorensen, J., Thain, N., & Vasserman, L. (2018, December). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 67-73).

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*(16), E3635-E3644.

Gonen, H., & Goldberg, Y. (2019). Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Hendricks, L. A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 771-787).

Hu, Z., & Strout, J. (2018). Exploring stereotypes and biased data with the crowd. *arXiv preprint arXiv:1801.03261*.

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Park, J. H., Shin, J., & Fung, P. (2018). Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Prates, M. O., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, *32*, 6363-6381.

Swinger, N., De-Arteaga, M., Heffernan IV, N. T., Leiserson, M. D., & Kalai, A. T. (2019, January). What are the biases in my word embedding?. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 305-311).

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, *6*, 605-617.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2018). Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335-340).

Zhao, J., Zhou, Y., Li, Z., Wang, W., & Chang, K. W. (2018). Learning gender-neutral word embeddings. *arXiv preprint arXiv:1809.01496*.