FEATURE SELECTION FOR MACHINE LEARNING WITH UNSTRUCTURED DATA: APPLICATIONS IN HEALTHCARE

KIMBERLY VILLALOBOS CARBALLO

Abstract. Multimodal machine learning models trained on unstructured data—such as clinical text and images—have achieved remarkable predictive performance in healthcare tasks ranging from diagnosis to outcome prediction [1]. However, these models remain fundamentally limited in their ability to provide *faithful* interpretations for their outputs. In particular, they cannot *reliably* answer questions such as why a prediction was made or what information was most influential. This motivates the study of the classical feature selection problem [2] in the context of unstructured data: identifying small, informative subsets of the input (e.g., sentences in a text) that are sufficient to explain and drive model predictions.

In this talk, we address this challenge by formulating feature selection as a learning problem that is solved jointly with the original predictive task, leading to a maximum-likelihood optimization problem over the parameters of both the predictive model and the feature selection function. While a direct formulation of the problem yields an intractable combinatorial problem, we introduce a sequential approximation based on conditional probabilities that produces sparse, interpretable models while remaining computationally tractable.

Empirical results on real-world healthcare datasets demonstrate the effectiveness of our approach. For end-of-stay mortality and next-day discharge prediction tasks using intensive care unit clinical notes from MIMIC-IV [3], the method yields highly sparse models that rely on only a small number of sentences per patient while achieving predictive performance comparable to models trained on the full text. In a separate experiment with data from the Framingham Heart Study for coronary heart disease risk prediction [4], we assess the clinical relevance of the most frequently selected sentences and find that they align closely with established clinical risk factors, demonstrating the method's ability to identify clinically meaningful predictive features from unstructured data.

Keywords: multimodal machine learning, feature selection, healthcare analytics, large language models, interpretable AI.

Mathematics Subject Classifications (2020): 90C15, 90C30, 90C59.

References

[1] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, and D. Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, 5(1):149, 2022.

[2] B. Venkatesh and J. Anuradha. A review of feature selection and its methods. *Cybern. Inf. Technol.*, 19(1):3–26, 2019.

[3] A. E. W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[4] C. Andersson, A. D. Johnson, E. J. Benjamin, D. Levy, and R. S. Vasan. 70-year legacy of the Framingham Heart Study. *Nature Reviews Cardiology*, 16(11):687–698, 2019.

Tandon School of Engineering, New York University, NY, USA, 11201
Email address: kimberly.v@nyu.edu