UMEÅ UNIVERSITET

Probabilistic Metric Space for Machine Learning: Data and Model Spaces

Mariam Taha

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av Teknologie doktorsexamen framläggs till offentligt försvar i HÖRSAL NAT.D. 360, byggnad Naturvetarhuset, fredag den 23 maj, kl. 09:00-12.00.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Susana Montes, Department of Statistics and Operational Research, Oviedo University, Spanien.

Department of Computing Science

Organization
Umeå University
Dept. Of Computing Sciecne

Document type
Doctoral thesis

Date of publication
30 APRIL 2025

Author
Mariam Taha

Title
Probabilistic Metric Space for Machine Learning: Data and Model Spaces.

Abstract
Machine learning models are inherently shaped by the data used to train them. Understanding the relationship between datasets and the models they generate is essential for tasks such as model selection, privacy metrics, and robustness evaluation. This thesis presents a rigorous mathematical framework for comparing machine learning models and algorithms by formalizing the interaction between two fundamental spaces: the database space, which captures possible datasets, and the model space, which contains the models or hypotheses derived from those datasets. A central motivation stems from the observation that different datasets can lead to the same or highly similar models. Such recurrent models—which arise frequently across diverse data sources—are particularly significant in privacy-sensitive applications. Their recurrence suggests reduced dependence on any specific data point or subgroup, thus offering inherent privacy and generalization benefits. By quantifying the relationship between models and their generating data, this work enables principled evaluation of a model's robustness and disclosure risk.

To formalize relationships between the two spaces, the thesis develops a family of probabilistic metric space constructions tailored to different aspects of the data–model interaction. The first contribution models database evolution as a Markov process and defines probabilistic distances between models based on the likelihood of transitioning between their generating datasets. The second contribution introduces F-space, a framework based on fuzzy measures that captures richer structural properties of the data—such as redundancy, synergy, and overlap among subsets. Building on this, the third contribution applies the F-space theory in practical machine learning scenarios. It demonstrates how fuzzy measures can be used to compare different linear regression algorithms trained over structured subsets of real datasets. The final contribution further generalizes the framework through Generalized F-spaces, where the model space itself is endowed with probabilistic structure—allowing uncertainty in both the datasets and the model outputs to be captured simultaneously.

Together, these constructions offer a principled alternative to traditional model comparison metrics. Rather than relying solely on pointwise loss or accuracy, the proposed framework incorporates the diversity, dynamics, and internal structure of the data that underlies each model—enabling more robust and privacy-aware assessments.

Keywords: Probabilstic metric space, space of data, space of models, fuzzy measures

Language
English

ISBN
print: 978-91-8070-680-3
PDF: 978-91-8070-681-0

ISSN
0348-0542

Number of pages
50 + 4 papers