Mathematics and Machine Learning

Equipo organizador

- Santiago Mazuelas (Basque Center for Applied Mathematics)
- Jose Antonio Lozano (Basque Center for Applied Mathematics)

Descripción

The mathematics of machine learning traces back to foundational work carried out by several mathematicians in the 1980s, including Vladimir Vapnik and Leslie Valiant. This special session brings together leading national researchers approaching machine learning from a mathematical perspective. The invited speakers are active contributors to the field, regularly publishing in top-tier conferences and journals in machine learning and artificial intelligence. The talks will span a range of research areas, including Bayesian methods, kernel techniques, generative models, and minimax approaches. The session has two main objectives: to promote broader engagement with the mathematics of machine learning within the national mathematical community, and to strengthen the currently small and scattered group of national researchers working on these topics at a highly competitive international level.

Palabras clave: Supervised learning; Performance guarantees for learning; Bayesian methods for learning; Generative models; Minimax approaches for learning; Reproducing kernel Hilbert spaces.

Programa

MARTES, 20 de enero

11:00 – 11:30	Andres Masegosa (University of Aalborg) *From PAC-Chernoff Bounds to Practice: Smooth Interpolators and Near-Optimal Generalization in Deep Learning*
11:30 – 12:00	Pablo Moreno Muñoz (Pompeu Fabra) *Towards Understanding Generalization in LLMs: A Probabilistic Perspective*
12:00 – 12:30	Santiago Mazuelas (BCAM) *Beyond Empirical Risk Minimization*
15:30 – 16:00	Alberto Gonzalez Sanz (Columbia University) *Learning on the Space of Probability measures*
16:00 – 16:30	Verónica Álvarez (Massachusetts Institute of Technology) *Adaptive Supervised Learning in Time-Dependent Environments*
16:30 – 17:00	Pablo Morales Alvarez (Universidad de Granada) *SM: Enhanced Localization in Multiple Instance Learning for medical Imaging Classification*
17:00 – 17:30	Aritz Pérez (BCAM) *Bayesian Networks for Ranking Data*
18:00 – 18:30	Pablo Martínez Olmos (Universida Carlos III) *Training Implicit Generative Models via an Invariant Statisitical Loss*
18:30 – 19:00	Daniel Hernandez Lobato (Universidad Autónoma de Madrid) *Joint Entropy Search for Multi-objective Bayesian Optimization with Constraints and Multiple Fidelities*

From PAC-Chernoff Bounds to Practice: Smooth Interpolators and Near-Optimal Generalization in Deep Learning

Andres Masegosa

University of Aalborg

arma@cs.aau.dk

Resumen. Why do some over-parameterized models generalize remarkably well while others overfit—even when both perfectly interpolate the training data? This talk presents a unified perspective grounded in Large Deviation Theory, combining theoretical insights and practical algorithms to better understand and improve generalization in modern deep learning.

In the first part, we introduce PAC-Chernoff bounds—a new class of distribution-dependent generalization bounds that remain tight even for interpolating models. These bounds identify a natural smoothness measure derived from the rate function of the loss distribution. We show that many regularization strategies—such as L2 penalties, input-gradient constraints, and distance-to-initialization—along with architectural choices like invariance and over-parameterization, implicitly promote smoother interpolators. Our framework explains why such choices often lead to better generalization, even in regimes where classical bounds fail.

In the second part, we operationalize these theoretical insights by proposing the Inverse-Rate Regularizer (IRR)—a practical and scalable estimator of the optimal regularizer implied by PAC-Chernoff theory. IRR uses overlapping-batch estimates of the cumulant generating function to approximate the inverse of the rate function, leading to a principled and adaptive reweighting of training samples. This connects to techniques in sample-adaptive loss functions and distributionally robust optimization. Empirical results on standard vision benchmarks demonstrate that IRR enhances generalization, calibration, and robustness over strong baselines.

Together, these two works bridge the gap between theory and practice: from understanding why certain interpolators generalize, to designing how we can learn them more effectively.

Referencias

[1] A. R. Masegosa, L. A. Ortega (2025). PAC-Chernoff Bounds: Understanding Generalization in the Interpolation Regime. *Journal of Artificial Intelligence Research*, 82, 503-562.

[2] J. Hu, L. A. Ortega, T. M. Laleg, A. R. Masegosa (2025). Towards Near-Optimal Regularization in Deep Learning via the Inverse-Rate Regularizer. Under review, submitted to NeurIPS 2025.

Towards Understanding Generalization in LLMs: A Probabilistic Perspective

Pablo Moreno Muñoz

Universidad Pompeu Fabra

pablo.moreno@upf.edu

Resumen. The proliferation of foundation models in society, specifically large language models (LLMs) such as Llama, ChatGPT, or Gemini, has put the spotlight on their remarkable generalization abilities. While their predictive performance is impressive at the moment, there are still open questions on why these types of neural network models generalize better than other methods. Answering them is of critical importance for the future, due to security reasons, trustworthiness, energy cost auditing, or fair regulation. Among all questions, we can draw two here: i) Is it because of the learning algorithm? ii) or due to its stochastic nature that implicitly maximizes the evidence of the model? To answer them, looking back at probabilistic learning principles is needed. One good example of this is given by masked pre-training, one of the initial variants of self-supervised learning used in natural language processing (Devlin et al., 2018). Its main strength comes from removing random input dimensions, for later learning a model that can predict the self-induced missing values. Empirical results indicate that this intuitive form of self-supervised learning yields models that generalize very well to new domains. By building a new proof that relies on a previous observation from Fong and Holmes (2020), where log-marginal likelihood equals the average of exhaustive cross-validation, recent results show that such a form of self-supervised learning implicitly performs stochastic maximization of the model's marginal likelihood (Moreno-Muñoz et al. 2023). The latter density is generally acknowledged as being an excellent measure of a model's ability to generalize in the probabilistic setting. Last but not least, this sort of analysis is also validated by empirical results, which also show new ways to understand LLMs, which might change the way we build and conceive AI models in the near future.

Referencias

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[2] E. Fong and C. C. Holmes (2020). On the marginal likelihood and cross-validation. Biometrika, 107(2):489–496.

[3] P. Moreno-Muñoz, P. G. Recasens and S. Hauberg (2023). On Masked Pre-training and the Marginal Likelihood, In Advances in Neural Information Processing Systems (NeurIPS).

Beyond Empirical Risk Minimization

Santiago Mazuelas

Basque Center for Applied mathematics, BCAM

smazuelas@bcamath.org

Resumen. The empirical risk minimization (ERM) approach for supervised learning chooses prediction rules that fit training samples and are "simple" (generalize). This approach has been the workhorse of machine learning methods and has enabled a myriad of applications. However, ERM methods strongly rely on the specific training samples available and cannot easily address scenarios affected by distribution shifts or corrupted samples. Robust risk minimization (RRM) is an alternative approach that does not aim to fit training examples and instead chooses prediction rules minimizing the maximum expected loss (risk). This talk presents a learning framework based on the generalized maximum entropy principle that leads to minimax risk classifiers (MRCs). In particular, MRCs can minimize worst-case expected 0-1 loss while providing performance guarantees, and are strongly universally consistent using feature mappings given by characteristic kernels. In addition, the methods presented can provide techniques that are robust to practical situations that defy conventional assumptions, e.g., training samples that follow distributions that change over time.

Referencias

[1] S. Mazuelas, M. Romero, and P. Grunwald (2023). Minimax risk classifiers with 0-1 loss. *Journal of Machine Learning Research* 1–48.

[2] S. Mazuelas, Y. Shen, and A. Perez (2022). Generalized maximum entropy for supervised classification. *IEEE Transactions on Information Theory*, 4, 2530-2550.

[3] P. Grunwald and A. P. Dawid (2004). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *The Annals of Statistics*, 32, 1367–1433.

Learning on the Space of Probability Measures

Alberto Gonzalez Sanz

Columbia University

ag4855@columbia.edu

Resumen. The distribution regression problem encompasses many important statistics and machine learning tasks, and arises in a large range of applications. Among various existing approaches to tackle this problem, kernel methods have become a method of choice. Indeed, kernel distribution regression is both computationally favorable, and supported by a recent learning theory. This theory also tackles the two-stage sampling setting, where only samples from the input distributions are available.
In this talk, we will present new advances in the learning theory of kernel distribution regression. Focusing on kernels based on Hilbertian embeddings, which cover most existing methods, we introduce a novel near-unbiasedness condition on these embeddings. This condition allows us to derive sharper error bounds for the two-stage sampling setting through a new analysis of the sampling effect. We show that this condition holds for key families of kernels, including those based on optimal transport and mean embeddings. As a result, we establish improved convergence rates for these widely used kernels. Finally, we will illustrate these findings with numerical experiments.

Referencias

[1] F. Bachoc, L. Béthune, A. González-Sanz, J-M. Loubes (2025). Improved learning theory for kernel distribution regression with two-stage sampling. Accepted in the Annals of Statistics.

[2] F. Bachoc, L. Béthune, A. González-Sanz and J-M. Loubes (2023). Gaussian Processes on Distributions based on Regularized Optimal Transport. Proceedings of The 26th International Conference on Artificial Intelligence and Statistics. pp. 4986-5010.

Adaptive Supervised Learning in Time-Dependent Environments

Verónica Álvarez

Massachusetts Institute of Technology

vealvar@mit.edu

Resumen. The statistical characteristics of instance-label pairs often change with time in practical scenarios of supervised classification. Conventional learning techniques adapt to such concept drift accounting for a scalar rate of change by means of a carefully chosen learning rate, forgetting factor, or window size. However, the time changes in common scenarios are multidimensional, i.e., different statistical characteristics often change in a different manner. We propose adaptive minimax risk classifiers (AMRCs) that account for multidimensional time changes by means of a multivariate and high-order tracking of the time-varying underlying distribution. In addition, differently from conventional techniques, AMRCs can provide computable tight performance guarantees. Experiments on multiple benchmark datasets show the classification improvement of AMRCs compared to the state-of-the-art and the reliability of the presented performance guarantees.

Referencias

[1] V. Álvarez, S. Mazuelas, J.A. Lozano (2025). Supervised Learning with Evolving Tasks and Performance Guarantees. *Journal of Machine Learning Research* 26 (17), 1-59.

[2] V. Álvarez, S. Mazuelas, J.A Lozano (2023). Minimax forward and backward learning of evolving tasks with performance guarantees. *Advances in Neural Information Processing Systems (Neurips)* 36, 65678-65702.

[3] V. Álvarez, S. Mazuelas, J.A. Lozano (2022). Minimax classification under concept drift with multidimensional adaptation and performance guarantees. *International Conference on Machine Learning (ICML)*, 486-499.

SM: Enhanced Localization in Multiple Instance Learning for Medical Imaging Classification

Pablo Morales Alvarez

Universidad de Granada

pablomorales@ugr.es

Resumen. Multiple Instance Learning (MIL) is widely used in medical imaging classification to reduce the labeling effort. While only bag labels are available for training, one typically seeks predictions at both bag and instance levels (classification and localization tasks, respectively). Early MIL methods treated the instances in a bag independently. Recent methods account for global and local dependencies among instances. Although they have yielded excellent results in classification, their performance in terms of localization is comparatively limited. We argue that these models have been designed to target the classification task, while implications at the instance level have not been deeply investigated. Motivated by a simple observation – that neighboring instances are likely to have the same label – we propose a novel, principled, and flexible mechanism to model local dependencies. It can be used alone or combined with any mechanism to model global dependencies (e.g., transformers). A thorough empirical validation shows that our module leads to state-of-the-art performance in localization while being competitive or superior in classification. Our code is publicly availabe at GitHub.

Referencias

[1] F.M Castro-Macías, P. Morales-Álvarez, Y Wu, R Molina, AK Katsaggelos (2024). Sm: enhanced localization in Multiple Instance Learning for medical imaging classification. *Advances in Neural Information Processing Systems 38 (NeurIPS 2024)*

[2] P. Morales-Álvarez, A. Schmidt, J.M. Hernández-Lobato, R. Molina (2024). Introducing instance label correlation in multiple instance learning. Application to cancer detection on histopathological images. *Pattern Recognition*

[3] A. Schmidt, P. Morales-Álvarez, R. Molina (2023). Probabilistic Modeling of Inter- and Intra-observer Variability in Medical Image Segmentation. *International Conference on Computer Vision (ICCV)*.

Bayesian Networks for Ranking Data

Aritz Pérez

BCAM

aperez@bcamath.org

Resumen. Modeling ranking data is a central task in many applications, from decision-making and recommendation systems to sports analytics and elections. A fully general model that represents a distribution over rankings requires a factorial number of parameters, making inference and learning computationally intractable. To address this, existing models introduce independence assumptions tailored to the structural constraints of ranking data. These assumptions reduce the number of parameters, simplify inference, and enhance the interpretability of the model. In this talk, I will introduce Graphical Ranking Models (GRMs), a framework to model ranking data based on the standard notions of conditional independence. GRMs rely on ranking variables—categorical variables that represent relative rankings and support the use of conditional independence. In essence, GRMs are Bayesian networks defined on ranking variables and represent probability distributions over partitions of the ranking space. I will show that GRMs generalize several existing models and independence notions for rankings—such as the repeated insertion model and the riffle-shuffle model. I will present the theoretical foundations of GRMs, highlight their expressive power and practical advantages, and discuss open challenges and research opportunities at the intersection of ranking models and probabilistic graphical models.

Training Implicit Generative Models via an Invariant Statisitical Loss

Pablo Martínez Olmos

Universida Carlos III

pamartin@ing.uc3m.es

Resumen. Implicit generative models have the capability to learn arbitrary complex data distributions. On the downside, training requires telling apart real data from artificially-generated ones using adversarial discriminators, leading to unstable training and mode-dropping issues. As reported by Zahee et al. (2017), even in the one-dimensional (1D) case, training a generative adversarial network (GAN) is challenging and often suboptimal. In this work, we develop a discriminator-free method for training one-dimensional (1D) generative implicit models and subsequently expand this method to accommodate multivariate cases. Our loss function is a discrepancy measure between a suitably chosen transformation of the model samples and a uniform distribution; hence, it is invariant with respect to the true distribution of the data. We first formulate our method for 1D random variables, providing an effective solution for approximate reparameterization of arbitrary complex distributions. Then, we consider the temporal setting (both univariate and multivariate), in which we model the conditional distribution of each sample given the history of the process. We demonstrate through numerical simulations that this new method yields promising results, successfully learning true distributions in a variety of scenarios and mitigating some of the well-known problems that state-of-the-art implicit methods present.

Joint Entropy Search for Multi-objective Bayesian Optimization with Constraints and Multiple Fidelities

Daniel Hernandez Lobato

Universidad Autónoma de Madrid

daniel.hernandez@uam.es

Resumen. Bayesian optimization (BO) methods can be used to solve efficiently problems with several objectives and constraints. Each objective and constraint is considered a black-box function that is expensive to evaluate, lacking also a closed-form expression. BO methods use a model of each black-box to guide the search for the problem's solution. Specifically, they make intelligent decisions about where each black-box function should be evaluated next with the goal of finding the solution using a few evaluations only. Sometimes, however, the black-boxes may be evaluated at different fidelity levels. A lower fidelity is simply a cheap proxy of the corresponding black-box. These lower fidelities correlate with the actual black-boxes to optimize and can, therefore, be used to reduce the overall cost of solving the optimization problem. Here, we propose Multi-fidelity Joint Entropy Search for Multi-objective Bayesian Optimization with Constraints (MF-JESMOC), a BO method for solving the aforementioned problems. MF-JESMOC chooses the next point, and fidelity level at which to evaluate the black-boxes, as the combination that is expected to reduce the most the joint entropy of the Pareto set and the Pareto front, normalized by the fidelity's cost. We use Deep Gaussian processes to model each black-box and the dependencies between fidelities. These are powerful probabilistic models that can learn the dependency structure among fidelity levels of each black-box. Several experiments show that MF-JESMOC outperforms other state-of-the-art methods for multi-objective BO with constraints and different fidelity levels in both synthetic and real-world problems.

Referencias

[1] D. Fernández-Sánchez, D. Hernández-Lobato (2024). Joint entropy search for multi-objective Bayesian optimization with constraints and multiple fidelities, *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

[2] E. C. Garrido-Merchán, D. Hernández-Lobato (2019). Predictive entropy search for multi-objective Bayesian optimization with constraints, *Neurocomputing* 361 50–68.

[3] C. Hvarfner, F. Hutter, L. Nardi (2022). Joint entropy search for maximally-informed Bayesian optimization. *Advances in Neural Information Processing Systems* (Neurips).