Berlin-Bielefeld-Paris Workshop on Early Stopping

Berlin, April 13/14th, 2023

Abstracts

Cross-validation for estimator selection

Sylvain Arlot
Université Paris-Saclay

Cross-validation is a widespread strategy because of its simplicity and its (apparent) universality. It can be used with two main goals: (i) estimating the risk of an estimator, and (ii) model selection or hyperparameter tuning, or more generally for choosing among a family of estimators. Many results exist on the performance of cross-validation procedures, which can strongly depend on the goal for which it is used.

This talk will show the big picture of these results, with an emphasis on the goal of estimator selection. In short, at first order (when the sample size goes to infinity), the key parameter is the bias of cross-validation, which only depends on the size of the training set. Nevertheless, "second-order" terms do matter in practice, and I will show recent results on the role of the "variance" of cross-validation procedures on their performance. As a conclusion, I will provide some guidelines for choosing the best cross-validation procedure according to the particular features of the problem at hand.

References:
Survey paper (with Alain Celisse): http://projecteuclid.org/euclid.ssu/1268143839
Paper about V-fold cross-validation in least-squares density estimation (with Matthieu Lerasle):
http://jmlr.org/papers/v17/14-296.html

Early stopping for conjugate gradients in statistical inverse problems

Laura Hucker
Humboldt-Universität zu Berlin

We consider estimators obtained by applying the conjugate gradient algorithm to the normal equation of a prototypical statistical inverse problem. For such iterative procedures, it is necessary to choose a suitable iteration index to avoid under- and overfitting. Unfortunately, classical model selection criteria can be prohibitively expensive in high dimensions. In contrast, it has been shown for several methods that sequential early stopping can achieve statistical and computational efficiency by halting at a data-driven index depending on previous iterates only. Residual-based stopping rules, similar to the discrepancy principle for deterministic problems, are well understood for linear regularization methods. In the case of conjugate gradients, the estimator depends nonlinearly on the observations, allowing for greater flexibility. This significantly complicates the error analysis. Our goal is to establish adaptation results in this setting.

Early stopping of untrained convolutional networks

Tim Jahn
Universität Bonn

In recent years new regularisation methods based on neural networks have shown promising performance for the solution of ill-posed problems, e.g., in imaging science. Due to the non-linearity of the networks, these methods often lack profound theoretical justification. In this talk we rigorously discuss convergence for an untrained convolutional network. Untrained networks are

particulary attractive for applications, since they do not require any training data. Its regularising property is solely based on the architecture of the network. Because of this, appropriate early stopping is essential for the success of the method. We show that the discrepancy principle is an adequate method for early stopping here, as it yields minimax optimal convergence rates.

Learning linear operators: Infinite-dimensional regression as a well-behaved non-compact inverse problem

Nicole Mücke
TU Braunschweig

We consider the problem of learning a linear operator θ between two Hilbert spaces from empirical observations, which we interpret as least squares regression in infinite dimensions. We show that this goal can be reformulated as an inverse problem for θ with the undesirable feature that its forward operator is generally non-compact (even if θ is assumed to be compact or of p-Schatten class). However, we prove that, in terms of spectral properties and regularisation theory, this inverse problem is equivalent to the known compact inverse problem associated with scalar response regression. Our framework allows for the elegant derivation of dimension-free rates for generic learning algorithms under Hölder-type source conditions. The proofs rely on the combination of techniques from kernel regression with recent results on concentration of measure for sub- exponential Hilbertian random variables. The obtained rates hold for a variety of practically- relevant scenarios in functional regression as well as nonlinear regression with operator- valued kernels and match those of classical kernel regression with scalar response. Joint work with Mattes Mollenhauer (FU Berlin), Tim Sullivan (U. Warwick).

Early stopping and regularization: From kernel methods to variable importance in neural networks

Garvesh Raskutti
University of Wisconsin-Madison

Early stopping of iterative methods is known to often induce implicit regularization. Perhaps the most well-understood example is the connection/approximate equivalence between kernel ridge regression and early stopping of gradient descent. In this talk, I explore this connection for the more modern problem of estimating variable importance for black-box methods such as neural networks. Two general approaches for estimating variable importance are: dropout methods which trains a single full model using all variables and dropping out the variable of interest from the trained full model and retrain methods that retrain the model every time the variable of interest is dropped out and compare the change in loss to the full model. Retrain methods are generally viewed as more accurate than dropout methods since they account for high dependence. However retrain methods are often computationally prohibitive since a new model needs to be trained for each variable. I introduce our lazy-variable (LazyVI) importance framework which bridges this computational-statistical tradeoff by using an ℓ_2-regularized first-order Taylor approximation centered around the original full model. Theoretical guarantees, a simulation study and a climate science application is provided to show that our LazyVI framework achieves accuracy quite close to retrain methods with a significant reduction in run-time. Potential connections to Shapley values are also discussed. Finally I discuss the connection between early stopping of gradient descent for neural networks (a widely used approach in practice) and our

LazyVI framework and potential future directions through understanding the eigen-spectra for the Neural Tangent Kernel for different neural network architectures.

Implicit regularization in statistical learning: An overview and some recent results

Patrick Rebeschini
University of Oxford

Recently, there has been a surge of interest in understanding the implicit regularization properties of iterative first-order methods applied to statistical learning problems. In this talk, we give an overview of some of the main ideas in this line of research and present recent results involving mirror descent, a generalization of gradient descent. In particular, we investigate the statistical guarantees on the excess risk achieved by early-stopped mirror descent on the unregularized empirical risk with the squared loss for linear models and kernel methods. We show that there is a direct link between the potential-based analysis of mirror descent from optimization theory and recent notions of localized Rademacher complexities from statistical learning theory. This link allows characterizing the statistical performance of the path traced by mirror descent in terms of offset complexities of function classes depending on the choice of the mirror map, initialization point, step size, and the number of iterations. (Based on joint works with Tomas Vaškevičius and Varun Kanade)

Early stopping for L^2-boosting in sparse high-dimensional linear models

Bernhard Stankewitz
Bocconi University

We consider L^2-boosting in a sparse high-dimensional linear model via orthogonal matching pursuit (OMP). For this greedy, nonlinear subspace selection procedure, we analyze a data-driven early stopping time τ, which is sequential in the sense that its computation is based on the first τ iterations only. Our approach is substantially less costly than established model selection criteria, which require the computation of the full boosting path.

We prove that sequential early stopping preserves statistical optimality in this setting in terms of a general oracle inequality for the empirical risk and recently established optimal convergence rates for the population risk. The proofs include a subtle ω-pointwise analysis of a stochastic bias-variance trade-off, which is induced by the greedy optimization procedure at the core of OMP. Simulation studies show that, at a significantly reduced computational cost, these types of methods match or even exceed the performance of other state of the art algorithms such as the cross-validated Lasso or model selection via a high-dimensional Akaike criterion based on the full boosting path.