Mathematical Foundations of Machine Learning

Balázs Csanád Csáji
Senior Researcher, MTA SZTAKI
balazs.csaji@sztaki.mta.hu

General Concept

The course aims at introducing the students to the mathematical background of machine learning, especially to the foundations of *statistical learning* and *reinforcement learning*, though some of the material is also important for *deep learning*. A number of standard mathematical concepts are overviewed, as well, denoted by the term "[Reminder]" below. These include the basics of linear regression, Markov chains and nonlinear optimization, which should ideally be known to the audience, and could be skipped depending on the students. The trade-off is that the more reminders are needed, the less proofs can be given for the main material.

[Reminder] Linear Regression

data fitting with basis functions ⋆ LS: least squares (ordinary, weighted, generalized, recursive, and least norm) ⋆ ridge regression ⋆ LASSO: least absolute shrinkage and selection operator ⋆ deterministic LS: normal equation ⋆ orthogonal projection ⋆ solution via QR factorization ⋆ SVD: singular value decomposition ⋆ low rank approximation ⋆ stochastic LS: mean and covariance ⋆ Gauss-Markov theorem ⋆ strong consistency ⋆ limiting distribution ⋆ LS and maximum likelihood estimation ⋆ statistical efficiency ⋆ confidence ellipsoids

[Reminder] Nonlinear Optimization

nonlinear optimization ⋆ conjugate function ⋆ Lagrangian duality ⋆ weak and strong duality ⋆ Karush-Kuhn-Tucker conditions ⋆ convex optimization ⋆ equivalent transformations ⋆ Slater's condition ⋆ Wolfe duality

Statistical Learning Theory (SLT)

classification and regression ⋆ loss ⋆ expected risk ⋆ empirical risk minimization ⋆ Bayes optimal classifier ⋆ consistency ⋆ no free lunch results ⋆ inductive bias ⋆ nearest neighbor classifiers ⋆ estimation error vs approximation error ⋆ bias-variance trade-off ⋆ underfitting vs overfitting ⋆ shattering ⋆ Vapnik-Chervonenkis (VC) dimension ⋆ generalization bounds ⋆ structural risk minimization ⋆ linear classification: canonical parametrization and support vectors ⋆ Vapnik's (hard and soft margin) support vector classification (SVC) ⋆ least-squares support vector machines ⋆ Wolfe dual of SVC ⋆ nonlinear SVC ⋆ inner product representation ⋆ kernel ridge regression ⋆ kernelized LASSO ⋆ (linear and nonlinear) support vector regression ⋆ reproducing kernel Hilbert spaces (RKHS) ⋆ Riesz-Fréchet representation theorem ⋆ reproducing property ⋆ typical kernels ⋆ Moore-Aronszajn theorem ⋆ representer theorem ⋆ McDiarmid's inequality ⋆ uniformly stable estimators ⋆ uniform convergence bounds ⋆ misclassification bounds ⋆ nearest centroid classifier ⋆ kernel mean embedding of distributions ⋆ universal and characteristic kernels ⋆ famous embeddings: moment generating and characteristic functions ⋆ induced metric on probability distributions ⋆ empirical estimation of mean embeddings ⋆ generalized strong law of large numbers with error bounds ⋆ weak convergence to Gaussian processes

[Reminder] Markov Chains

discrete (countable) Markov chains ⋆ transition kernels ⋆ initial distribution ⋆ Chapman-Kolmogorov equations ⋆ communicating classes ⋆ closed and absorbing classes ⋆ recurrence and transience ⋆ passage times ⋆ expected return times ⋆ irreducibility and aperiodicity ⋆ invariant distributions ⋆ existence of and convergence to the stationary distribution ⋆ positive and null recurrence ⋆ ergodic theorem ⋆ Poisson equation

Markov Decision Processes (MDPs)

equivalent definitions of MDPs ⋆ control policies ⋆ sufficiency of Markov policies ⋆ value functions ⋆ partial ordering of policies ⋆ state augmentation ⋆ finite horizon problems ⋆ stochastic shortest path problems ⋆ discounted problems ⋆ average cost problems ⋆ Bellman operators ⋆ optimality equation ⋆ dynamic programming principle ⋆ famous examples: asset selling, inventory control and linear-quadratic regulator ⋆ underlying contractions ⋆ monotonicity ⋆ constant shifts ⋆ value iteration (and variants: asynchronous, approximate, Gauss-Seidel, relative) ⋆ policy iteration (and variants: approximate, optimistic, and generalized) ⋆ error bounds ⋆ linear programmig formulation ⋆ Blackwell optimality ⋆ unbounded costs ⋆ partial observability

Reinforcement Learning (RL)

model-free solutions to MDPs ⋆ actor-critic methods ⋆ Monte Carlo policy evaluations ⋆ temporal difference (TD) learning ⋆ first-visit, every-visit, online, offline TD variants ⋆ TD(λ) ⋆ strong consistency of TD ⋆ action-value functions ⋆ SARSA ⋆ Bellman equation for Q-factors ⋆ underlying contractions ⋆ Q-learning ⋆ strong consistency ⋆ exploration vs exploitation ⋆ stochastic bandits ⋆ pseudo-regret ⋆ sub-Gaussian distributions ⋆ concentration bounds ⋆ explore-then-commit algorithm ⋆ optimism principle ⋆ UCB algorithm

Stochastic Approximation (SA)

adaptive algorithms ⋆ fixed point and root finding problems ⋆ Robbins-Monro algorithm ⋆ Kiefer-Wolfowitz algorithm ⋆ policy gradient ⋆ SPSA (simultaneous perturbation stochastic approximation) ⋆ stochastic gradient descent (SGD) ⋆ momentum acceleration ⋆ Polyak averaging ⋆ asymptotic analysis for martingale difference noises ⋆ consistency based on Lyapunov functions ⋆ examples: SGD with Lipschitz continuous gradient and Euclidean norm pseudo-contractions ⋆ consistency based on contraction and monotonicity properties

Recommended Literature

- Vapnik, V. N.: *Statistical Learning Theory*, John Wiley & Sons, 1998.
- Schölkopf, B., & Smola, A. J.: *Learning with Kernels*: Support Vector Machines, Regularization, Optimization and Beyond, The MIT Press, 2002.
- Bertsekas, D. P., & Tsitsiklis, J. N.: *Neuro-Dynamic Programming*, Athena Scientific, 1996.
- Bertsekas, D. P.: *Dynamic Programming and Optimal Control*, 4th ed., Athena Scientific, 2017.
- Lattimore, T., & Szepesvári, Cs.: *Bandit Algorithms*, Cambridge University Press, 2018.
- Friedman, J. H., Tibshirani, R., & Hastie, T.: *Elements of Statistical Learning*, 2nd edition, Springer Science & Business Media, 2008.
- Manton, J. H., & Amblard, P. O.: *A Primer on Reproducing Kernel Hilbert Spaces*, Foundations and Trends in Signal Processing, Now Publisher, 2015.
- Muandet, K., Fukumizu, K., Sriperumbudur, B., & Schölkopf, B.: *Kernel Mean Embedding of Distributions*: A Review and Beyond, Foundations and Trends in Machine Learning, Now Publisher, 2017.

Background Material for the Reminders

- DeGroot, M. H., & Schervish, M. J.: *Probability and Statistics*, 4th ed. Pearson Education, 2012.
- Boyd, S., & Vandenberghe, L.: *Convex Optimization*, Cambridge University Press, 2004.
- Norris, J. R.: *Markov Chains*, Cambridge University Press, 1998.
- Lax, P. D.: *Functional Analysis*, John Wiley & Sons, 2002.