Fondements théoriques du deep learning

Sébastien Gerchinovitz, François Malgouyres, Édouard Pauwels, Nicolas Thome

Objectifs :

L'objectif principal de ce cours est de présenter la formulation mathématique des réseaux de neurones profonds, dans le cadre de leur utilisation pour la classification et la régression. Nous décrirons les différents problèmes, ainsi que des résultats mathématiques représentatifs de la recherche actuelle, concernant l'optimisation des paramètres du réseaux, les propriétés des réseaux de neurones, leur expressivité, leur complexité ainsi que des phénomènes encore inexpliqués comme la double descente et la régularisation implicite. Enfin nous ferons le lien avec la mise en pratique de ces algorithmes, dans un objectif de robustesse (incertitude, stabilité).

Les résultats mathématiques seront complétés par des travaux pratiques permettant aux étudiants de comprendre intuitivement certaines propriétés des réseaux profonds et de mettre en oeuvre des algorithmes d'apprentissage pour résoudre des problèmes concrets. L'objectif est que les étudiants fassent le lien entre les différents formalismes et situent les propriétés mathématiques vues dans la première partie du cours dans leur contexte applicatif.

Équipe enseignante:

- Sébastien Gerchinovitz, chercheur à l'IRT Saint-Exupéry.
  `https://www.math.univ-toulouse.fr/~sgerchin/`
- François Malgouyres, Pr à l'Institut de Mathématiques de Toulouse.
  `https://www.math.univ-toulouse.fr/~fmalgouy/`
- Édouard Pauwels, McF à l'Institut de Recherche en Informatique de Toulouse.
  `https://www.irit.fr/~Edouard.Pauwels/`
- Nicolas Thome, Pr à Sorbonne Université, laboratoire ISIR.
  `https://thome.isir.upmc.fr/`

Calendrier:

- Francois Malgouyres: jeudi 05/10 (14h-17h, visio), jeudi 12/10 (14h-17h, ENS), vendredi 13/10 (9h-12h, ENS)
- Edouard Pauwels : Vendredi 20/10 (9h-12h, visio)
- Nicolas Thome: jeudi 26/10 (14h-17h, ENS), jeudi 9/11 (14h-17h, ENS), jeudi 16/11 (14h-17h, ENS)
- Sébastien Gerchinovitz: jeudi 23/11 (14h-17h, ENS), vendredi 24/11 (9h-12h, ENS), vendredi 1/12 (9h-12h, visio)

Evaluation: TP : 30 %, Examen : 70%

Programme: Chacune des séances ci-dessous dure $3h$.

Séance 1: F. Malgouyres – Introduction

- Généralités sur la classification et la régression
- Erreur de généralisation, d'approximation et d'optimisation. Régularisation implicite et double descente.
- Réseaux feed-forward et convolutionels. Aperçu de la zoologie: transformers, RNN, GAN, FairGAN, apprentissage de représentation...
- Une prédiction affine par morceaux en fonction de X

Séance 2: F. Malgouyres – La prédiction, comme une fonction des paramètres,

- Une prédiction polynomiale par morceaux en fonction de θ
- B-A-BA des propriétés de la fonction coût (non-convexité, non-coercivité)
- Back-propagation: Explosion/disparition du gradient

Séance 3: F. Malgouyres – Paysage de la fonction objectif

- Les propriétés du paysage et leur impact sur l'optimisation [3]
- Paysage dans le cas linéaire, régularisation implicite [4]
- Cas non-linéaire: Bonnes propriétés de la fonction coût pour les réseaux profonds larges [25]

Séance 4: E. Pauwels – Optimisation non-convexe

- Descente de gradient et convergence vers des minima locaux pour les fonctions de Morse [3].
- Propriétés algébriques et géométriques des fonctions de coût en deep learning [20,21].
- Méthodes de gradient sous hypothèse KL [1].
- Application à la convergence globale, gradient stochastique, moindre carrés sous hypothèse de surjectivité.

Séance 5: N. Thome – Deep Learning robuste 1 (cours puis TP)

- Contexte de la robustesse décisionnelle, enjeux théoriques et appliqués pour le systèmes critiques (healthcare, autonomous driving)
- Stabilité [12,13,14]
- Uncertitude décisionnelle: typologie, présentation des modèles Bayésiens (postérieure et distribution prédictive) [15, 16]

- Régression linéaire Bayésienne, extension à la régression non-linéaires (plongements explicites)
- TP: régression linéaire Bayésienne sur données simulées

Séance 6: N. Thome – Deep Learning robuste 2 (cours puis TP)

Réseaux de neurones Bayésiens :

- Régression logistique : approximation de Laplace
- Réseaux de neurone Bayésiens : approximations variationnelles [17]
- Monte Carlo Dropout [18]: lien entre la régularisation et l'inférence variationnelle
- TP: approximation de Laplace pour la classification, mise en place d'une couche LinearVariationelle en `PyTorch`, application à des modèles linéaires et à des percpetron multi-couches, comparaison avec MCDroupout

Séance 7: N. Thome – Deep Learning robuste 3 (cours puis TP)

- Métriques d'incertitudes pour la régression
- Ensembling, modèles évidentiels
- Incertitude décisionnelle en pratique
  - Calibration [19], prédiction d'échec [20]
  - Détection de points hors distribution (OOD) [21]
- TP: analyse de l'incertitude avec un réseau de neurones convolutif pour la classification sur MNIST, détection d'échec (ConfidNet), détection d'OOD (ODIN)

Séance 8: S. Gerchinovitz – Expressivité des réseaux feed-forward

- Un théorème d'approximation universelle [26]
- Exemples de résultats plus quantitatifs : un résultat fondateur en norme L^2 [8], phénomène de 'depth separation' [9], super-approximation par des réseaux profonds [27].

Séance 9: S. Gerchinovitz – Erreur de généralisation des réseaux feed-forward

- Rappels statistiques en classification binaire : bornes de risque, VC dimension, complexité de Rademacher
- Quelques résultats de complexité : VC dimension de réseaux feed-forward [28], complexité de Rademacher de réseaux feed-forward [10]

Séance 10: S. Gerchinovitz – Questions ouvertes sur la généralisation en deep learning

- Un rapide tour d'horizon de questions et pistes de recherche [11]

- Le phénomène de double descente [29,30]

References:

1. Bolte, Sabach, Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems", Math. Prog. Vol 146, no 1-2, pp 59-594, 2014.
2. Panageas, Piliouras, "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions", ITCS 2017
3. Lee, Simchowitz, Jordan, Recht, "Gradient descent only converges to minimizers", COLT 2016.
4. Baldi, Hornik, "Neural networks and principal component analysis: Learning from examples without local minima", Neural networks, 2(1):53–58, 1989.
5. Venturi, Bandeira, Bruna, "Neural Networks with Finite Intrinsic Dimension have no Spurious Valleys", ArXiv, February 2018.
6. Safran, Shamir, "On the quality of the initial basin in overspecified neural networks", ICML, pages 774–782, 2016.
7. Malgouyres, Landsberg, "Multilinear compressed sensing: application to deep linear networks", ArXiv, 2017.
8. Barron, "Universal approximation bounds for superpositions of a sigmoidal function", IEEE Transactions on Information Theory:39(3), pages 930–945, 1994.
9. Eldan, Shamir, "The power of depth for feedforward neural networks", 29th Annual Conference on Learning Theory, PMLR 49:907-940, 2016.
10. Golowich, Rakhlin, Shamir, "Size-Independent Sample Complexity of Neural Networks", Proceedings of COLT'18, 297-299, 2018.
11. Kawaguchi, Kaelbling, Bengio, "Generalization in Deep Learning". In "Mathematical Aspects of Deep Learning" (pp. 112-148), Cambridge University Press, 2022.
12. "Invariant scattering convolution networks", J. Bruna and S. Mallat. IEEE T-PAMI.,35(8):1872–1886, 2013.
13. "Group Invariance and Stability to Deformations of Deep Convolutional Representations", A. Bietti, J. Mairal. In NIPS, 2017.
14. "Understanding black-box predictions via influence functions", Pang Wei Koh and Percy Liang. In ICML, 2017.
15. "Pattern Recognition and Machine Learning", Bishop, C. M., Springer-Verlag, Berlin, Heidelberg, 2006.

16. "Machine Learning: A Probabilistic Perspective", Murphy, K. P., The MIT Press, 2012

17. "Weight uncertainty in neural network", Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. ICML 2015

18. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning", Yarin Gal, Zoubin Ghahramani. Proceedings of the 33th International Conference on Machine Learning (ICML'16).

19. "On calibration of modern neural networks", Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. ICML 2017

20. "Confidence estimation via auxiliary models", C Corbière, N Thome, A Saporta, TH Vu, M Cord, P Pérez. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021

21. "Enhancing the reliability of out-of-distribution image detection in neural networks", Liang, S., Li, Y., and Srikant, R. ICLR 2018

22. Kurdyka, K. (1998, January). On gradients of functions definable in o-minimal structures. In Annales de l'institut Fourier (Vol. 48, No. 3, pp. 769-784). Chartres: L'Institut, 1950-.

23. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M. (2007). Clarke subgradients of stratifiable functions. SIAM Journal on Optimization, 18(2), 556-572.

24. Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J. D. (2018). Stochastic subgradient method converges on tame functions. arXiv preprint arXiv:1804.07795.

25. Nguyen, Q. and Hein, M., "The loss surface of deep and wide neural networks", arXiv preprint arXiv:1704.08045, 2017.

26. Cybenko, "Approximation by superpositions of a sigmoidal function", Math. Control Signal Systems, 2, 303-314, 1989.

27. Yarotsky, "Optimal approximation of continuous functions by very deep ReLU networks", Proceedings of COLT'18, 639-649, 2018.

28. Bartlett, Harvey, Liaw, Mehrabian, "Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks", 20(63):1-17, 2019.

29. Belkin, Hsu, Ma, Mandal, "Reconciling modern machine-learning practice and the classical bias-variance trade-off", Proceedings of the National Academy of Sciences, Volume 116, Issue 32, p.15849-15854, 2019.

30. Hastie, Montanari, Rosset, Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation", The Annals of Statistics,50:949-986, 2022.