Analyse des transitions de phase en théorie statistique de la discrimination Sujet de thèse 2021

Yann Guermeur[1], Fabien Lauer[1]
[1] ABC Team, Université de Lorraine, CNRS, LORIA Nancy
(firstname.lastname@loria.fr)

22 avril 2021

1 Exposé scientifique du projet

1.1 Etat de l'art

L'émergence récente de nouveaux systèmes discriminants, comme les multiples succès obtenus par les modèles de l'apprentissage profond [8], mettent sous tension les résultats classiques de la théorie statistique de l'apprentissage [28]. Le phénomène vertueux de convergence entre théorie et pratique, initié par Vapnik et ses collaborateurs au milieu des années 90 avec l'introduction des méthodes à noyau [26], a connu un coup d'arrêt marqué. A présent, les arguments classiquement mis en avant pour encadrer la pratique ne portent plus, et l'utilisation par le praticien de modèles fortement sur-paramétrés tend à se généraliser, ce qui ne manque pas de poser de graves problèmes, d'ailleurs bien identifiés par la littérature [27]. Pour tenter de reprendre l'initiative, la théorie emprunte principalement deux voies, les contributions généralistes [9, 25, 19, 22, 11] alternant avec des contributions dédiées à des classifieurs spécifiques, comme la méthode des plus proches voisins [18] ou les machines à vecteurs support multi-classes (M-SVM) [6, 29]. Récemment, une communication de Bartlett et co-auteurs [2] a véritablement lancé la théorie statistique des réseaux de neurones profonds. Depuis, les contributions se sont multipliées (voir [20] pour un état de l'art). Cependant, le sujet demeure encore largement ouvert. Plus généralement, la théorie statistique des systèmes discriminants multi-classes à marge, fondée il y a près de vingt ans [17] avec les outils de l'époque pour l'étude des classifieurs de l'époque, doit s'adapter aux nouveaux défis qui lui sont soumis, dans le cadre d'une démarche synthétique intégrant les diverses contributions déjà disponibles.

1.2 Objectifs principaux

L'objectif de cette thèse est de faire progresser la théorie statistique de la discrimination à catégories multiples à travers la dérivation de nouvelles bornes sur la probabilité d'erreur, encore nommées *risques garantis*. Ces bornes permettront de munir les principales familles de classifieurs de la littérature d'une topologie. Il s'agit plus concrètement d'établir des risques garantis pour lesquels la dépendance de l'intervalle de confiance aux trois paramètres fondamentaux : la taille m de l'échantillon, le nombre C de catégories et le paramètre de marge γ, présente une amélioration par rapport à l'état de l'art. A notre connaissance, cette prise en

compte globale des trois paramètres est récente [12], la littérature s'étant jusqu'alors concentrée sur la dépendance aux deux premiers. L'originalité de notre démarche réside dans le point de vue adopté : celui des *transitions de phase*, introduit par Mendelson [23, 24]. Il s'agit de concentrer l'étude sur les petites variations des configurations des classifieurs induisant un changement de dépendance à l'un des paramètres fondamentaux. C'est précisément sur ces variations et leurs conséquences que s'appuiera notre topologie. Ainsi, les classifieurs considérés trouveront leur place dans l'espace à trois dimensions engendré par les paramètres fondamentaux.

1.3 Structuration en tâches

La réalisation des objectifs annoncés doit s'appuyer sur plusieurs tâches, qui trouvent leur inspiration dans nos travaux antérieurs. Nous les exposons à présent, en commençant par celles qui constituent une poursuite directe de nos recherches passées ou actuelles, pour finir par celles qui font intervenir les verrous scientifiques les plus importants.

1.3.1 Performances en généralisation de la machine à noyau hyperbolique

La première tâche programmée doit aborder le phénomène de transition de phase à travers l'étude d'une nouvelle machine à noyau que nous avons introduite dans [5] : la machine à noyau hyperbolique (HKM). Deux outils classiques de l'inférence empirique ont inspiré sa conception : le classifieur du plus proche centroïde (NCC) et la sphère englobante de rayon minimal. Elle engendre dans l'espace de Hilbert à noyau autoreproduisant (RKHS) associé au noyau des frontières de décision quadratiques. A notre connaissance, une seule autre machine à noyau est non linéaire/quadratique dans le RKHS : l'extension « kernélisée » de l'analyse discriminante (quadratique) de Fisher (KFD).

D'un point de vue purement paramétrique, la capacité de la HKM se situe entre celles du NCC et de la KFD. L'étude conjointe des performances en généralisation de ces trois classifieurs est ainsi susceptible de faire apparaître des transitions de phase. Pour la conduire, nous améliorerons et étendrons les majorants déjà obtenus de la complexité de Rademacher [3] de la HKM. Ces développements s'appuieront sur l'utilisation de nouveaux résultats de la théorie des processus empiriques.

1.3.2 Performances en généralisation des réseaux de neurones profonds

Notre principale contribution à la théorie des systèmes discriminants multi-classes à marge est l'introduction de mesures de capacité combinatoires dédiées : les γ-Ψ-dimensions [9]. C'est précisément l'emploi d'un ancêtre de ces mesures, la γ-dimension [15], qui avait permis à Mendelson de mettre en évidence le phénomène de transition de phase (voir le théorème 18 de [24]). Les bornes les plus récentes impliquant des γ-Ψ-dimensions [12] s'appliquent à des classifieurs définis de manière très abstraite, à travers les seules propriétés statistiques des classes de fonctions sous-jacentes. Ces classes sont simplement supposées être Glivenko-Cantelli uniformes (uGC) [7]. Cependant, la dédication de nos bornes à une famille de classifieurs particuliers, comme les M-SVM, permet d'obtenir des garanties plus fines. Il apparaît donc intéressant de les appliquer aux réseaux de neurones, et comparer les résultats ainsi obtenus à ceux de l'état de l'art [20]. Au préalable, le principal facteur limitant la portée de cette application, la trop forte dépendance au paramètre d'échelle du résultat combinatoire dédié au couple (norme L_2, dimension de Natarajan à marge), aura été éliminé par une mise en œvre plus appropriée du principe de petite déviation.

Il est connu de longue date qu'un moyen simple d'observer d'importantes transitions de phase chez les réseaux de neurones à propagation avant [1] consiste à modifier la fonction d'activation (voir par exemple [16]). L'étude comparative évoquée au paragraphe précédent sera donc conduite en tenant compte de ce phénomène. Du point de vue de la topologie sur l'espace des classifieurs à marge, il est fondamental de conserver à l'esprit le fait qu'il est non seulement possible de comparer les réseaux de neurones entre eux, mais encore de les comparer à des machines à noyau. Pour ce faire, nous utiliserons comme point de départ la contribution très originale de Jacot et co-auteurs [14].

1.3.3 Propriétés statistiques des M-SVM à coût quadratique

Les machines à vecteurs support à coût quadratique [13, 10] ont été conçues de manière à disposer d'une borne « rayon-marge » sur l'erreur de validation croisée *leave-one-out* [4]. Cela simplifie la mise en œuvre de la sélection de modèle par parcours du chemin de régularisation. Pour ces machines, différentes questions théoriques demeurent ouvertes, comme la Fisher consistance [21]. Dans le cas particulier des M-SVM, un prolongement naturel des travaux précédemment décrits consiste à enrichir la topologie en y incorporant des « dimensions » supplémentaires, comme la Fisher consistance. Il est important de souligner le fait que l'absence d'une expression analytique de la fonction de perte des M-SVM à coût quadratique représente un verrou scientifique majeur.

Références

[1] M. Anthony and P.L. Bartlett. *Neural Network Learning : Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.

[2] P.L. Bartlett, D.J. Foster, and M. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *NIPS 31*, pages 6241–6250, 2017.

[3] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities : Risk bounds and structural results. *Journal of Machine Learning Research*, 3 :463–482, 2002.

[4] R. Bonidal, S. Tindel, and Y. Guermeur. Model selection for the ℓ_2-SVM by following the regularization path. *Transactions on Computational Collective Intelligence*, XIII :83–112, 2014.

[5] A. El Dakdouki, Y. Guermeur, and N. Wicker. Hyperbolic kernel machine. 2020. (in revision).

[6] U. Doğan, T. Glasmachers, and C. Igel. A unified view on multi-class support vector classification. *Journal of Machine Learning Research*, 17(45) :1–32, 2016.

[7] R.M. Dudley, E. Giné, and J. Zinn. Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability*, 4(3) :485–510, 1991.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, Cambridge, MA, 2016.

[9] Y. Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8 :2551–2594, 2007.

[10] Y. Guermeur. A generic model of multi-class support vector machine. *International Journal of Intelligent Information and Database Systems*, 6(6) :555–577, 2012.

[11] Y. Guermeur. L_p-norm Sauer-Shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89 :450–473, 2017.

[12] Y. Guermeur. Combinatorial and structural results for γ-ψ-dimensions. Technical report, arXiv:1809.07310v3, 2020.

[13] Y. Guermeur and E. Monfrini. A quadratic loss multi-class SVM for which a radius-margin bound applies. *Informatica*, 22(1) :73–96, 2011.

[14] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel : convergence and generalization in neural networks. In *NIPS 32*, 2018.

[15] M.J. Kearns and R.E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3) :464–497, 1994.

[16] P. Koiran and E. Sontag. Neural networks with quadratic VC dimension. *Journal of Computer and System Sciences*, 54(1) :190–198, 1997.

[17] V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1) :1–50, 2002.

[18] A. Kontorovich and R. Weiss. Maximum margin muliclass nearest neighbors. In *ICML'14*, 2014.

[19] V. Kuznetsov, M. Mohri, and U. Syed. Multi-class deep boosting. In *NIPS 27*, pages 2501–2509, 2014.

[20] A. Ledent, Y. Lei, and M. Kloft. Norm-based generalisation bounds for multi-class convolutional neural networks. Technical report, arXiv:1905.12430v3, 2020.

[21] Y. Liu. Fisher consistency of multicategory support vector machines. In *AISTATS 11*, pages 289–296, 2007.

[22] A. Maurer. A vector-contraction inequality for Rademacher complexities. In *ALT'16*, pages 3–17, 2016.

[23] S. Mendelson. Rademacher averages and phase transitions in Glivenko-Cantelli classes. *IEEE Transactions on Information Theory*, 48(1) :251–263, 2002.

[24] S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A.J. Smola, editors, *Advanced Lectures on Machine Learning*, chapter 1, pages 1–40. Springer-Verlag, Berlin, Heidelberg, New York, 2003.

[25] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, 2012.

[26] B. Schölkopf and A.J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, Cambridge, MA, 2002.

[27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *ICLR'14*, 2014.

[28] V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.

[29] Y.Lei, U. Doğan, D.-X. Zhou, and M. Kloft. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5) :2995–3021, 2019.