Machine Learning

Massih-Reza Amini

Université Grenoble Alpes, Grenoble INP, CNRS
Laboratoire d'Informatique de Grenoble

Abstract

Cet ouvrage présente les fondements scientifiques de la théorie de l'apprentissage supervisé, les algorithmes les plus répandus développés suivant ce domaine ainsi que les deux cadres de l'apprentissage semi-supervisé et de l'ordonnancement, à un niveau accessible aux étudiants de master et aux élèves ingénieurs.

La première édition, connue sous le nom Apprentissage machine, fut traduite en chinois par les éditions iTuring. Dans cette deuxième édition, un nouveau chapitre est dédié au Deep Learning, sur les réseaux de neurones artificiels, et nous avons réorganisé les autres chapitres pour un exposé cohérent reliant la théorie aux algorithmes développés dans cette sphère.

Vous trouverez également dans cette édition quelques programmes des algorithmes classiques, écrits en langages Python et C (langages à la fois simples et populaires), et à destination des lecteurs qui souhaitent connaître le fonctionnement de ces modèles désignés parfois comme des boites noires. Ces programmes libres (GPLv3) essentiels au développement de solutions big data sont déposés progressivement sur ce gitlab :
https://gricad- gitlab.univ-grenoble-alpes.fr/aminima/machine-learning-tools

À qui s'adresse ce livre ?

- Aux élèves ingénieurs, étudiants de master et doctorants en mathématiques appliquées, algorithmique, recherche opérationnelle, gestion de production, aide à la décision.
- Aux ingénieurs, enseignants-chercheurs, informaticiens, industriels, économistes et décideurs ayant à résoudre des problèmes de classification, de partitionnement et d'ordonnancement à large échelle.

Préface

Depuis quelques années, dans les domaines scientifiques, industriels et personnels, la presence de données numériques et leur utilisation ont explosé. Certaines de ces données sont massives, nécessitent des outils et architectures spécifiques, comme en astronomie ou pour les moteurs de recherche, et constituent les problèmes dits de «big data».

D'autres données ne sont pas si massives, comme les photos ou vidéos familiales, mais constituent toujours un défi algorithmique. Le grand changement récent est non seulement la taille, mais aussi le côté omniprésent de ces données, qui sont utilisées quotidiennement.

Depuis une vingtaine d'années, l'apprentissage statistique («machine learning» en anglais) s'est considérablement développé, à l'interface entre l'informatique et les statistiques, et constitue le cœur méthodologique des algorithmes modernes de traitement de données. Même si les recherches en apprentissage sont toujours en plein essor, un socle méthodologique et algorithmique a émergé.

Ce livre constitue une introduction équilibrée aux concepts et outils les plus importants de l'apprentissage supervisé et de ses extensions. Un accent remarquable est mis sur des résultats théoriques élégants, simples mais puissants, des algorithmes efficaces qui ont montré leurs preuves en pratique, et des codes permettant de reproduire les expériences.

Francis Bach

octobre 2014

Table des matières

Avant-propos

L'apprentissage machine est l'un des domaines phares de l'intelligence artificielle. Il concerne l'étude et le développement de modèles quantitatifs permettant à un ordinateur d'accomplir des tâches sans qu'il soit explicitement programmé à les faire. Apprendre dans ce contexte revient à reconnaître des formes complexes et à prendre des décisions intelligentes. Compte tenu de toutes les entrées existantes, la difficulté d'accomplir cette tâche réside dans le fait que l'ensemble des décisions possibles est généralement très complexe à énumérer. Pour contourner cette difficulté, les algorithmes en apprentissage machine ont été conçus dans le but d'acquérir de la connaissance sur le problème à traiter en se basant sur un ensemble de données limitées issues de ce problème.

Concepts étudiés

Pour illustrer ce principe, considérons le cadre de l'apprentissage supervisé que nous allons en partie traiter dans cet ouvrage. Suivant ce cadre, la décision à prendre sur une entrée donnée est prise d'après la sortie d'une fonction de prédiction qui est inférée en utilisant un ensemble d'exemples étiquetés (ou base d'entraînement), où chacun de ces exemples est une paire constituée du vecteur représentatif d'une observation dans un espace vectoriel donné, et d'une réponse associée à l'exemple (aussi appelée sortie désirée ou sortie réelle). Après la phase d'estimation ou d'apprentissage, la fonction renvoyée par l'algorithme doit permettre de prédire la réponse associée à de nouvelles observations. L'hypothèse sous-jacente dans ce cas est que les exemples sont, d'une façon générale, représentatifs du problème de prédiction sur lequel la fonction sera appliquée. En pratique, une fonction d'erreur mesure l'écart entre la prédiction du modèle sur un exemple et sa sortie désirée. À partir d'un ensemble d'entraînement donné, l'algorithme d'apprentissage choisit alors une fonction, issue d'un ensemble de fonctions défini au préalable, qui réalise l'erreur moyenne la plus faible sur les exemples de la base d'entraînement.

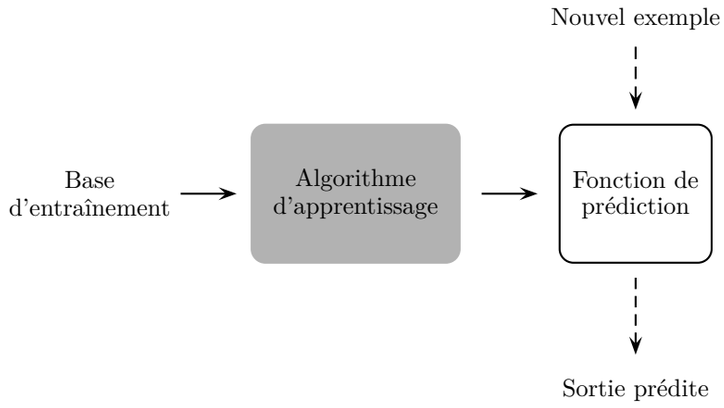

Figure I - Illustration des deux phases d'un problème d'apprentissage supervisé. Dans la phase d'apprentissage (schématisée par les traits pleins), une fonction minimisant l'erreur empirique sur une base d'entraînement est trouvée parmi une classe de fonctions prédéfinies. Dans la phase de test (schématisée par les traits pointillés), les sorties de nouveaux exemples sont prédites par la fonction de prédiction.

Cette erreur n'est généralement pas représentative de la performance de l'algorithme sur de nouveaux exemples. Il est alors nécessaire de disposer d'un second ensemble d'exemples étiquetés, ou une base de test, auxquels l'algorithme n'avait pas accès et d'estimer l'erreur moyenne de la fonction produite lors de la phase d'estimation qui sera cette fois représentative de son erreur de généralisation. On attend de l'algorithme d'apprentissage qu'il trouve une fonction ayant de bonnes performances en généralisation et non celle qui sera capable de reproduire parfaitement les réponses associées aux exemples d'entraînement (voir figure I). Les garanties d'apprenabilité du procédé de la minimisation du risque empirique ont été étudiées dans la théorie de l'apprentissage machine largement initiée par (1). Elles dépendent de la taille de la base d'entraînement et de la complexité de la classe de fonctions où on cherche la fonction de prédiction.

Historiquement, les deux tâches principales du cadre de l'apprentissage supervisé étaient la classification et la régression. Ces tâches sont similaires à la différence de l'espace des sorties désirées des exemples. Dans le cas de la classification, l'espace de sortie est discret alors qu'en régression cet espace est réel.

À la fin des années 1990 et sous l'impulsion de nouvelles technologies, notamment celles liées au développement d'Internet, de nouveaux cadres d'apprentissage ont vu le jour. Un de ces cadres est l'apprentissage avec des données partiellement

étiquetées, ou l'apprentissage semi-supervisé, dont le développement est motivé par l'effort qu'il faut consentir à constituer des bases d'apprentissage étiquetées et le constat que les données étiquetées sont chères à obtenir alors que les données non étiquetées sont foison et qu'elles contiennent de l'information sur le problème que l'on cherche à résoudre. De ce constat sont nés plusieurs travaux qui avaient pour objectif d'employer une petite quantité de données étiquetées, simultanément avec une grande quantité de données non étiquetées, pour apprendre une fonction de prédiction.

L'autre cadre qui a suscité de nombreux travaux dans la communauté d'apprentissage depuis les années 2000 concerne le développement de modèles d'ordonnancement. Ce cadre a formalisé dans un premier temps les problèmes de la Recherche d'Information et il a été par la suite étendu à d'autres problèmes plus généraux.

Depuis de nombreuses années, les algorithmes d'apprentissage développés suivant ces cadres ont été appliqués avec succès à une grande variété de problèmes, incluant la reconnaissance de la parole et de l'écriture manuscrite, la vision par ordinateur, la prédiction de la structure des protéines, les systèmes de recommendations, la classification documentaire, les moteurs de recherche, etc.

Organisation du livre

Cet ouvrage présente les fondements scientifiques de la théorie de l'apprentissage supervisé, les algorithmes les plus répandus développés suivant ce cadre ainsi que les deux cadres de l'apprentissage évoqués plus haut, à un niveau accessible aux étudiants de master et aux élèves ingénieurs. Notre souci a été de proposer un exposé cohérent reliant la théorie aux algorithmes développés dans ce domaine. En outre, cette étude ne se limite pas à l'exposé de ces fondements, mais présente aussi quelques programmes des algorithmes classiques proposés dans ce manuscrit, écrits dans un langage informatique simple et populaire qui est le langage C [1], et à destination des lecteurs qui cherchent à connaître le fonctionnement de ces modèles désignés parfois comme des boîtes noires. Ce livre est organisé en six chapitres principaux et deux annexes. L'enchaînement des idées présentées dans chacun d'eux est le suivant :

- Dans le chapitre 1, nous décrivons les concepts fondamentaux de la théorie de l'apprentissage statistique de (1). Nous exposons la notion de consistance du principe de la minimisation du risque empirique selon lequel la plupart des algorithmes en apprentissage supervisé ont été développés. L'étude de cette consistance nous mènera à l'exposé du second principe fondamental en apprentissage qui est la minimisation du risque structurel, ouvrant le champ au développement de nouveaux modèles en apprentissage machine. En particulier,

1. http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html

nous présentons dans ce chapitre la notion de borne sur l'erreur de généralisation en décrivant les hypothèses et les outils nécessaires pour l'obtenir.

- Dans le chapitre 2, nous allons présenter des bornes sur l'erreur de généralisation qui peuvent être estimées sur une base d'entraînement qui sert à apprendre un modèle. Ces bornes sont basées sur une notion de complexité de classes dépendante des données, appelée complexité de Rademacher. Avec cette notion il est aussi possible de dériver facilement des bornes sur l'erreur de généralisation pour la classification multi-classe. Nous exposons les approches multi-classe basées sur la classification binaire, appelées approches combinées, et dérivons une borne sur l'erreur de généralisation des classifieurs linéaires dans ce cas.
- C'est dans le chapitre 3 que nous nous intéressons aux algorithmes de base issus du domaine de l'optimisation pour la minimisation d'une fonction de risque convexe majorant le risque empirique appelés *algorithmes à direction de descente*. En particulier, nous présentons les conditions nécessaires à vérifier un algorithme à direction de descente pour converger vers le minimiseur d'une fonction objectif convexe et nous décrivons quelques variantes simples et efficaces de cet algorithme.
- Le chapitre 4 présente les principaux modèles qui sont issus des travaux sur la modélisation numérique des neurones du système nerveux, appelés plus communément réseaux de neurones artificiels ou *Deep Learning*. Ces modèles étaient précurseurs au développement des méthodes quantitatives en Intelligence Artificielle et un soin particulier a été porté à la présentation de ces modèles dans leur contexte historique.
- Dans le chapitre 5, nous présentons les séparateurs à vaste marge (SVM) qui sont issus du principe de la minimisation du risque structurel. Ces modèles sont devenus très populaires grâce à leurs justifications théoriques. En particulier, nous verrons comment utiliser l'astuce de noyau pour plonger l'espace d'entrée dans un espace de plus grande dimension dans lequel le problème d'apprentissage devient plus simple à résoudre et nous présentons l'extension de ces au cas multi-classe.
- Le chapitre 6 présente l'algorithme Adaboost. Cet algorithme combine plusieurs classifieurs de base, appelés apprenants faibles, pour construire un classifieur final, appelé apprenant fort, qui est plus performant que chacun de ces classifieurs de base. En particulier, nous ferons le lien entre cet algorithme et le principe de la minimisation du risque empirique énoncé par (1). Nous exposons en outre l'extension de cet algorithme au cas multi-classe.
- Nous exposons ensuite le cadre de l'apprentissage semi-supervisé dans le chapitre 7. Nous commençons ce chapitre par présenter les algorithmes EM et CEM développés dans le cadre de l'apprentissage non supervisé, en détaillant quelques cas particuliers menant à des modèles non supervisés bien connus comme l'algorithme des k-moyennes. Nous présentons ensuite les hypothèses

de base en apprentissage semi-supervisé en détaillant les trois approches génératives, discriminantes et graphiques développées suivant ce cadre.

- C'est dans le chapitre 8 que nous décrivons formellement le cadre de l'apprentissage de fonctions d'ordonnancement (ou learning to rank en anglais) en focalisant sur deux formes particulières d'ordonnancement appelées ordonnancement d'alternatives et d'ordonnancement d'instances. Nous exposons ensuite quelques algorithmes développés suivant les approches classiques de l'apprentissage de fonctions d'ordonnancement. Nous terminons ce chapitre par montrer la réduction de quelques problèmes d'ordonnancement à la classification binaire de paires d'observations. Cette réduction ouvre la voie à l'apprentissage de classifieurs avec des exemples interdépendants que nous analysons avec le résultat de (2).
- En annexe A, nous donnons quelques rappels des outils de base en probabilité que nous employons dans cet ouvrage.
- En annexe B, nous donnons les codes programmes de quinze algorithmes présentés dans les différents chapitres, en détaillant les structures de données utilisées et en liant les différentes parties des programmes aux points correspondants abordés dans cet ouvrage.

Bibliographie

[1] V. N. Vapnik, *The nature of statistical learning theory (second edition)*. Springer-Verlag, 1999.

[2] S. Janson, "Large deviations for sums of partly dependent random variables," *Random Structures and Algorithms*, vol. 24, no. 3, pp. 234–248, 2004.

[3] M. Amini and N. Usunier, "A contextual query expansion approach by term clustering for robust text summarization," in *Proceedings of the 7th Document Understanding Conference*, (Rocherster - USA), 2007.

[4] J.-F. Pessiot, Y.-M. Kim, M. Amini, and P. Gallinari, "Improving document clustering in a learned concept space," *Information Processing & Management*, vol. 46, no. 2, pp. 180–192, 2010.

[5] M. Amini, N. Usunier, and F. Laviolette, "A transductive bound for the voted classifier with an application to semi-supervised learning," in *Advances in Neural Information Processing Systems 21*, pp. 65–72, 2009.

[6] N. Usunier, M. Amini, and P. Gallinari, "A data-dependent generalisation error bound for the AUC," in *ICML'05 workshop on ROC Analysis in Machine Learning*, 2005.

[7] M. Amini, "Interactive learning for text summarization," in *Proceedings of the PKDD/MLTIA Workshop on Machine Learning and Textual Information Access*, (Lyon - France), 2000.

[8] Y.-M. Kim, M.-R. Amini, C. Goutte, and P. Gallinari, "Multi-view clustering of multilingual documents," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pp. 821–822, 2010.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York, USA : Academic Press, 1972.

[10] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2001.

[11] B. Schölkopf and A. J. Smola, *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002.

[12] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification : a survey of some recent advances," *ESAIM : Probability and Statistics*, pp. 323–375, 2005.

[13] M. R. Genesereth and N. J. Nilsson, *Logical Foundations of Artificial Intelligence*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1987.

[14] O. Bousquet, S. Boucheron, and G. Lugosi, "Introduction to statistical learning theory," in *Advanced Lectures on Machine Learning*, pp. 169–207, 2003.

[15] J. Langford, "Tutorial on practical prediction theory for classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, Dec. 2005.

[16] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, pp. 13–30, 1963.

[17] P. L. Tchebychev, "Des valeurs moyennes," *Journal de mathématiques pures et appliquées*, vol. 2, no. 12, pp. 177–184, 1867.

[18] V. N. Vapnik and A. J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and its Applications*, vol. 16, pp. 264–280, 1971.

[19] N. Sauer, "On the density of families of sets," *Journal of Combinatorial Theory*, vol. 13, no. 1, pp. 145–147, 1972.

[20] S. Shelah, "A combinatorial problem : Stability and order for models and theories in infinity languages," *Pacific Journal of Mathematics*, vol. 41, pp. 247–261, 1972.

[21] H. Brönnimann and M. T. Goodrich, "Almost optimal set covers in finite vc-dimension.," *Discrete and Computational Geometry*, vol. 14, no. 4, pp. 463–479, 1995.

[22] N. Cesa-Bianchi and D. Haussler, "A graph-theoretic generalization of the sauer-shelah lemma," *Discrete Applied Mathematics*, vol. 86, pp. 27–35, 1998.

[23] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. MIT Press, 2012.

[24] V. N. Vapnik and A. J. Chervonenkis, "Theory of pattern recognition," *Nauka*, 1974.

[25] V. I. Koltchinskii and D. Panchenko, "Rademacher processes and bounding the risk of function learning," in *High Dimensional Probability II* (E. Gine, D. Mason, and J. Wellner, eds.), pp. 443–459, 2000.

[26] V. I. Koltchinskii, "Rademacher penalties and structural risk minimization," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1902–1914, 2001.

[27] P. Massart, "Some applications of concentration inequalities to statistics," *Annales de la faculté des sciences de Toulouse*, vol. 9, no. 2, pp. 245–303, 2000.

[28] P. L. Bartlett and S. Mendelson, "Rademacher and gaussian complexities : risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2003.

[29] J. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. New York, USA : Cambridge Press University, 2004.

[30] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the vapnik-chervonenkis dimension," *Journal of the ACM*, vol. 36, pp. 929–965, 1989.

[31] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant, "A general lower bound on the number of examples needed for learning," *Information and Computation*, vol. 82, pp. 247–261, 1989.

[32] E. Giné, "Empirical processes and applications : an overview," *Bernoulli*, vol. 2, no. 1, pp. 1–28, 1996.

[33] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics*, vol. 141, pp. 148–188, 1989.

[34] M. Ledoux and M. Talagrand, *Probability in Banach Spaces : Isoperimetry and Processes*. Springer Verlag, 1991.

[35] A. Antos, B. Kégl, T. Linder, and G. Lugosi, "Data-dependent margin-based generalization bounds for classification," *Journal of Machine Learning Research*, vol. 3, pp. 73–98, 2003.

[36] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the* 14^{th} *International Joint Conference on Artificial Intelligence (IJCNN)*, pp. 1137—1143, 1995.

[37] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, classification, and risk bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, 2006.

[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA : Cambridge University Press, 2004.

[39] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer, 2006.

[40] D. D. Lewis, Y. Yang, T. Rose, and F. Li, "RCV1 : A new benchmark collection for text categorization research," *Journal of Machine Learning Research*, vol. 5, pp. 361–397, 2004.

[41] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2004.

[42] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[43] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of Artificial Intelligence Research*, vol. 2, pp. 263–286, 1995.

[44] E. L. Allwin, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary : A unifying approach for margin classifiers," *Journal of Machine Learning Research*, vol. 1, pp. 113 – 141, 2000.

[45] Y. Guermeur, *SVM Multiclasses, Théorie et Applications*. Habilitation à diriger des recherches, Université Nancy 1, 2007.

[46] Y. Guermeur, "Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs," *Communications in Statistics - Theory and Methods*, vol. 39, no. 3, pp. 543–557, 2010.

[47] D. E. Rumelhart, G. E. Hinton, and R. Williams, "Learning internal representations by error propagation," *Parallel Distributed Processing : Explorations in the Microstructure of Cognition*, vol. I, 1986.

[48] L. Bottou, "Online algorithms and stochastic approximations," in *Online Learning and Neural Networks* (D. Saad, ed.), Cambridge, UK : Cambridge University Press, 1998. revised, oct 2012.

[49] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)* (Y. Lechevallier and G. Saporta, eds.), (Paris, France), pp. 177–187, Springer, August 2010.

[50] F. Bach and E. Moulines, "Non-strongly-convex smooth stochastic approximation convergence rate $o(\frac{1}{n})$," in *Advances in Neural Information Processing Systems (NIPS 26)*, pp. 773–781, 2013.

[51] A. S. Nemirovski and D. B. Yudin, *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

[52] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *Journal of the Society for Industrial and Applied Mathematics on Optimization (SIOPT)*, vol. 19, no. 4, pp. 1574–1609, 2009.

[53] P. E. Gill and M. W. Leonard, "Reduced-hessian quasi-newton methods for unconstrained optimization," *Journal of the Society for Industrial and Applied Mathematics on Optimization (SIOPT)*, vol. 12, no. 1, 2001.

[54] P. Deuflhard, *Newton Methods for Nonlinear Problems : Affine Invariance and Adaptive Algorithms*. Springer Verlag, 2004.

[55] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. Sagastizàbal, *Numerical optimization, theoretical and numerical aspects*. Springer Verlag, 2006.

[56] R. Fletcher, *Practical methods of optimization*. New York, USA : John Wiley & Sons, 1987.

[57] W. C. Davidon, "Variable metric method for minimization," *Journal of the Society for Industrial and Applied Mathematics on Optimization (SIOPT)*, vol. 1, no. 1, pp. 1–17, 1991.

[58] E. Polak, *Computational methods in optimization*. Academic press, 1971.

[59] P. Wolfe, "Convergence conditions for ascent methods," *SIAM Review*, vol. 11, no. 2, pp. 226–235, 1966.

[60] L. Armijo, "Minimization of functions having lipschitz continuous first partial derivatives," *Pacific Journal of Mathematics*, vol. 16, no. 1, pp. 1–3, 1966.

[61] G. Zoutendijk, "Some recent development in nonlinear programming," in *5th Conference on Optimization Techniques, Part 1*, pp. 407–417, 1973.

[62] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Classics in Applied Mathematics, 16)*. Soc for Industrial & Applied Math, 1996.

[63] K. A. Atkinson, *An introduction to numerical analysis*. John Wiley and Sons, 1988.

[64] M. R. Hestenes and E. Stiefel, "Methods of conjugate gradients for solving linear systems," *Journal of Research of the National Bureau of Standards*, vol. 49, pp. 409–436, 1952.

[65] E. Polak and G. Ribiere, "Note sur la convergence de méthodes de directions conjuguées," *ESAIM : Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, vol. 3, no. R1, pp. 35–43, 1969.

[66] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *The Computer Journal*, vol. 7, no. 2, pp. 149–154, 1964.

[67] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biophysics*, vol. 5, pp. 115–133, 1943.

[68] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, pp. 433–460, 1950.

[69] D. Hebb, *The Organization of Behavior*. John Wiley, 1949.

[70] M. Minsky, "A neural-analogue calculator based upon a probability model of reinforcement," *Harvard University Psychological Laboratories, Cambridge, Massachusetts*, 1952.

[71] F. Rosenblatt, "The perceptron : A probabilistic model for information storage and organization in the brain," *Psychological Review*, vol. 65, pp. 386–408, 1958.

[72] A. B. Novikoff, "On convergence proofs on perceptrons," *Symposium on the Mathematical Theory of Automata*, vol. 12, pp. 615–622, 1962.

[73] G. Widrow and M. Hoff, "Adaptive switching circuits," *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record*, vol. 4, pp. 96–104, 1960.

[74] N. J. Nilsson, *Learning machines ; foundations of trainable pattern-classifying systems*. McGraw-Hill, 1965.

[75] M. Minsky and S. Papert, *Perceptrons : An Introduction to Computational Geometry*. MIT Press, 1969.

[76] J. A. Anderson and E. Rosenfeld, *Neurocomputing : Foundations of Research*. MIT Press, 1988.

[77] J. J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," in *Proceedings of the National Academy of Sciences USA*, pp. 3088–3092, 1984.

[78] Y. LeCun, L. Bottou, and Y. Bengio, "Reading checks with multilayer graph transformer networks," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, 1997.

[79] S. Hocreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[80] V. N. Vapnik, *The nature of statistical learning theory*. 1995.

[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, pp. 1097–1105, 2012.

[82] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine Learning Journal*, vol. 37, pp. 277–296, 1999.

[83] J. Truett, J. Cornfield, and W. Kannel, "A multivariate analysis of the risk of coronary heart disease in framingham," *Journal of Chronic Diseases*, vol. 20, no. 7, pp. 511–524, 1967.

[84] J. A. Anderson, "Logistic discrimination," *Handbook of Statistics*, vol. 2, pp. 169–191, 1982.

[85] M. Kupperman, "Probabilities of hypotheses et information-statistics in sampling from exponential-class populations," *Annals of Mathematical Statistics*, vol. 9, no. 2, pp. 571–575, 1958.

[86] P. J. Werbos, *Beyond Regression : New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.

[87] D. Parker, "Learning logic," Tech. Rep. TR-87, Cambridge, MA : Center for Computational Research in Economics and Management Science, MIT, 1985.

[88] L. Bottou, *Une Approche théorique de l'Apprentissage Connexionniste : Applications à la Reconnaissance de la Parole*. PhD thesis, Université de Paris XI, Orsay, France, 1991.

[89] D. Hubel and T. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *Journal of Physiology*, vol. 195, no. 1, pp. 215–243, 1968.

[90] K. Fukushima, "Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological Cybernetics*, vol. 36, no. 4, 1980.

[91] A. Waibel, "Phoneme recognition using time-delay neural networks," in *Meeting of the Institute of Electrical, Information and Communication Engineers*, 1987.

[92] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[93] A. J. Robinson and F. Fallside, "The utility driven dynamic error propagation network," tech. rep., Engineering Department, Cambridge University, Cambridge, UK, 1987.

[94] R. Williams and D. Zipser, "Gradient-based learning algorithms for recurrent networks and their computational complexity," in *Backpropagation : theory, architectures, and applications* (B. Schölkopf, C. Burges, and A. Smola, eds.), ch. 4, p. 433–486, 1995.

[95] M. D. Richard and R. P. Lippman, "Neural network classifiers estimate bayesian a posteriori probabilities," *Neural Computation*, vol. 3, no. 4, pp. 461–483, 1991.

[96] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[97] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, 2010.

[98] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[99] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $o(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.

[100] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 1139–1147, 2013.

[101] S. Ioffe and C. Szegedy, "Batch normalization : Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 448–456, 2015.

[102] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.

[103] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout : A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[104] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992.

[105] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods - Support Vector Learning* (B. Schölkopf, C. Burges, and A. Smola, eds.), ch. 11, pp. 169–184, Cambridge, MA : MIT Press, 1999.

[106] R. Fan, K. Chang, C. Hsieh, X. Wang, and C.-J. Lin, "LIBLINEAR : A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[107] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos : primal estimated sub-gradient solver for svm," *Mathematical Programming*, vol. 127, no. 1, pp. 3–30, 2011.

[108] J. Mercer, "Functions of positive and negative type and their connection with the theory of integral equations," *Philosophical Transactions of the Royal Society*, vol. 209, pp. 415–446, 1909.

[109] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *European Symposium on Artificial Neural Netwroks (ESANN)*, pp. 219 –224, 1999.

[110] K. Crammer and Y. Singer, "On the algorithmic implementation of multi class kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[111] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines : Theory and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.

[112] L. Valiant, "The theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[113] D. Kearns and L. Valiant, "Learning boolean formulae or finite automata is as hard as factoring," Tech. Rep. TR-14-88, Harvard University Aiken Computation Laboratory, 1988.

[114] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[115] R. E. Schapire, "Theoretical views of boosting and applications," in *Proceedings of the 10th International Conference on Algorithmic Learning Theory*, pp. 13–25, 1999.

[116] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[117] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin : a new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1680, 1998.

[118] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[119] D. M. Titterington, A. F. M. Smith, and U. E. Smith, *Statistical Analysis of Finite Mixture Distributions*. Wiley, New York, 1985.

[120] M. J. Symons, "Clustering criteria and multivariate normal mixtures," *Biometrics*, vol. 37, no. 1, pp. 35–43, 1981.

[121] G. Celeux and G. Govaert, "A classification em algorithm for clustering and two stochastic versions," *Computational Statistics and Data Analysis*, vol. 14, no. 3, pp. 315–332, 1992.

[122] T. Zhang and F. J. Oles, "A probability analysis on the value of unlabeled data for classification problems," in *17th International Conference on Machine Learning*, 2000.

[123] F. G. Cozman and I. Cohen, "Unlabeled data can degrade classification performance of generative classifiers," in *Fifteenth International Florida Artificial Intelligence Society Conference*, pp. 327–331, 2002.

[124] M. Seeger, "Learning with labeled and unlabeled data," tech. rep., 2001.

[125] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 27 – 34, 2002.

[126] G. J. Machlachlan, *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Interscience, 1992.

[127] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances in Neural Information Processing Systems (NIPS 17)*, pp. 529–536, MIT Press, 2005.

[128] O. Chapelle, B. Schölkopf, and A. Zien, eds., *Semi-Supervised Learning*. Cambridge, MA : MIT Press, 2006.

[129] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning Journal*, vol. 39, no. 2 - 3, pp. 103–134, 2000.

[130] M.-R. Amini and É. Gaussier, *Recherche d'Information - applications, modèles et algorithmes*. Eyrolles, 2013.

[131] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang, "Semisupervised learning of classifiers : Theory, algorithms, and their application to human-computer interaction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 12, pp. 1553–1567, 2004.

[132] S. C. Fralick, "Learning to recognize patterns without a teacher," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 57–64, 1967.

[133] E. A. Patrick, J. P. Costello, and F. C. Monds, "Decision-directed estimation of a two-class decision boundary," *IEEE Transactions on Information Theory*, vol. 9, no. 3, pp. 197–205, 1970.

[134] G. Tür, D. Z. Hakkani-Tür, and R. E. Schapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2, pp. 171–186, 2005.

[135] R. Urner, S. Shalev-Shwartz, and S. Ben-David, "Access to unlabeled data can speed up prediction time," in *28th International Conference on Machine Learning, ICML 2011*, pp. 641–648, 2011.

[136] P. Derbeko, E. El-Yaniv, and R. Meir, "Error bounds for transductive learning via compression and clustering," in *Advances in Neural Information Processing Systems (NIPS 15)*, pp. 1085–1092, 2003.

[137] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the* 16^{th} *International Conference on Machine Learning*, pp. 200–209, 1999.

[138] T. Joachims, *Learning to Classify Text Using Support Vector Machines : Methods, Theory and Algorithms*. Norwell, MA, USA : Kluwer Academic Publishers, 2002.

[139] Z. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization theory and applications*, vol. 72, no. 1, pp. 7–35, 1992.

[140] M.-R. Amini, N. Usunier, and F. Laviolette, "A transductive bound for the voted classifier with an application to semi-supervised learning," in *Advances in Neural Information Processing Systems (NIPS 21)*, pp. 65–72, 2009.

[141] G. Dantzig, "Maximization of a linear function of variables subject to linear inequalities," in *Activity Analysis of Production and Allocation* (T. Koopmans, ed.), p. 339–347, Wiley, New York, 1951.

[142] F. Bach, R. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proceedings of the Twenty-first International Conference on Machine Learning*, 2004.

[143] V. Sindhwani, P. Niyogi, and M. Belkin., "A co-regularization approach to semi-supervised learning with multiple views," in *ICML-05 Workshop on Learning with Multiple Views*, pp. 74–79, 2005.

[144] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Learning Theory*, pp. 92–100, 1998.

[145] B. Leskes, "The value of agreement, a new boosting algorithm," in *Proceedings of Conference on Learning Theory (COLT)*, pp. 95–110, 2005.

[146] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep. CMU-CALD-02-107, Carnegie Mellon University, 2002.

[147] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in 20^{th} *International Conference on Machine Learning*, pp. 912–919, 2003.

[148] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS 16)*, pp. 321–328, MIT Press, 2004.

[149] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling.* ASA-SIAM Series on Statistics and Applied Probability, Philadelphia, Pa. SIAM, Society for Industrial and Applied Mathematics Alexandria, Va. ASA, American Statistical Association, 1999.

[150] M. Szummer and T. Jaakkola, "Partially labeled classification with markov random walks," in *Advances in Neural Information Processing Systems (NIPS 14)*, pp. 945–952, 2002.

[151] E. Montroll, "Random walks in multidimensional spaces, especially on periodic lattices," *Journal of the Society for Industrial and Applied Mathematics (SIAM)*, vol. 4, no. 4, pp. 241 – 260, 1956.

[152] W. W. Cohen, R. E. Schapire, and Y. Singer, "Learning to order things," in *Advances in Neural Information Processing Systems (NIPS 10)*, pp. 451–457, 1998.

[153] C. Rudin, C. Cortes, M. Mohri, and R. E. Schapire, "Margin-based ranking meets boosting in the middle," in *Conference On Learning Theory (COLT)*, 2005.

[154] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth, "Generalization bounds for the area under the roc curve," *Journal of Machine Learning Research*, vol. 6, pp. 393–425, 2005.

[155] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," in *Advances in Neural Information Processing Systems (NIPS 16)*, pp. 313–320, 2004.

[156] G. Salton, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[157] S. Hill, H. Zaragoza, R. Herbrich, and P. Rayner, "Average Precision and the Problem of Generalisation," in *SIGIR Workshop on Mathematical and Formal Methods in Information Retrieval*, 2002.

[158] S. E. Robertson and S. Walker, "Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval," in *SIGIR'94, conference on Research and development in information retrieval*, pp. 232–241, 1994.

[159] S. Clinchant and É. Gaussier, "Information-based models for *a*d hoc IR," in *SIGIR'10, conference on Research and development in information retrieval*, 2010.

[160] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[161] K. Crammer and Y. Singer, "Pranking with ranking," in *Advances in Neural Information Processing Systems (NIPS 14)*, pp. 641–647, 2002.

[162] A. Shashua and A. Levin, "Ranking with large margin principle : Two approaches," in *Advances in Neural Information Processing Systems (NIPS 15)*, pp. 961–968, 2003.

[163] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," in *22th International Conference on Machine Learning, ICML 2005*, pp. 145–152, 2005.

[164] T. Qin, T.-Y. Liu, and H. Li, "A general approximation framework for direct optimization of information retrieval measures," Tech. Rep. MSR-TR-2008- 164, Microsoft Research, November 2008.

[165] M. Taylor, J. Guiver, S. Robertson, and T. Minka, "Softrank : Optimising non-smooth rank metrics," in *WSDM 2008*, 2008.

[166] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, "A support vector method for optimizing average precision," in *SIGIR '07, conference on Research and Development in Information Retrieval*, pp. 271–278, 2007.

[167] J. Xu and H. Li, "Adarank : A boosting algorithm for information retrieval," in *SIGIR '07, conference on Research and Development in Information Retrieval*, pp. 391 – 398, 2007.

[168] J. Xu, T.-Y. Liu, M. Lu, H. Li, and W.-Y. Ma, "Directly optimizing evaluation measures in learning to rank," in *SIGIR '08, conference on Research and Development in Information Retrieval*, pp. 107–114, 2008.

[169] C. Calauzènes, N. Usunier, and P. Gallinari, "On the (non-)existence of convex, calibrated surrogate losses for ranking," in *Advances in Neural Information Processing Systems (NIPS 25)*, pp. 197–205, 2012.

[170] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of Machine Learning Research*, vol. 4, pp. 933–969, 2003.

[171] N. Usunier, *Apprentissage de fonctions d'ordonnancement : une étude théorique de la réduction à la classification et deux applications à la Recherche d'Information*. PhD thesis, Université Pierre & Marie Curie, 2006.

[172] S. K. Wong and Y. Y. Yao, "Linear structure in information retrieval," in *SIGIR'88, conference on Research and Development in Information Retrieval*, pp. 219–232, 1988.

[173] R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer, "Learning preference relations for information retrieval," in *Proceedings of the AAAI Workshop Text Categorization and Machine Learning, Madison, USA.*, 1998.

[174] A. Rakotomamonjy, "Optimizing area under roc curve with SVMs.," in *1st International workshop on ROC Analysis in Artificial Intelligence*, pp. 71–80, 2004.

[175] N. Usunier, M.-R. Amini, and P. Gallinari, "Generalization error bounds for classifiers trained with interdependent data," in *Advances in Neural Information Processing Systems (NIPS 18)*, pp. 1369–1376, 2006.

[176] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[177] P. Barbé and M. Ledoux, *Probabilité*. EDP Sciences, 2007.

[178] W. Feller, *An Introduction to Probability Theory and Its Applications*. Wiley, 1968.

[179] T. Bayes, "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London*, vol. 53, pp. 370–418, 1763.

[180] P. S. Laplace, "Mémoire sur la probabilité des causes par les Événements," *Académie Royale des sciences de Paris (Savants étrangers)*, vol. 6, pp. 621–656, 1771.

[181] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities : A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Index