UCLouvain

ldats2470

2024

Statistical Machine Learning and High Dimensional Data Analysis

3.00 crédits	15.0 h	Q2

Enseignants	Hafner Christian ;
Langue d'enseignement	Anglais
Lieu du cours	Louvain-la-Neuve
Préalables	Concepts et outils équivalents à ceux enseignés dans les UEs LSTAT2020 Logiciels et programmation statistique de base LSTAT2120 Linear models LSTAT2110 Analyse des données
Thèmes abordés	1. Partitioning methods for clustering 2. Statistical approaches for dimension reduction and feature extraction 3. Regularization methods in high dimensions, including linear and nonlinear shrinkage 4. Applications
Acquis d'apprentissage	
Modes d'évaluation des acquis des étudiants	Project using a real data set, and an oral exam
Méthodes d'enseignement	The lectures provide the theoretical material, give many practical examples, and show how to implement the methods in common programming packages.
Contenu	1. Partitioning methods for clustering • k-means and variants • Nonlinear k-means with kernels • Support Vector Machines and other multiple kernel learning machines • Spectral clustering 2. Statistical approaches for dimension reduction and feature extraction • Factor models and probabilistic PCA • Kernels for non-linear PCA • Kernels for non-linear ICA 3. Regularization methods in high dimensions, including linear and nonlinear shrinkage 4. Applications

Bibliographie	A syllabus will be written based on the following sources (not exhaustive): Amari, S. and Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. Neural Networks, 12(6):783-789. Chitta, R., Jin, R., Havens, T.C. and Jain, A.K. (2011). Approximate kernel k-means: Solution to large scale kernel clustering. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 895-903. ACM. Devroye, L., Gyorfi, L. and Lugosi, G. (2013). A probabilistic theory of pattern recognition, volume 31. Springer Science & Business Media. Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models, The Annals of Statistics, 147, 186–197. Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements, Journal of the Royal Statistical Society: Series B, 75, 603-680. Gonen, M. and Alpaydyin, E. (2011). Multiple kernel learning algorithms. Journal of machine learning research, 12:2211-2268. Grandvalet, Y. and Canu, S. (2003). Adaptive scaling for feature selection in SVMs. In: Advances in neural information processing systems, pages 569-576. Guyon, I. and Elissee, A. (2006). An introduction to feature extraction. Feature extraction, pages 1-25. Fine, S. and Scheinberg, K. (2001). Efficient SVM training using low-rank kernel representations. Journal of Machine Learning Research, 2(Dec):243-264. Hagen, L. and Kahng, A.B. (1992). New spectral methods for ratio cut partitioning and clustering. IEEE transactions on computer-aided design of integrated circuits and systems, 11(9):1074-1085. Hardle, W., Dwi Prastyo, D. and Hafner, C.M. ¨ (2014). Support Vector Machines with Evolutionary Feature Selection for Default Prediction, Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics, Oxford UP, edited by A. Ullah, J. Racine and L. Su. Hardle, W. and Simar, L. ¨ (2015). Applied Multivariate Statistical Analysis, Springer Verlag. Jain, A.K. (2010). Data clustering: 50 years beyond k-means. Pattern recognition letters, 31(8): 651-666. Johnson, W.B. and Lindenstrauss, J. (1984). Extensions of lipschitz mappings into a hilbert space. Contemporary mathematics, 26(189-206):1. Keerthi, S.S. and Lin, Ch-J. (2003). Asymptotic behaviors of support vector machines with gaussian kernel. Neural computation, 15(7):1667-1689. Kloft, M., Brefeld, U., Laskov, P., Muller, K.-R., Zien, A., and Sonnenburg, S. (2009). Efficient and accurate lp-norm multiple kernel learning. In Advances in neural information processing systems, pages 997-1005. Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, Journal of Multivariate Analysis, 88, 365-411. Ledoit, O. and Wolf, M. (2012). Nonlinear shrinkage estimation of large-dimensional covariance matrices, Annals of Statistics, 40, 1024-1060. Ledoit, O. and Wolf, M. (2015). Spectrum estimation: a unified framework for covariance matrix estimation and PCA in large dimensions, Journal of Multivariate Analysis, 139, 360-384. Ledoit, O. and Wolf, M. (2020). Direct nonlinear shrinkage estimation of large dimensional covariance matrices, Annals of Statistics. Lee, Y.-J. and Huang, S.-Y. (2007). Reduced support vector machines: A statistical theory. IEEE Transactions on Neural Networks, 18(1):1-13. Lee, S.-W. and Bien, Z. (2010). Representation of a Fisher criterion function in a kernel feature space. IEEE transactions on neural networks, 21(2):333-339. Mohar, B., Alavi, Y., Chartrand, G. and Oellermann, OR (1991). The laplacian spectrum of graphs. Graph theory, combinatorics, and applications, 2(871-898):12. Neumann, J., Schnorr, C. and Steidl, G. (2005). Combined svm-based feature selection and classification. Machine learning, 61(1):129-150. Ng, A.-Y., Jordan, M.I. and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems, pages 849-856. Peters, G. W., Statistical Machine Learning and Data Analytic Methods for Risk and Insurance (Version 8, 2017). Available at SSRN: https://ssrn.com/abstract=3050592. Scholkopf, B. and Smola, A.J. (2001). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001. Yao, J., Zheng, S. and Bai, Z. (2015). Large sample covariance matrices and highdimensional data analysis, Cambridge UP
Faculté ou entité en charge:	LSBA

Programmes / formations proposant cette unité d'enseignement (UE)

Intitulé du programme	Sigle	Crédits	Prérequis	Acquis d'apprentissage
Master [120] en science des données, orientation statistique	DATS2M	3		
Master [120] en statistique, orientation générale	STAT2M	3		
Certificat d'université : Statistique et science des données (15/30 crédits)	STAT2FC	3		