Chair Data Science and Artificial Intelligence for Digitalized Industry and Services

Internship project

Subject

Kernel-based multi-scale learning with the alpha-procedure

Possibility to continue as a PhD candidate

YES (Funding to be confirmed)

About the chair

The Chair Data Science and Artificial Intelligence for Digitalized Industry and Services (DSAIDIS), lead by Florence d'Alché-Buc, a Professor in the department Image, Data, Signal of Telecom Paris, unites five industrial partners: Airbus Defence & Space, Engie, Idemia, Safran et Valeo. It's general objective is to develop, in collaboration with the partners, teaching and research of the international level.

Its four principal research directions are:

1. Building predictive analytics on time series and data streams.
2. Exploiting large scale, heterogeneous, partially labeled data.
3. Machine Learning for trusted and robust decision.
4. Learning through interactions with environment.

Description of the internship

Supervision

Pavlo Mozharovskyi (https://perso.telecom-paristech.fr/mozharovskyi/)

Location and dates of the internship

Address : Télécom Paris, 19 Place Marguerite Perey, 91120 Palaiseau

Date of the beginning of the internship : beginning 2024

Team where the thesis will be written

Department IDS, Team Signal, Statistique et Apprentissage (S2A)

Keywords

Alpha-procedure, reproducing kernel Hilbert space, Gram-Schmidt orthogonalization, robustness, projection pursuit, visualization.

Detailed subject

The goal of this project is to develop a method for supervised classification inspired by the DDalpha-classifier [2], a proper combination of the depth [7,10,4] embedding [3] and of the alpha-procedure [8]. The DDalpha-classifier is fully non-parametric and robust due to the depth embedding but also fast and indicator-risk driven in the low-dimensional Euclidean space. It has been implemented in the R-package ddalpha [6] and tested satisfactorily with a number of depths on numerous data sets [5] for speed and quality of classification. Having shown promising results when coupled with data depth, the

alpha-procedure itself requires further investigations and development, which are in the focus of the proposed internship.

The main purpose of this internship is to explore a variation of the alpha-procedure aimed on construction of flexible frontiers between the patterns to be separated. The alpha-procedure [see 8, 7] is an iterative heuristics that synthesizes a finite-dimensional space of relevant features. Due to its iterative nature, practically any loss can be minimized without substantial increase of computing time, on a full data set or its sub-sample. Here, we propose to employ the alpha-procedure in a properly constructed reproducing kernel Hilbert space (RKHS), *i.e.* after a proper kernel-based mapping of observations. To provide a finite-dimensional basis in RKHS, scalar products with the observations of the training sample are used, which can be further made orthogonal by the Gram-Schmidt method based on the so-called kernel trick. Further, the multi-scale property shall be introduced, which consists in choosing different kernels (or kernel parameters) for different observations of the training sample. While the parameters shall be chosen by (cross-)validation, this can be built in the step of the choice of observation/kernel without additional computational cost. The reverse Gram-Schmidt process shall further allow for a simple multi-scale classification rule represented by the weighted observations of the training sample [see also 1]. (Due to the multi-scale nature only a few coefficients should differ from zero here.)

The proposed method is naturally robust by preserving the original (indicator) loss on all its stages. Since all iteration steps are performed in low-dimensional (up to bivariate) spaces, the classification rule as well as the learning process can both be visualized, which allows for on-line control of the training procedure. The interpretability is further enhanced by representation in terms of observations of the training set, which also allows to order them with respect to their importance. Finally, due to built-in per-step (cross-)validation, the method is scalable for larger data sets and suited for parallel computing.

Though the proposal is on an early stage and needs both theoretical and implementation development, an experimental prototype exists and is ready to preliminary testing. Nevertheless, this should first be improved in sense of option flexibility. Further, an experimental study, involving both synthetic and real-world data sets, which illustrates the method's performance under different settings, should be conducted to reveal (close to) optimal settings configuration and provide guidance for the choice of parameters for an applicant.

Candidate profile

Student having master 2 research

- Statistical learning, bases of probability

- Good level of programming (R, C/C++, Python)

- Good command of English

Application

To send on pavlo.mozharovskyi@telecom-paris.fr:

- Curriculum Vitae

- Personalized motivation letter that explains interest of the candidate in the subject (can be directly in the body of the email)

- Grade reports for recent years

- Contact of a person willing to give recommendation

Incomplete applications will not be considered.

References

[1] Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning 20, 273–297.

[2] Lange, T., Mosler, K., and Mozharovskyi, P. (2014). Fast nonparametric classification based on data depth. Statistical Papers 55, 49–69.

[3] Li, J., Cuesta-Albertos, J.A., and Liu, R.Y. (2012). DD-classifier: Nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association 107, 737–753.

[4] Mosler, K. (2013). Depth Statistics. In: Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather, Becker, C., Fried, R. and Kuhnt, S. Eds., Springer, Berlin, 17–34.

[5] Mozharovskyi, P., Mosler, K., and Lange, T. (2015). Classifying real-world data with the DDα-procedure. Advances in Data Analysis and Classification 9, 287–314.

[6] Pokotylo, O., Mozharovskyi, P., and Dyckerhoff, R. (2018). Depth and depth-based classification with R-package ddalpha. Journal of Statistical Software, in press.

[7] Tukey, J.W. (1975). Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, Volume 2, James, R.D. Ed., Canadian Mathematical Congress, 523–531.

[8] Vasil'ev, V.I. (2003). The reduction principle in problems of revealing regularities I. Cybernetics and Systems Analysis 39, 686–694.

[9] Vasil'ev, V.I. and Lange, T. (1998). The duality principle in learning for pattern recognition (in Russian). Kibernetika i Vytschislit'elnaya Technika 121, 7–16.

[10] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. The Annals of Statistics 28(2), 461–482.