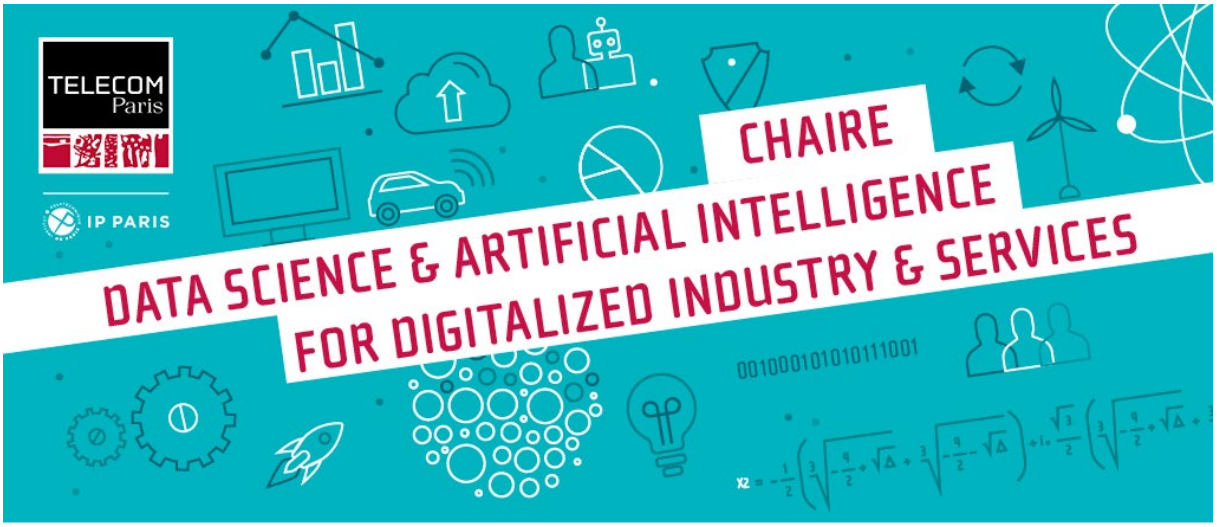

Offre de stage

Prédiction de liens pour des graphes non Poissoniens

Possibilité de poursuivre sur une thèse

A propos de la chaire

La Chaire Data Science and Artificial Intelligence for Digitalized Industry and Services, portée par Florence d'Alché-Buc, enseignante-chercheuse dans le département Image, Données, Signal de Télécom Paris, la chaire DSAIDIS réunit cinq partenaires industriels : Airbus Defence & Space, Engie, Idemia, Safran et Valeo. Son objectif général est de développer, en liaison étroite avec les partenaires, une formation et une recherche de niveau international.

Ses quatre principaux axes de recherche sont :

1. Analyse et prévision de séries temporelles (Predictive Analytics on Time Series) ;

2. Exploitation de données hétérogènes, massives et partiellement étiquetées (Exploiting Large Scale and Heterogeneous, Partially Labelled Data) ;

3. Apprentissage pour une prise de décision robuste et fiable (Learning for Trusted and Robust Decision) ;

4. Apprentissage dans un environnement dynamique (Learning through Interactions with a Changing Environment).

Plus d'information sur la chaire : **www.telecom-paris.fr/dsaidis**

Description du stage

Encadrement

Stephan Clémençon

Lieu et dates du stage

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

Date de début du stage : début 2020

Équipe(s) d'accueil de la thèse

Département IDS, équipe S2A

Mots clés

Reconstruction de graphe, prédiction de lien, théorie des graphes, théorie de l'apprentissage statistique

Sujet détaillé

The problem of predicting connections between a set of data points finds many applications, in systems biology and social network analysis among others. This statistical learning problem is motivated by a variety of applications such as systems biology (e.g., inferring protein-protein interactions or metabolic networks, see [7, 9] and social network analysis (e.g., predicting future connections between users, see [10].It has recently been the subject of a good deal of attention in the machine learning literature (see [17, 18, 15]), and is also known as supervised link prediction [11,15]. The learning task is formulated as the minimization of a reconstruction risk, whose natural empirical version is the average prediction error over the n (n - 1)/2 pairs of nodes in a training graph of size n. Under standard complexity assumptions on the set of candidate prediction rules, excess risk bounds of the order O ($1/\sqrt{n}$) for the empirical risk minimizers have been established by [18] based on a representation of the objective functional very similar to the first Hoeffding decomposition for second-order U-statistics. However, [18] ignored the computational complexity of finding an empirical risk minimizer, which scales at least as O (n^2) since the empirical graph reconstruction risk involves summing up over n (n - 1)/2 terms. This makes the approach impractical when dealing with large graphs commonly found in many applications. sed on a different decomposition of the excess of reconstruction risk of any decision rule candidate, involving the second Hoeffding representation of a U-statistic approximating it, as well as appropriate maximal/concentration inequalities, [19] establishes universal fast rates for this problem, that is, rate of order O (log n/n) are always achieved by empirical reconstruction risk minimizers, in absence of any restrictive condition imposed on the data distribution.

This is much faster than the O ($1/\sqrt{n}$) rate established by [18].

Performance of minimizers of computationally cheaper Monte-Carlo estimates of the empirical reconstruction risk, built by averaging over B pairs of vertices drawn with replacement are also investigated in order to scale up the empirical risk minimization procedure. The rate bounds obtained highlight that B plays the role of a tuning parameter to achieve an effective trade-off between statistical accuracy and computational cost.

While the result obtained in [18, 19] rely on independence assumption on the feature of each nodes, the goal of the present research subject is to investigate how the results established in [19] translate when relaxing this hypothesis and assuming that the degree distribution of the graph follows a powerlaw, as in a scale-free network.

Profil du candidat

Etudiant titulaire d'un master 2 recherche

- Apprentissage statistique / reconnaissance des formes

- Bon niveau en programmation (Java, C/C++, Python)

- Bon niveau d'anglais

Candidatures

A envoyer à stephan.clemencon@telecom-paris.fr

- Curriculum Vitae

- Lettre de motivation personnalisée expliquant l'intérêt du candidat sur le sujet (directement dans le corps du mail)

- Relevés de notes des années précédentes

- Contact d'une personne de référence

Les candidatures incomplètes ne seront pas examinées.

Références

[1] Bellet, A., Habrard, A., and Sebban, M. (2015). Metric Learning. Morgan & Claypool Publishers.

[2] Boucheron, S., Bousquet, O., Lugosi, G., and Massart, P. (2005). Moment inequalities for functions of independent random variables. Ann. Stat., 33(2):514–560.

[3] Clémençon, S. (2014). A statistical view of clustering performance through the theory of U-processes. Journal of Multivariate Analysis, 124:42 – 56.

[4] Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and Empirical Minimization of U-statistics. Ann. Stat., 36(2):844–874.

[5] Cukierski, W., Hamner, B., and Yang, B. (2011). Graph-based features for supervised link prediction. In IJCNN.

[6] De la Pena, V. and Giné, E. (1999). Decoupling : from dependence to independence. Springer.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics, 19:293–325.

[7] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A., Snyder, M., Greenblatt, J., andGerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science, 302(5644):449–453.

[8] Janson, S. and Nowicki, K. (1991). The asymptotic distributions of generalized U-statistics with applications to random graphs. Probability Theory and Related Fields, 90:341–375.

[9] Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data. Pharmacogenomics, 2(4):373–385.

[10] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In CIKM.

[11] Lichtenwalter, R., Lussier, J., and Chawla, N. (2010). New perspectives and methods in link prediction. In KDD.

[12] Mammen, E. and Tsybakov, A. (1999). Smooth discrimination analysis. Ann. Stat., 27(6):1808–1829.

[13] Massart, P. and Nédélec, E. (2006). Risk bounds for statistical learning. Ann. Stat., 34(5).

[14] Mattick, J. and Gagen, M. (2005). Accelerating networks. Science, 307(5711):856–858.

[15] Shaw, B., Huang, B., and Jebara, T. (2011). Learning a Distance Metric from a Network. In NIPS.

[16] Vert, J.-P., Qiu, J., and Noble, W. S. (2007). A new pairwise kernel for biological network inference with support vector machines. BMC Bioinformatics, 8(10).

[17] Vert, J.-P. and Yamanishi, Y. (2004). Supervised graph inference. In NIPS, pages 1433–1440.

[18] Biau G. and Bleakley, K. (2006). Statistical Inference on Graphs. Statistics & Decisions,

[19] Papa G., Bellet A., Clémençon S. (2016) On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability, NIPS