The English-speaking candidacies are accepted

Fiche Stage

Intitulé du Stage

Décomposition des classes et constructions de variables pour le classifieur naïf de bayes

Mission:

Le contexte général du stage est la classification supervisée multi-classes avec pour classifieur un Selective Naive Bayes (SNB) MODL [1].

Dans ce contexte le thème du stage porte la volonté d'améliorer les performances du classifieur (SNB) dans le cas où les données sont sous forme monotable (tabulaire). Deux axes d'études seront considérés :

- le premier axe concerne la transformation de la variable à prédire à l'aide de la décomposition des classes [2][3]. On cherchera à reproduire et à vérifier les résultats de l'article scientifique. On étudiera si les gains affichés dans le cas du naïve bayes utilisé dans l'article se vérifient pour le SNB
- le deuxième axe concerne l'ajout de variables explicatives dans le vecteur d'entrée présenté au SNB. Dans certains challenges [4] récents la construction de variables [5] a été un élément clef pour le ranking au sein des résultats des classifieurs employés. Au cours du stage on s'intéressera par exemple à la construction de variables à l'aide d'une technique de coclustering sans écarter d'autres possibilités.

Profil : Data scientist avec goût pour l'informatique et les mathématiques appliquées

Contact : Vincent Lemaire (vincent.lemaire@orange.com)

http://vincentlemaire-labs.fr/

Pour postuler merci de fournir :

- **un CV détaillant vos compétences en mathématiques, data science et informatique**
- **une lettre de motivation**
- **un rapport rédigé par vous dans le passé (idéalement rapport de stage ou de projet)**
- **le contenu de vos cous M1, M2 et les notes correspondantes**
- **une lettre de recommandation si possible**

Références:

[1] "Compression-Based Averaging of Selective Naive Bayes Classifiers", M. Boullé. Journal of Machine Learning Research, 8:1659-1685, 2007

[2] "Class decomposition via clustering: A new framework for low-variance classifiers." R. Vilalta, M.-K. Achari, and C. F. Eick, in International conference on Data Mining (ICDM), 2003

[3] "An empirical study of the suitability of class decomposition for linear models: When does it work well?" F. Ocegueda-Hernandez and R. Vilalta, in SDM SIAM, 2013, pp. 432–440

[4] (IJCRS'15 Data Mining Competition) Janusz A. et al. (2015) Mining Data from Coal Mines: IJCRS'15 Data Challenge. In: Yao Y., Hu Q., Yu H., Grzymala-Busse J. (eds) Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing. Lecture Notes in Computer Science, vol 9437. Springer, Cham

[5] "Feature Construction Methods: A Survey", Parikshit Sondhi

The English-speaking candidacies are accepted

Profil:

- Le profil souhaité est BAC + 5, Master Industriel informatique et/ou statistiques ou école d'ingénieur.
- Intérêt pour les aspects applicatifs et théoriques du sujet.

Compétences

- Le profil souhaité est BAC + 5, Master Industriel informatique et/ou statistiques ou école d'ingénieur.
- Intérêt pour les aspects applicatifs et théoriques du sujet.
- Les connaissances en Matlab (ou équivalent) et **Python** seraient les bienvenues.
- Des connaissances en **statistiques**, mathématiques et/ou apprentissage statistique sont un plus.

Modalités

5 ou 6 mois, printemps-été 2016, Lannion (Bretagne)
Rémunération : de l'ordre de 1000€/mois

Le plus de l'offre

Proche de la mer, vous serez dans l'équipe de traitement des données d'Orange Labs directement en lien avec des problématiques opérationnelles du groupe Orange sur le CRM et l'Audience. Le stagiaire évoluera dans un contexte très recherche sur un sujet très porteur. Il sera intégré dans l'URD au sein d'une équipe recherche. Rémunération de l'ordre de 1000€ net.

Contacts

Vincent Lemaire – vincent.lemaire@orange.com
http://www.vincentlemaire-labs.fr/

The English-speaking candidacies are accepted