Journée Thématique « Modèles Graphiques »

organisée par le GRCE – Groupe de Recherche en Communication Ecrite

Date : Lundi 20 Juin 2011
Lieu : Telecom ParisTech – amphithéâtre B312
46 rue Barrault, 75013 Paris

Les modèles graphiques sont un outil de modélisation qui permet de représenter dans un formalisme unifié les liens entre les données ou entre différentes sources de connaissances. Cette journée fera le point sur les modèles graphiques récents : réseaux Bayesiens et champs de Markov et nous les étendrons aux réseaux de neurones récurrents et réseaux profonds. Nous présenterons les applications de ces modèles dans le domaine de l'analyse de documents et de la reconnaissance de l'écriture.

Organisateurs

Laurence Likforman-Sulem (Institut Telecom/Telecom ParisTech)
Laurent Heutte (Univ. de Rouen/LIRIS)
Laurent Wendling (Univ. Paris Descartes/LIPADE)

Programme

9h45-10h : accueil

10h-11h : Une introduction aux modèles graphiques probabilistes
Philippe Leray (Ecole Polytechnique de l'université de Nantes)

11h-12h : Tutorial « Réseaux de neurones récurrents » (présentation en anglais)
Marcus Eichenberger-Liwicki, (German Research Center for Artificial Intelligence)

12h-12h45 : Approche Markovienne discriminante pour l'analyse de séquences
Thierry Artières (Université Pierre et Marie Curie, LIP6)

12h45-14h15 pause déjeuner

14h15-15h : Machines de Boltzmann Restreintes pour la classification de documents, Jérôme Louradour (A2iA)

15h-15h30 Modèles HMM en contexte pour la reconnaissance de l'écriture cursive.
Anne-Laure Bianne-Bernard (Telecom ParisTech et A2iA)

15h30-16h Modélisation hiérarchique par Champ Aléatoire Conditionnel pour l'analyse de structure de documents
Florent Montreuil (Université Paris Descartes)

16h-16h30 Champ Aléatoire Conditionnel linéaire pour l'extraction de structures dans les journaux d'archives
David Hebert (LITIS, Rouen)

Résumé des contributions

Titre : Une introduction aux modèles graphiques probabilistes (Philippe Leray).

A la croisée de la théorie des graphes et de celles des probabilités, les modèles graphiques probabilistes (MGP) sont des outils de modélisation de connaissance permettant de représenter de manière concise une distribution de probabilité jointe. Le but de cette présentation est d'introduire la notion de MGP, des modèles non dirigés (Champs de Markov) aux modèles dirigés (Réseaux bayésiens) en passant par des modèles hybrides (Chain Graphs). Nous donnerons aussi un certain nombre de pointeurs vers les algorithmes de référence permettant de construire de tels modèles, puis d'effectuer de l'inférence probabiliste.

Titre : Tutorial « Réseaux de neurones récurrents » (Marcus Eichenberger-Liwicki)

In this presentation the recently introduced Bidirectional Recurrent Neural Networks will be described.They allow for a direct recognition of raw pixel data. In this talk I will speak about the mathematical background and successful applications of this novel NN-architecture. Promising experimental results on various datasets from different countries are presented. A toolkit implementing the networks is freely available for public.

Titre : Machines de Boltzmann Restreintes pour la classification de documents (Jérôme Louradour)

L'exposé porte sur les Machines de Boltzmann Restreintes (RBM en anglais) et leur application à des problèmes de classification de documents écrits. Les RBM sont des modèles probabilistes graphiques à l'origine du succès de l'apprentissage de réseaux profonds, grâce à des techniques d'apprentissage efficaces (Hinton, 2006). Aussi les RBMs en tant que tels (réseaux non profonds) constituent déjà des classifieurs très compétitifs pour les problèmes à très haute dimension et

parcimonieux, comme la classification de texte. Nous présentons une variante des RBM adaptée à la classification, pouvant être entraînée de manière discriminative, générative ou hybride. Nous introduisons aussi une nouvelle extension adaptée à la classification d'ensembles de taille variable, et l'appliquons à la classification de courriers en tant qu'ensembles d'un nombre variable de pages. Des expérimentations comparant ces deux variantes avec d'autres méthodes de classification état de l'art (SVM et AdaBoost) montrent que l'approche par RBM est prometteuse.

Titre : Modèles HMM en contexte pour la reconnaissance de mots cursifs. (Anne-Laure Bianne-Bernard)

Nous présentons un système à base de HMMs pour la reconnaissance hors-ligne de mots manuscrits. Afin de prendre en compte les liaisons entre caractères, leurs modèles sont considérés dépendants de leur contexte, passant alors de monographes à trigraphes. Une telle modélisation augmente de manière considérable le nombre de modèles en contexte à estimer. Pour pallier ce problème, nous effectuons un partage des paramètres par un clustering sur chaque position d'état des trigraphes associés à un même monographe. Ce clustering est basé sur des arbres de décision dont les critères correspondent à des propriétés morphologiques des caractères. Nous présentons nos résultats sur les bases publiques Rimes et IAM.

Titre : Modélisation hiérarchique par Champ Aléatoire Conditionnel pour l'analyse de structure de documents (Florent Montreuil)

Le travail présenté est relatif à l'implémentation d'un modèle hiérarchique de champ aléatoire conditionnel (CAC) dans le contexte de l'analyse d'images de documents. Ces modèles CAC permettent de prendre en compte la variabilité et de contextualiser les connaissances, tout en bénéficiant des avantages apportés par les techniques d'apprentissage. De plus, la modélisation hiérarchique permet de lier simultanément l'adjacence entre les régions et la hiérarchie des régions du document. Cela permet de créer un modèle unifié combinant différentes sources d'informations : locales et globales ; bas et hauts niveaux sémantiques ; ... qui sont utilisées pour extraire conjointement les structures physiques et logiques des documents.

Titre : Champ Aléatoire Conditionnel linéaire pour l'extraction de structures dans les journaux d'archives (David Hébert)

Le travail présenté concerne l'utilisation d'un modèle de type Champ Aléatoire Conditionnel (CAC) 1D tel que défini par les auteurs originaux en 2001, pour l'extraction d'entités structurelles dans des images de journaux. La tâche de segmentation 2D est réalisée par un modèle de ligne, permettant ainsi l'utilisation des capacités de ces modèles de séquences qui utilisent un algorithme de décodage optimal. L'utilisation de feature functions binaires tel que définies en 2001 permet une représentation symbolique de données qui n'est pas appropriée aux valeurs numériques que sont les caractéristiques physiques généralement extraites pour la segmentation d'images. La conversions des données est réalisée par une quantification multi-échelle, ne nécessitant pas d'étape d'apprentissage et qui ne prend aucune décision, cette tâche étant exclusivement réalisée par le modèle CAC