Projets cours MALAP 2016

Objectif des projets

Les projets sont à faire en trinôme. Chaque groupe de trois élèves choisit un thème et étudie en détail un article du thème ; il sera généralement utile d'avoir consulté plus d'une référence dans la liste donnée pour chaque thème. Vous devez montrer au travers de votre compte rendu écrit et de la restitution orale que vous avez compris la problématique scientifique et l'apport de l'article dans ce domaine. Il vous est demandé d'expliquer et d'illustrer numériquement les aspects principaux de l'article, ce qui suppose d'implémenter au moins une partie des algorithmes proposés et, le cas échéant, d'exposer les grandes lignes des arguments et formulations clés. Au vu des difficultés rencontrées, le travail peut se focaliser sur un seul aspect de l'article. Il est essentiel néanmoins de mettre en relation les idées rencontrées dans l'article avec les concepts vus en cours. Par ailleurs, dans les expériences numériques que vous réalisez, veillez à faire les comparaisons qui s'imposent avec des méthodes simples ou de références et à vous posez les bonnes questions sur la structure des modèles appris, les choix des hyperparamètres, etc. Un travail d'analyse critique des résultats est attendu au regard des concepts vus en cours.

Présentation orale

La soutenance aura lieu le 2 juin à l'horaire du cours sous la forme d'une session poster. Les modalités exactes vous seront communiquées ultérieurement. Il vous sera demandé de faire une courte présentation orale sur votre projet.

Rapport écrit

Vous devrez rendre un compte rendu de 6 pages A4 (avec éventuellement un appendice d'illustrations complémentaires des expériences d'au plus 4 pages) le 02 juin à soumettre sur Educnet. Si vous travaillez en Latex pourrez ajuster la taille des marges en insérant `\usepackage{fullpage}` dans votre en-tête de façon à avoir des marges standard.

Données

— Un certain nombre d'articles présentent des expériences basées sur des jeux de données référencées dans l'article que vous pourrez utiliser pour vos propres expériences.
— Vous pourrez sinon utiliser des données que vous trouverez sur l'UCI repository http://archive.ics.uci.edu/ml/
— Vous trouverez aussi un certain nombre de base de données open data en cherchant sur le web.

Liste des thèmes

Analyse canonique des corrélations

L'analyse en composantes principales permet de trouver un sous-espace qui explique la plus grande fraction de la variance des données. Elle permet aussi de faire de la réduction de dimension linéaire en

minimisant la distorsion des données. Lorsqu'on a deux types de données différentes comme par exemple une liste de marqueurs génétiques et une liste de susceptibilités à certains médicaments on peut souhaiter analyser la covariance entre ces deux types de données, par exemple trouver la combinations de marqueurs génétiques qui induit la réponse la plus forte aux médicament ou encore trouver une typologie des profils qui expliquent les différentes réponses aux médicament observées. C'est ce que permet de faire l'Analyse Canonique des Corrélations (ACC).

— Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis : An overview with application to learning methods. Neural computation, 16(12), 2639-2664. http://eprints.soton.ac.uk/259225/1/tech_report03.pdf

Régression linéaire bayésienne

L'apprentissage et les statistiques bayésiennes proposent de poser le problème d'apprentissage ou d'estimation comme le calcul d'une distribution sur les paramètres possibles au lieu de produire une valeur unique pour les paramètres. Le but de se projet serait de vous familiariser avec l'approche bayésienne dans le cadre d'un exemple simple : celui de la régression linéaire.

— Des notes de cours sur les statistiques bayésiennes https://www.ceremade.dauphine.fr/~xian/mr081.pdf

— La page wikipédia : http://en.wikipedia.org/wiki/Bayesian_linear_regression

— Thomas P. Minka (2001) Bayesian Linear Regression, Microsoft research web page˙ http://research.microsoft.com/~minka/papers/linear.html

Processus gaussiens

Si la régression à noyaux donne un cadre mathématique élégant pour faire l'apprentissage ou l'estimation d'une fonction non-paramétrique, les processus gaussiens sont d'une certaine manière leur pendant bayésien. Comme les méthodes bayésiennes en général les processus gaussiens permettent d'apprendre une fonction sous la forme d'une distribution sur les fonctions possible. Cette distribution est une généralisation à un espace de fonctions de la gaussienne multivariée. Le but de ce projet sera de comprendre comment fonctionnent les processus gaussiens et de les appliquer à un problème de régression non-linéaire en les comparant à la régression à noyaux.

— Un site dédié avec un certain nombre de ressources : http://www.gaussianprocess.org/

— En particulier, les chapitres 1 et 2 du livre de Carl Rasmussen : http://www.gaussianprocess.org/gpml/chapters/

Réduction de dimension non-linéaire

L'analyse en composantes principales (ACP) permet de proposer une représentation approximative des données dans un espace de dimension plus faible. Elle est bien adaptée s'il parait plausible que les données soient proches d'un sous-espace de dimension faible. L'ACP est limité par le fait qu'elle cherche une transformation linéaire des données. Il existe donc un certain nombre de réduction de dimension non-linéaires. Les plus connues sont l'ACP à noyaux, MDS, Isomap, LLE et le plongement laplacien (connu aussi sous le nom de cartes de diffusion). Le but de ce projet sera de comprendre au moins 2 des ces méthodes et de les appliquer sur des données comme les données MNIST.

— http://en.wikipedia.org/wiki/Nonlinear_dimensionality_reduction

— http://www.math.uwaterloo.ca/~aghodsib/courses/f06stat890/readings/tutorial_stat890.pdf

— Scholkopf, B., Smola, A., & Müller, K. R. (1999). Kernel principal component analysis. In Advances in kernel methods-support vector learning. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.128.7613

— Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 15(6), 1373-1396.
http://www.math.pku.edu.cn/teachers/yaoy/Fall2011/Laplacian.pdf
— MNIST database http://yann.lecun.com/exdb/mnist/

Classification par diffusion

Une des méthodes d'apprentissage les plus simples est la méthode des k-plus proches voisins. Elle suppose néanmoins que les données soient plongées dans un espace métrique. Dans certains cas, plutôt qu'une représentation explicite des données, on a un graphe de similarités entre données. Les arêtes de ce graphe sont typiquement pondérés par les valeurs de similarité. Si on suppose que certaines données sont étiquetées et d'autres non, on a à résoudre un problème de classification dit "transductif". Une façon de l'aborder est de spécifier quelle est la probabilité que deux noeuds soient dans la même classe en fonction de leur similarité et ce simultanément pour tous les noeuds. Pour se faire, on utilise typiquement la théorie des modèles graphique. Un cas particulier est plus simple : celui où on considère un modèle gaussien. Il conduit à une formulation où les étiquettes connues "diffusent" sur le graphe en suivant les même principes que celle de la diffusion de la chaleur. Le but de ce projet sera de comprendre la méthode (éventuellement de faire le lien avec le clustering spectral) et d'appliquer la méthode aux données utilisées dans l'article ou à de nouvelles données.

— Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In The 20th International Conference on Machine Learning (ICML), 2003. ICML 10-Year Classic Paper Prize.
http://pages.cs.wisc.edu/~jerryzhu/pub/zgl.pdf

Le modèle naïf de Bayes

Lorsque les ressources computationnelles ne permettent pas de résoudre un SVM ou une régression logistique, il est utile d'avoir des modèles plus simple à apprendre. C'est le cas du modèle naïf de Bayes qui est quelquefois très efficace. (Contrairement à ce que dit la page wikipédia en français, il ne s'agit pas d'une méthode bayésienne).

— http://en.wikipedia.org/wiki/Naive_Bayes_classifier
— Naive Bayes : A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In AAAI-98 Workshop on Learning for Text Categorization, 1998.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.9324&rep=rep1&type=pdf
— Book Chapter : Naive Bayes text classification, Introduction to Information Retrieval
http://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html

Boosting

Le "Boosting" pour la classification est considéré par certain comme "le meilleur algorithme de classification". Il s'agira ici de comprendre les mécanismes sur lesquels ce type de procédure repose, d'en maîtriser certaines variantes (selon le type de fonction de perte considérée, l'utilisation éventuelle de la randomisation, etc.) et de les mettre en pratique sur données simulées et réelles (on pourra utiliser les données proposées à l'adresse http ://archive.ics.uci.edu/ml/ à cet effet).

— Chapitre 10 de "The Elements of Statistical Learning", T. Hastie, R. Tibshirani et J. Friedman, Springer 2007.
http://statweb.stanford.edu/~tibs/ElemStatLearn/
— Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression : a statistical view of boosting. The Annals of Statistics, 28(2), 337-407.
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/boost.pdf

— Friedman, J. H. (2001). Greedy function approximation : a gradient boosting machine. Annals of statistics, 1189-1232.
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/trebst.pdf

Arbres et Forêts

Deux types de techniques ont été proposées comme développement à partir des arbres de décision : les forêts aléatoires et le boosting d'arbres de décision. Le but de ce projet est de comparer les niveaux de performances obtenu avec ces deux méthodes. On s'appuiera sur

— les chapitres 10 et 15 du livre "Elements of Statistical Learning Theory" de Hastie, Tibshirani et Friedman
http://statweb.stanford.edu/~tibs/ElemStatLearn/ et

— Leo Breiman, Random forests
http://oz.berkeley.edu/~breiman/randomforest2001.pdf

Pairwise coupling et round-robin classification

Dans un certain nombre de contextes, pour faire de la classification multiclasse, il est plus efficace de combiner plusieurs classifieur binaires que d'apprendre directement une régression logistique multi-classe (ou un SVM multi-classe). Le but du projet est de comparer plusieurs façons de combiner les sorties de classifieurs binaires probabilistes pour obtenir un classifieur multi-classe.

— Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. The Journal of Machine Learning Research, 5, 975-1005. http://www.csie.ntu.edu.tw/~cjlin/papers/svmprob/svmprob.pdf

— Fürnkranz, J. (2002). Round robin classification. The Journal of Machine Learning Research, 2, 721-747.

Un algorithme incrémental pour l'apprentissage supervisé à grande échelle

Un certain nombre de problème d'apprentissage se formulent naturellement comme des problèmes d'optimisation. Pour la minimisation du risque empirique régularisé, on utilise classiquement un algorithme de descente de gradient qui calcule le gradient du risque empirique à chaque itération. Mais le calcul de ce gradient nécessite de parcourir toutes les données d'apprentissage à chaque itération. Dans un contexte "Big Data", cela n'est pas efficace, d'autant plus que comme les données viennent de la même distribution, elles sont assez redondantes. Un certain nombre d'algorithme d'optimisation incrémentaux, c'est-à-dire qui ne prennent en considération qu'un petit nombre d'exemple d'apprentissage à la fois, on été proposés ou étudiés dans les dernières années. Le but du projet est de comprendre et d'implémenter l'un d'entre-eux : Stochastic Dual Coordinate Ascent (SDCA)

— Yu, H. F., Huang, F. L., & Lin, C. J. (2011). Dual coordinate descent methods for logistic regression and maximum entropy models. Machine Learning, 85(1-2), 41-75. http://www.csie.ntu.edu.tw/~cjlin/papers/maxent_dual.pdf

Détection d'anomalie et détection de nouveauté

Dans de nombreux domaines d'applications, les techniques d'apprentissage peuvent être utilisées pour vérifier que les nouvelles données sont "normales" au sens où elle suivent la distribution habituelle. Un exemple simple est la surveillance de l'activité sur un réseau informatique, où il s'agit de détecter l'éventuel comportement anormal de certaines machines dues à la présence d'un virus ou d'un bot. On parle de détection de nouveauté quand on ne dispose de données qui soient représentatives du type d'anomalie qui peut survenir où lorsque qu'il peut y avoir des anomalies nouvelles.

Les problèmes d'apprentissage pour la détection d'anomalie et de nouveautés et pour la classification à partir de données d'une seule classe

— Khan, S. S., & Madden, M. G. (2013). One-Class Classification : Taxonomy of Study and Review of Techniques. arXiv preprint arXiv :1312.0049.
http://arxiv.org/abs/1312.0049
— Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., & Platt, J. C. (1999). Support Vector Method for Novelty Detection. In NIPS (Vol. 12, pp. 582-588).
http://users.cecs.anu.edu.au/~williams/papers/P126.pdf
— Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471.
http://research.microsoft.com/pubs/69731/tr-99-87.pdf

Classification sans exemple étiqueté négatif

Comme il est souvent plus facile de démontrer qu'une chose est vraie que de démontrer qu'elle est fausse, dans certains problèmes de classification on est capable de donner des exemples positifs mais on ne peut pas assurer que les autres exemples sont négatifs. Par exemple, on a déterminé que certaines protéines peuvent se lier avec une petite molécule, mais pour les autres on ne sait pas car on n'a pas la possibilité de tester toutes les conditions chimiques dans lesquelles une liaison pourrait se former. Ce problème a en commun avec la détection d'anomalie qu'on ne dispose pas d'exemples étiquetés négatifs. En revanche, comme en apprentissage semi-supervisé on dispose d'exemples non-étiquetés.

— Khan, S. S., & Madden, M. G. (2013). One-Class Classification : Taxonomy of Study and Review of Techniques. arXiv preprint arXiv :1312.0049.
http://arxiv.org/abs/1312.0049
— Blanchard, G., Lee, G., & Scott, C. (2010). Semi-supervised novelty detection. Journal of Machine Learning Research, 11, 2973-3009.
http://jmlr.org/papers/volume11/blanchard10a/blanchard10a.pdf
— Thiran, J. P., Gass, V., Borgeaud, M., Tuia, D., & de Morsier, F. (2013). Semi-Supervised Novelty Detection using SVM entire solution path. IEEE Transactions on Geoscience and Remote Sensing, 51, 1939-1950.
http://infoscience.epfl.ch/record/175357/files/SSNDNCSSVM_demorsier_infoscience.pdf&version=1

Régression logistique à noyau pour la reconnaissance des instruments de musique

On étudiera les extensions du modèle de régression logistique visant à le munir des mêmes atours que les machines à vecteurs de support (SVM) : utilisation de noyaux et régularisation. On clarifiera en particulier la relation entre ces deux modèles. Dans un deuxième temps, on les appliquera à une tâche de reconnaissance automatique des instruments de musique et on comparera leurs comportements vis-à-vis de variations sur le problème de classification (nombre de classes, déséquilibre entre classes, présence d'outliers,...). On pourra travailler avec les données de l'article 2.

— Zhu, J., & Hastie, T. (2002). Support vector machines, kernel logistic regression and boosting. In Multiple Classifier Systems (pp. 16-26). Springer Berlin Heidelberg.
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/kernel-log-regression-svm-boosting.pdf
— Lardeur, M., Essid, S., Richard, G., Haller, M., & Sikora, T. (2009, April). Incorporating prior knowledge on the digital media creation process into audio classifiers. In Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on (pp. 1653-1656). IEEE.
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/MLICASSP-09.pdf

Fonctionnelles de coût pour la factorisation en matrices non-négatives et application à l'estimation de notes de musique

On se propose d ?étudier la factorisation en matrices non-négatives (NMF) du point de vue de la fonctionnelle de coût utilisée (euclidienne, divergence de Kullback-Leibler, ou divergence d ?Itakura-Saito). On s ?intéressera en particulier à son interprétation probabiliste. On testera ensuite la méthode pour l ?estimation de notes de musique dans un mélange polyphonique (tel que décrit dans l ?article 2).

— Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562).
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/nmfconverge.pdf

— Févotte, C., Bertin, N., & Durrieu, J. L. (2009). Nonnegative matrix factorization with the Itakura-Saito divergence : With application to music analysis. Neural computation, 21(3), 793-830.
http://perso.telecom-paristech.fr/~clemenco/Projets_ENPC_files/FevotteBertinDurrieu-2009.pdf

Apprentissage de distance

Dans la plupart des problèmes en apprentissage, les données sont soit représentées comme des vecteurs de descripteurs, soit représentés comme des objets combinatoires (comme des arbres, des graphes, ou des séquences de caractères). Dans tous les cas, il est souvent difficile de dire a priori quelles caractéristiques des données sont celles qui sont les plus pertinentes pour une tâche d'apprentissage donné, quelle est l'importance relative qu'il faut donner à chacune d'entre et comment mesurer correctement la similarité entre les différent descripteurs. Un autre façon de voir le problème est qu'on ne sait pas a priori quelle est la bonne géométrie à considérer, quelle est la bonne mesure de distance entre les données. Les techniques d'apprentissage de la distance ou de la métrique ont pour objet de résoudre ce problème.

— Kulis, B. (2012). **Metric learning : a survey**. Found. and Trends in Machine Learning, 5(4), 287-364.
http://www.cse.ohio-state.edu/~kulis/pubs/ftml_metric_learning.pdf

— Bellet, A., Habrard, A., & Sebban, M. (2013). **A Survey on Metric Learning for Feature Vectors and Structured Data.** arXiv preprint arXiv :1306.6709.
http://arxiv.org/abs/1306.6709

"Learning from the crowd"

Dans un certain nombre de problèmes, on ne dispose pas vraiment de vérité terrain et les étiquettes des données peuvent donc être bruitées ou fausses. Par exemple, pour la détection de tumeurs à partir d'images radiologiques, les meilleurs experts font eux-mêmes des prédictions qui peuvent être divergentes. Lorsqu'on dispose de plusieurs experts et donc de plusieurs étiquetages bruités on peut espérer apprendre à faire la tâche mieux que les experts en apprenant simultanément lesquels on tendance à avoir un avis divergent par rapport à l'opinion générale. On est évidement confronté à ce type problème lorsqu'on utilise du crowdsourcing, car l'erreur est humaine...

— Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, Linda Moy ; 11(Apr) :1297 ?1322, 2010.
http://jmlr.org/papers/volume11/raykar10a/raykar10a.pdf

— Yan Yan, Rómer Rosales, Glenn Fung, Mark Schmidt, Gerardo Hermosillo, Luca Bogoni, Linda Moy, and Jennifer Dy Modeling Annotator Expertise : Learning when Everybody Knows a Bit of Something In Proc. International Conference on Artificial Intelligence and Statistics (AISTATS) 2010.
http://people.csail.mit.edu/romer/papers/AIStatsAnnotExpertise.pdf

Clustering spectral

Pour faire de la classification non supervisée, des algorithmes comme k-means ou les mélanges de Gaussiennes font l'hypothèse que les clusters sont ronds ou bien de forme ellipsoïdale. En pratique, les composantes connexes du support d'une distribution peuvent être beaucoup plus irrégulières. Une des méthodes qui permet d'identifier de telles composantes est le clustering spectral.

— Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and Computing, 17(4), 395-416. http://www.cyberneum.de/fileadmin/user_upload/files/publications/luxburg06_TR_v2_4139[1].pdf

Dropout

Le dropout est une technique introduite en 2012 pour éviter le surapprentissage dans les réseaux de neurones, et qui consiste à "éteindre" certains neurones aléatoirement pendant l'apprentissage.

La technique de dropout peut s'appliquer aussi à des modèles à une seule couche comme la régression logistique ou même la régression linéaire. Elle peut s'interpréter comme une technique de régularisation. Le dropout ressemble en effet un peu à un phénomène connu depuis longtemps et qui est que le fait de bruiter les données d'entrée a un effet de régularisation type ridge (voir le deuxième article). On pourra donc se poser la question de savoir (pour la régression linéaire) ce que fait le dropout en espérance, c'est-à-dire quelle est l'espérance du risque empirique régularisé pour la régression logistique lorsque le dropout est utilisé.

Le but de ce projet serait, de comprendre les idées principales des deux articles suggérés, de comparer la régression ridge, le bruitage Gaussien des données d'entrée et le dropout dans le cas des modèles à une seule couche régression linéaire/logistique et ensuite d'appliquer la technique de droptout à des modèles à plusieurs couches (sans que ce soient des réseaux profonds). Le projet peut être entrepris en Matlab, Theano ou Torch (ce dernier uniquement sous Unix/Linux)

— http://www.cs.toronto.edu/~rsalakhu/papers/srivastava14a.pdf (On ne travaillera pas sur la partie de l'article qui traite des Restricted Boltzman machines)

— http://papers.nips.cc/paper/4882-dropout-training-as-adaptive-regularization.pdf

Auto-encodeurs

Les auto-encodeurs sont des réseaux de neurones qui fournissent des généralisations non-linéaires de l'analyse en composantes principales et des modèles de factorisation de matrice. Le principe général est de construire une version compressée des données qui permette de les reconstruire au mieux. La version compressée correspond à la couche cachée d'un réseau à deux couches et l'entrée et la sortie sont les données à encoder.

Le but du projet est de comprendre les liens existants entre ACP, approximation de rang faible, factorisation de matrice et auto-encodeurs. On pourra implémenter un modèle d'ACP ou de factorisation de matrice et un auto-encodeur et l'évaluer sur des données réelles telles que des images pour les comparer, par exemple sur une tâche de débruitage.

— https://www.cs.toronto.edu/~hinton/science.pdf

— Chapitre 4.6 de http://www.iro.umontreal.ca/~bengioy/papers/ftml_book.pdf

References on matrix factorization (and its relation to PCA)

— https://en.wikipedia.org/wiki/Low-rank_approximation

— A paper on the relation of auto-encoders with PCA :
http://ace.cs.ohiou.edu/~razvan/courses/dl6890/papers/bourlard-kamp88.pdf

— Relation between matrix factorization and PCA :
Sections 1 and 2.1 of https://people.csail.mit.edu/tommi/papers/SreJaa-aim03.pdf

— http://www.niss.org/sites/default/files/tr185.pdf

“Profondeur” en deep learning et transfert de représentation

Le succès du deep learning est dû à la possibilité d'entraîner des réseaux de neurones profonds à partir de bases de données massives, qui apprennent des représentations très non-linéaires mais avec les bonnes propriétés d'invariance.

Le but de ce projet est d'évaluer (a) dans quelle mesure un nombre important de couches est utile (la profondeur du réseau) et (b) dans quelle mesure les représentations apprises, par exemple dans le cadre de la vision, permettent d'apprendre efficacement d'autre tâches visuelles.

Un possibilité serait de travailler sur les données MNIST (http://yann.lecun.com/exdb/mnist/). Le protocole pourrait être le suivant : on entraine un réseau profond à convolution (CNN) avec les données MNIST correspondant aux digits 0 à 4 pour résoudre le problème de classification à 5 classes correspondant. Ensuite dans un deuxième temps on enlève les k dernières couches du réseau appris et soit (a) on réapprend les ces k couches pour résoudre le problème de classification des digits 5 à 10 soit (b) on utilise les k premières couches pour calculer un vecteur de descripteurs que l'on utilise comme données d'entrées dans un SVM linéaire. Dans les deux cas l'idée est de tester si les k premières couches du réseau on permis d'apprendre une représentation des données visuelle qui est pertinente pour une tâche proche et à quel niveau dans le réseau la représentation n'est pas encore trop spécialisée pour la tâche choisie initialement pour l'apprentissage. Il faudra se poser la question du nombre de couches et de leur structure (convolution, max-pooling, etc) On pourra travailler avec Theano et s'appuyer sur http://deeplearning.net/tutorial/.

Un autre possibilité est de s'appuyer sur un réseau profond déjà appris avec de grandes quantités de données (et qu'on ne pourrait pas apprendre entièrement dans le cadre d'un projet de cours). De tels réseaux sont disponibles dans Torch http://torch.ch/ qu'il faudra préalablement installer (Attention : l'installation de Torch n'est pas possible sous Windows et requiert de correctement installer et lier un certain nombre de librairies sous Linux/Unix. A n'entreprendre que si vous êtes prêt à faire face aux difficultés associées.). Ensuite seulement les premières couches du réseau appris sont gardées et elles sont utilisées pour calculer une représentation non-linéaire des données. Un problème de classification d'image est ensuite considéré avec des données suffisamment différentes des données qui ont été utilisées pour l'apprentissage du réseaux. Un classifieur multi-classe est appris à partir de la représentation fournie par le réseau tronqué en utilisant soit (a) des combinaisons de SVM ou de régression logistiques binaires, soit (b) une régression logistique multiclasse, soit encore (c) un réseau de neurones multicouches.

Le travail fourni devra analyser le rôle des différentes couches et dire quel est l'influence du nombre de couches utilisées pour l'apprentissage, du nombre de couches gardées pour le transfert, quel est le rôle de la structure du réseau, etc.

— A web page with references on deep learning.
http://deeplearning.net/tutorial/
— Bengio, Y. (2009). Learning deep architectures for AI. Foundations and trend in Machine Learning, 2(1), 1-127.
http://www.iro.umontreal.ca/~bengioy/papers/ftml.pdf
— Bengio et al. Representation Learning : A Review and New Perspectives
http://arxiv.org/pdf/1206.5538v2.pdf