PROJET IA COMPUTATIONNELLE

Application des méthodes d'apprentissage automatique

Objectifs

Ce projet destiné aux élèves ingénieurs GSI et MI en deuxième année a comme objectif de permettre l'approfondissement de la matière de l'intelligence artificielle computationnelle et, plus généralement, de permettre à l'élève de se confronter à un problème technique avec des données réelles, non simulées.

Le projet est composé de trois sous-projets :

1. application des réseaux de neurones artificiels
2. application des machines à vecteur de support.
3. synthèse des résultats de deux sous-parties précédentes et conclusion générale.

Outils & logiciels utilisés

Pour la réalisation du projet, les élèves utiliseront des logiciels de leur choix. L'École vous propose JavaNNS (ou SNNS) pour les réseaux des neurones, en association avec le programme IACompu en Scilab pour la partie analyse de données et calcul de l'influence des entrées aux réseaux sur les résultats.

Pour les SVM, on propose libSVM soit comme librairie de Scilab, soit comme librairie de Weka.

Étapes de réalisation

Le projet se fera en groupe de trois personnes. Chaque groupe choisira un ensemble de données relatif à un domaine d'expertise précis. Nous décrivons dans la suite les différentes parties du projet.

Partie étude et analyse des données

La **première partie** du projet comportera les étapes suivantes :

1. ***Analyse Préliminaire du problème et des besoins*** : est-ce un problème de classification ou de régression ?
2. ***Etude approfondie des données*** : Quelle est la nature des données ? S'agit-il des valeurs qualitatives ou quantitatives, discrètes ou continues ? Cette étude permettra d'avoir plus de précision sur le choix des architecture et des paramètres à utiliser.
3. ***Pré-traitement des données*** : cette étape est composée de plusieurs phases qui permettent de préparer les données aux différentes méthodes envisagées, à savoir :

(a) examen la structure des données et élimination les points qui sont aberrants ou ont des données manquantes ;

(b) binarisation des sorties des réseaux des neurones ;

(c) normalisation les données d'entrée ;

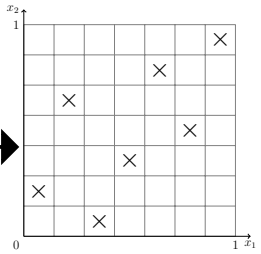
(d) partage en trois sous-ensembles (apprentissage, test et validation).

Partie apprentissage

Cette partie contient les étapes suivantes :

4. ***Choisir la méthode de validation :*** pour pouvoir comparer les résultats des différentes méthodes d'apprentissage utilisées, il faut se fixer un ensemble d'apprentissage, un ensemble de test et un ensemble de validation. L'ensemble d'apprentissage permet de construire le modèle. L'ensemble de test permet de tester le modèle déjà construit sur des nouvelles données. Enfin l'ensemble de validation permet de valider le modèle en utilisant des données qui n'ont pas participer directement ou indirectement à l'élaboration du modèle.
5. ***Choix des paramètres des algorithmes d'apprentissage*** : un choix judicieux des paramètres des algorithmes d'apprentissage nécessite un nombre important d'essais, consistant à faire varier les valeurs de ces paramètres pour la construction d'un modèle à partir d'une base d'apprentissage. La variation des paramètres a pour objectif de trouver les valeurs permettant d'obtenir les meilleurs résultats en terme de pourcentage d'erreur. Ce dilemme est par exemple traité via une variation itérative des valeurs des paramètres selon un pas constant par l'algorithme `adaboost`. Faire varier les paramètres selon un pas constant engendre un nombre considérable d'essais quand le nombre de paramètres augmente. Par exemple, si 3 paramètres varient sur 10 valeurs, cela fait 10^3 combinaisons de valeurs possibles. Nous proposons donc deux autres méthodes pour déterminer les variations des paramètres à tester basée sur des hypercubes latins. Il s'agit de faire varier les paramètres des algorithmes d'apprentissage de manière homogène en choisissant pour valeurs chaque ligne de l'hypercube latin. Nous l'appliquerons soit directement aux paramètres des RNA ou des SVM, soit aux paramètres de départ d'AdaBoost. Ceci permet de diminuer le nombre d'essais tout en gardant le même nombre de valeurs testées pour chaque paramètre. Dans l'exemple, ci-dessus, la variation de 3 paramètres sur 10 valeurs selon un hypercube latin engendre uniquement 10 combinaisons.
6. ***Définition d'un hypercube latin :*** un hypercube latin $LH(n, d)$ est une matrice de taille $n * d$ où d est le nombre de paramètres à faire varier et n est le nombre d'essais telle que chaque colonne est une permutation de $\{0, \cdots, n-1\}$.
   Exemple : $LH(2, 7)$

$$\begin{pmatrix} 0 & 1 \\ 1 & 4 \\ 2 & 0 \\ 3 & 2 \\ 4 & 5 \\ 5 & 3 \\ 6 & 6 \end{pmatrix} \xrightarrow{\text{Normalisation}} \begin{pmatrix} 0 & 1/6 \\ 1/6 & 2/3 \\ 1/3 & 0 \\ 1/2 & 1/3 \\ 2/3 & 5/6 \\ 5/6 & 1/2 \\ 1 & 1 \end{pmatrix}$$



On note que chaque paramètre est testé sur n valeurs différentes (équiréparties sur chaque axe) et qu'il y a en tout n combinaisons de valeurs.

Les hypercubes latins sont disponibles sur le site `www.spacefillingdesign.nl`. Une simple transformation permet de ramener les valeurs de l'hypercube latin dans $[0, n-1]$ aux valeurs des paramètres dans $[min; max]$. Par exemple, pour un SVM avec noyau gaussien, on fera varier le paramètre de régularisation C dans $[2^{-5}; 2^{15}]$ et le paramètre γ dans $[2^{-15}; 2^{3}]$.

7. ***Appliquer les différentes méthodes d'apprentissage*** : l'objectif de ce travail consiste à comparer l'algorithme de meta-apprentissage `adaboost` avec la méthode expérimentale basée sur la variation des paramètres d'apprentissage selon un hypercube latin. Chacune des méthodes d'apprentissage devra être appliquée, d'une part sur un réseau de neurones et d'autre part sur un SVM.
8. ***Effectuer une étude comparative entre les résultats*** obtenus par les différentes méthodes et algorithmes utilisés pour votre jeux de données, en appliquant la méthode de validation ci-dessus. En particulier il vous est demandé d'analyser et de commenter les différents résultats en terme de pourcentage d'erreurs.

Partie Conclusion générale

La **troisième partie** du projet est consacrée à une discussion qui explique l'intérêt de chaque méthode utilisée. En particulier et en fonction des résultats obtenus, il faut d'une part présenter les avantages et les inconvénients de ces méthodes tels qu'ils en ressortent de votre étude et, d'autre part, faire des suggestions pour des utilisations futures des ces méthodes.

De plus, on doit répondre, entre autre, aux questions suivantes :

- Y-a-t-il de redondance et/ou de complémentarité entre les différentes méthodes que vous avez choisies ?
- Quels sont les perspectives de votre travail ?
- Pourrait-on envisager le travail autrement pour améliorer certains aspects ? Lesquels ?

Modalités

Chaque groupe doit fournir

- un fichier avec les données utilisées après pré-traitement avec une présentation de la méthode de pré-traitement utilisé. Il faut aussi fournir les trois fichiers des ensembles d'apprentissage, de test et de validation utilisées pour l'apprentissage et le test (cf. supra) ;
- l'architecture des réseaux des neurones et les valeurs des paramètres.
- les valeurs des paramètres pour les machines à support vectoriel ;
- un rapport écrit en latex, qui doit contenir
  - Une description de la réalisation de la première partie du projet ; chaque étape doit être bien expliquée et détaillée. Présentation et discussion des résultats
  - Une description de la réalisation de la deuxième partie du projet ; chaque étape doit être bien expliquée et détaillée. Présentation et discussion des résultats
  - Une synthèse des résultats des première et deuxième parties sous forme de tableaux montrant bien la variation des paramètres et leur impact sur la qualité de l'apprentissage :

NB : Si une des parties demandées est absente, la note sera 0 ***(zéro) pour la totalité du projet.***

Probleme	Modele(RN/SVM)	Methode	Erreur	Temps/Iteration	NbVecteur(SVM)
PB1	RN	AdaBoost			
		LHS			
		LHS AdaBoost			
	SVM	AdaBoost			
		LHS			
		LHS AdaBoost			
DIM 10	RN	AdaBoost			
		LHS			
		LHS AdaBoost			
	SVM	AdaBoost			
		LHS			
		LHS AdaBoost			
DIM 20	RN	AdaBoost			
		LHS			
		LHS AdaBoost			
	SVM	AdaBoost			
		LHS			
		LHS AdaBoost			

– Une conclusion finale détaillée sous forme de discussion qui explique l'intérêt de chaque méthode utilisée.

Livrables – Délais

Il y aura un seul livrable qui sera défendu lors d'une soutenance de 15 minutes.
Délai : **lundi 25 mars 2013**.