Examen
Module Apprentissage et Fouille de Données
Master Université Paris-Sud, 2004-2005

28 février 2005

L'examen comprend 3 parties. Vous pouvez nous faciliter la correction en rédigeant les trois parties sur trois copies distinctes (2 points de présentation). Le barême (sur 40 points, à diviser par deux) est indicatif. Il est conseillé de lire toutes les questions avant de commencer.

1 Première Partie : Principes (10pts)

- Structure sur l'espace des hypothèses (4 points) Pourquoi vaut-il mieux avoir une relation d'ordre dans H plutôt que seulement une notion de distance ? À quoi correspond la relation d'ordre utile pour l'induction ? Comment se situe l'apprentissage par réseaux de neurones : utilise-t-il une relation d'ordre ou une notion de distance dans H ?
- Le surapprentissage (3 points) Qu'est-ce que le surapprentissage ? Comment se manifeste-t-il par exemple dans le cas des arbres de décision ? Que fait-on pour y remédier dans ce cas ?
- Validation de l'apprentissage (3 points) Pourquoi l'erreur en test est-elle en général supérieure à l'erreur en apprentissage ? À quoi sert l'ensemble de validation ?

2 Seconde Partie : Programmation Logique Inductive (10pts)

Soit les 6 diagrammes décrits à la figure 1, dits *diagrammes de Bongard*.

1. Pourquoi ne pouvez-vous pas les représenter dans une représentation attribut-valeur classique, où un exemple est représenté par une seule ligne d'un tableau dont les attributs sont fixes? (1pt)

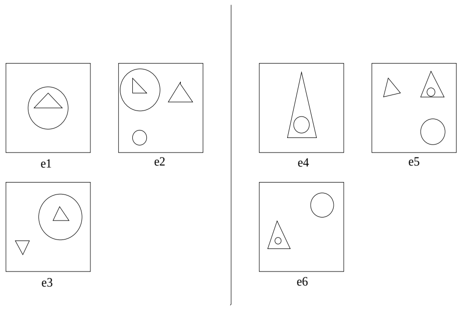

Figure 1: 3 exemples positifs e_1, e_2 et e_3 (à gauche) et 3 exemples négatifs e_4, e_5 et e_6 (à droite)

2. Décrivez les exemples en vous servant *exclusivement* des prédicats $part\_of/2$, $cercle/1$, $triangle/1$, $a\_gauche/2$, $a\_droite/2$, $au\_dessus/2$, $au\_dessous/2$, $a\_l\_interieur/2$. $cercle(X)$ note que X est un cercle, $a\_gauche(X, Y)$ note que l'objet X est à la gauche de l'objet Y.

   (**Indication** : le premier exemple positif se décrit par la clause: $positif(e_1) \leftarrow part\_of(c_1, e_1), part\_of(t_1, e_1), triangle(t_1), cercle(c_1), a\_l\_interieur(t_1, c_1)$, où e_1, t_1 et c_1 sont trois constantes notant respectivement le premier diagramme, le triangle et le cercle qu'il contient). Vous devez décrire les exemples aussi complètement que possible étant donné le vocabulaire qui vous est imposé. Justifiez vos choix. (2pts).

3. Pouvez-vous écrire une clause utilisant uniquement les prédicats précédents, permettant d'exprimer (1pt):

   - le diagramme est un exemple positif si il contient deux objets qui sont des triangles.
   - la diagramme est un exemple positif si il contient au moins un cercle et un triangle.

4. Calculez le moindre généralisé (lgg) sous θ-subsomption des exemples positifs (1,5pts). Réduisez-la (1,5pts). Cette clause couvre-t-elles les trois exemples positifs ? Couvre-t-elle un des exemples négatifs ? (1pt).

5. Décrivez le comportement du système FOIL sur des exemples décrits dans le même vocabulaire (Rappel : par rapport à la représentation précédente, un exemple est formé pour FOIL du seul littéral de tête de la clause et le corps de la clause fait maintenant partie de la théorie du domaine) (2pts)

3 Troisième Partie : Méthodologie

3.1 Théorème des schémas (5 points)

Considérons un espace binaire $\Omega = \{0,1\}^p$. Un schéma est un sous-ensemble de Ω, qui peut être décrit par un élément de $\{0,1,*\}^p$, où $*$ est mis pour 0 ou 1.

Soit $\mathcal{F}$ une fonction (fitness) à maximiser définie sur Ω, et considérons un algorithme génétique canonique (sélection proportionnelle, probabilité p_m de mutation par bit, probabilité p_c de croisement à un point).

Soit $n_t(H)$ le nombre d'individus de la population génétique à la génération t, qui appartiennent au schéma H.

Enoncer le théorême des schémas, fournissant une borne inférieure sur l'espérance de $n_{t+1}(H)$ en fonction de $n_t(H)$.

Discutez-en les implications.

3.2 Codage et opérateurs de variation (5 points)

S'appuyant sur une interprétation du théorême des schémas, il a été recommandé d'utiliser une représentation binaire des réels ou des entiers dans le cadre des algorithmes génétiques.

Discutez le choix de la représentation par rapport aux opérateurs de croisement et de mutation.

Indication : pour chacun des codages, définissez l'action du croisement et de la mutation, et discutez-en les propriétés d'exploration et d'exploitation.

3.3 Mise en situation (10 points)

Vous avez trois mois pour concevoir et mener à bien une application de fouille de données.

- Décrivez l'application que vous choisissez (champ libre : peut aller de la construction d'une stratégie pour War Craft, à l'aide à l'identification d'airs musicaux à partir du sifflotement de l'utilisateur, en passant par la génération automatique de questions d'examen à partir de l'enregistrement des cours de Master).
- Quelles sont les données ? Comment les représenter ?
- Qu'est-ce qu'une solution ?
- Quel est le critère ? (i.e. quels sont les signes qui permettent de déterminer si une solution est bonne).
- Quels algorithmes pourra-t-on utiliser ? (famille d'algorithme).
- Comment intervient l'expert dans le processus ? Quelles pourraient être ses indications ? Comment les prendre en compte ?