9 mai 2007 – Durée 3h

Modalités

Les documents de cours et travaux dirigés ainsi que l'usage de calculatrices sont autorisés. **Aucun** livre ne l'est[1]. Il est recommandé de bien séparer les questions en faisant clairement apparaître leur numéro.

Les durées figurant auprès des titres des exercices sont des durées indicatives quant à la réalisation des exercices concernés. Le respect de ces durées pour la réalisation des exercices n'est bien entendu pas exigé.

1 Cours

1.1 Classification multi-classes – 20 mn

La régression logistique et les machines à vecteurs de support sont des exemples de classifieurs dont la formulation initiale est dédiée *a priori* à la classification binaire (deux classes), à l'inverse, par exemple, des perceptrons multi-couches (pour lesquels on définit généralement autant de neurones de sortie qu'il y a de classes au problème considéré). Il est cependant possible d'utiliser ces méthodes, ainsi que toutes les méthodes de classification binaire dans le cas de la classification multi-classes. En vous basant sur le nom des stratégies utilisées pour réaliser cette tâche, décrivez-en le principe et discutez leurs avantages et inconvénients :

1. *one-versus-all*
2. *one-versus-one*

1.2 Descente de gradient – 20 mn

Expliquez en quelques phrases le principe de l'optimisation par descente de gradient. Pour illustrer votre réponse, considérez la fonction f définie par :

$$f(x) = x^2 - 5x + 1$$

et détaillez les 4 premières itérations du processus de descente de gradient en prenant comme valeur initiale $x_0 = 0$ et comme pas de gradient (ou pas d'apprentissage) $\eta = 0.2$.

2 Arbres de décision – 1 h

Dans cet exercice on utilisera le **critère d'entropie** pour la construction des arbres de décision. On rappelle qu'étant donné une distribution de probabilité $p_1, \ldots, p_n$ définie sur n modalités, l'entropie de cette distribution est

$$Ent(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

où $\log_2$ est le logarithme en base 2 (qui est donc défini par $\log_2(x) = \ln(x)/\ln(2)$).

Exemple : si l'on demande à 10 personnes de choisir leur couleur préférée parmi les $n = 3$ couleurs `rouge`, `bleu` et `vert` et que l'on obtient la répartition des réponses selon le tableau suivant

`rouge`	`bleu`	`vert`
5	3	2

alors l'entropie de la distribution des couleurs préférées est

$$\begin{aligned} Ent\left(\frac{5}{10}, \frac{3}{10}, \frac{2}{10}\right) &= -\frac{5}{10}\log_2\frac{5}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{2}{10}\log_2\frac{2}{10} \\ &= 1.48 \end{aligned}$$

(On remarquera que dans le cours et en travaux dirigés nous avons travaillé avec $n = 2$ puisque les problèmes de classification considérés étaient binaires, i.e. impliquaient deux classes.)

1. On se propose de construire un arbre de décision à partir du tableau 1 (page 2). Ce tableau répertorie les résultats de différents sportifs (pratiquant des sports éventuellement différents) décrits selon l'importance plus ou moins grande de l'épreuve à laquelle ils ont participé, de leur statut de favori ou non dans cette épreuve et de leur nationalité.

[1] Excepté, éventuellement, les dictionnaires de traduction

Épreuve	Statut	Nationalité	Résultat
majeure	favori	francais	defaite
majeure	outsider	francais	defaite
intermediaire	outsider	francais	victoire
intermediaire	favori	francais	defaite
mineure	favori	francais	defaite
majeure	favori	autre	victoire
majeure	outsider	autre	defaite
mineure	favori	autre	victoire
intermediaire	outsider	autre	defaite
intermediaire	favori	autre	victoire

Tab. 1 – Résultats des sportifs

(a) Construire un arbre de décision correspondant à ce jeu de données permettant d'estimer le résultat `victoire` ou `defaite` d'un sportif (rappel : utiliser le critère d'entropie). Fournir et expliquer les détails des calculs.

(b) Expliquer ce qu'est une feuille pure. Dans quelle situation est-il impossible de construire un arbre n'ayant que des feuilles pures ? Proposer une instance qui, ajoutée à l'ensemble des données du tableau 1 (page 2) conduit à cette situation.

2. Supposons qu'une feuille d'un arbre de décision ne soit pas pure ; comment classifier une instance dont la description conduit à cette feuille ? Proposer une méthode (très simple) pour quantifier la confiance qu'on peut avoir dans la classification d'une instance qui se fait à partir d'une feuille impure (indice : cette confiance prendra une valeur entre 0 et 1) ?
3. Quelle sera le résultat d'Alimee Muasermo, au tournoi de tennis majeur de Roland Garros, sachant que c'est une joueuse française qui sera favorite de la compétition ?
4. **Une question sur votre capacité de généralisation !** Dans cette question il est demandé de construire un arbre de décision pour un problème de classification à trois classes. Plus précisément, le but est de construire un arbre de décision à partir des données du tableau 2 (page 3) permettant de déterminer la nature de lentille de contact (`aucune`, `souple`, `dure`) qu'un individu peut porter en fonction de caractéristiques physiologiques (âge, type d'anomalie de la vue, présence d'astigmatie, production de larmes). Construire l'arbre de décision associé à cet échantillon d'apprentissage.

3 Noyaux

3.1 Noyaux polynomiaux – 20 mn

Dans cet exercice, on désigne par $\langle \cdot, \cdot \rangle_d$ l'application

$$\langle \cdot, \cdot \rangle_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$$
$$(\mathbf{u}, \mathbf{v}) \mapsto \langle \mathbf{u}, \mathbf{v} \rangle_p = \sum_{i=1}^{d} u_i v_i$$

1. Soit le noyau k_2 défini par
$$k_2(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle_2^2$$
Proposez une application $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ telle que
$$k_2(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_3$$
2. Soit le noyau k_p défini par
$$k_p(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v} \rangle_2^p$$
Proposez une application $\phi_p : \mathbb{R}^2 \to \mathbb{R}^q$ telle que
$$k_p(\mathbf{u}, \mathbf{v}) = \langle \phi_p(\mathbf{u}), \phi_p(\mathbf{v}) \rangle_q$$
Vous déterminerez q et donnerez par exemple la valeur de la i-ème coordonnée du vecteur $\phi_p(\mathbf{u})$.

On rappelle que, selon la formule du binôme de Newton : $(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}$.

Age	Type anomalie	Astigmatie	Production larmes	Prescription
jeune	myope	non	réduite	aucune
jeune	myope	non	normale	souple
jeune	myope	oui	réduite	aucune
jeune	myope	oui	normale	dure
jeune	hypermétrope	non	réduite	aucune
jeune	hypermétrope	non	normale	souple
jeune	hypermétrope	oui	réduite	aucune
jeune	hypermétrope	oui	normale	dure
adulte	myope	non	réduite	aucune
adulte	myope	non	normale	souple
adulte	myope	oui	réduite	aucune
adulte	myope	oui	normale	dure
adulte	hypermétrope	non	réduite	aucune
adulte	hypermétrope	non	normale	souple
adulte	hypermétrope	oui	réduite	aucune
adulte	hypermétrope	oui	normale	aucune
senior	myope	non	réduite	aucune
senior	myope	non	normale	aucune
senior	myope	oui	réduite	aucune
senior	myope	oui	normale	dure
senior	hypermétrope	non	réduite	aucune
senior	hypermétrope	non	normale	souple
senior	hypermétrope	oui	réduite	aucune
senior	hypermétrope	oui	normale	aucune

Tab. 2 – Prescription de lentilles

3.2 Somme de noyaux de Mercer – (à faire juste avant l'exercice facultatif) 20 mn

1. Etant donné un échantillon $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ d'exemples de $\mathcal{X}$, dire ce qu'est la matrice de Gram K d'un noyau de Mercer k défini par rapport à $\mathcal{S}$ (taille de cette matrice, terme général). Donner une propriété caractéristique de cette matrice.
2. Soit k_1 et k_2 deux noyaux de Mercer définis sur $\mathcal{X} \times \mathcal{X}$.
   (a) Etant donné un échantillon $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\}$ d'exemples de $\mathcal{X}$, donner la matrice de Gram associée à la fonction $k_1 + k_2$.
   (b) Montrer que puisque k_1 et k_2 sont des noyaux de Mercer, $k_1 + k_2$ est également un noyau de Mercer.

4 Perceptron – 40 mn

Soit l'ensemble d'apprentissage

$$\mathcal{S} = \left\{ \left(\mathbf{x}_1 = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, +1 \right), \left(\mathbf{x}_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, -1 \right), \left(\mathbf{x}_3 = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, +1 \right), \left(\mathbf{x}_4 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, -1 \right) \right.$$

On rappelle l'algorithme d'apprentissage du perceptron présenté en cours :
- Classification binaire, $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$, $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$
- Initialisation $\mathbf{w} = \mathbf{0}$
- Répéter jusqu'à convergence ou bien atteinte d'un nombre max d'itérations
  - pour tous les exemples $(\mathbf{x}_p, y_p)$ faire
    - si $\sigma(\mathbf{w} \cdot \tilde{\mathbf{x}}_p) = y_p$
      ne rien faire
    - sinon
      $\mathbf{w} \leftarrow \mathbf{w} + y_p \tilde{\mathbf{x}}_p$

avec $\sigma(x) = 1$ si $x > 0$ et $\sigma(x) = -1$ sinon.

1. Représenter sur le plan les points d'apprentissage $\mathbf{x}_1, \ldots, \mathbf{x}_4$
2. Peut-on espérer que l'algorithme du perceptron fournisse une solution au problème d'apprentissage donné ? Justifier.

3. Au lieu de considérer les vecteurs $\mathbf{x}_1, \ldots, \mathbf{x}_4$, nous allons appliquer l'algorithme du perceptron sur leurs transformations $\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_4)$ où ϕ est l'application de $\mathbb{R}^2$ dans $\mathbb{R}^3$ telle que

$$\phi\left(\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}\right) = \begin{bmatrix} u_1^2 \\ u_2^2 \\ u_1 u_2 \end{bmatrix}.$$

   Donner les coordonnées des images des $\mathbf{x}_i$ par ϕ.

4. Dans la description de l'algorithmes, les vecteurs $\tilde{\mathbf{x}}_i$ sont obtenus en ajoutant aux vecteurs $\mathbf{x}_i$ une composante fixée à 1. Quelle est l'utilité de cette composante ?
5. Appliquer l'algorithme du perceptron aux données transformées (moins d'une dizaine de *mises à jour* sont nécessaires).
6. En déduire une expression analytique, sous la forme d'une forme quadratique, de la surface de décision ainsi générée (cette expression qui doit être du type $f(x, y) = 0$ vous rappellera les equations de coniques vues en terminale).
7. Soit g le classifieur que vous avez obtenu par l'apprentissage précédent. Pour mesure l'erreur de généralisation de ce classifieur, une stratégie est d'utiliser m nouveaux exemples de la distribution qui a permis de générer les exemples d'apprentissage et de mesurer le nombre d'erreurs que g fait sur ces m exemples. Plus m est grand plus la proportion d'erreurs mesurées est proche de l'erreur de généralisaton, c'est-à-dire la probabilité p de se tromper sur un exemple nouveau. Si on considère que le fait que g fasse une erreur sur un exemple tiré au hasard est la réalisation d'une variable de Bernoulli de paramètre p, combien faut-il d'exemples m pour estimer avec une précision ϵ et une confiance $1 - \alpha$ l'erreur de généralisation de g, c'est-à-dire p. Ce nombre d'exemple s'écrira en fonction de p, ϵ et α.

5 Perceptron linéaire (facultatif)

Etant donné n entrées booléennes, $x_1, \ldots, x_n$ et un entier i entre 0 et n, construire un perceptron à fonction d'activation linéaire à seuil qui retourne 1 si et seulement si le nombre d'entrées égales à 1 est supérieur ou égal à i (il y aura bien sûr une cellule biais). Déduire de cette construction un perceptron à fonction d'activation linéaire à seuil possédant une couche cachée de n neurones et calculant la fonction *parité*, égale à 1 si le nombre d'entrées égales à 1 est pair et égale à 0 sinon.