1^{er} juillet 2005 – 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 notamment 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 – 50 mn

1. En quoi consiste la catégorisation de textes ? Donnez un exemple.
2. Quelle est le phénomène étonnant (allant à l'encontre de la théorie) que l'on peut constater lorsqu'on utilise le classifieur naïf de Bayes (*Naive Bayes*) en catégorisation (ou classification) de textes ?
3. Expliquez en quelques phrases le principe de l'optimisation par descente de gradient. Pour illustrer votre réponse, considérez la fonction f définie par :

   $$f(x) = 2x^2 - 7x + 1$$

   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) $\eta = 0.1$.
4. Que sont les mesures de *rappel* et *précision* ?

2 Arbre de décision – 1h

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 $\log_2(x) = \ln(x)/\ln(2)$).

Une société a réalisé une campagne publicitaire en envoyant des prospectus à plusieurs familles qui sont décrites par leur lieu d'habitation (banlieue,ville,campagne), leur type de logement (villa,maison,appartement), leur niveau de revenu (élevé,faible) et le fait que cette famille soit déjà cliente ou non de cette société. Cette société désire être capable de déterminer en fonction de ces caractéristiques si une famille est susceptible de répondre positivement (classe +) ou non (classe −) à la campagne publicitaire. Les résultats de cette campagne auprès de 14 familles sont donnés dans le tableau 1 (page 2).

1. Construire un arbre de décision correspondant à ce jeu de données et permettant d'estimer la positivité de la réponse d'une famille à la campagne publicitaire (rappel : utiliser le critère d'entropie). Fournir et expliquer les détails des calculs.
2. Supposons qu'on ajoute à l'ensemble d'apprentissage précédent une famille décrite par

Lieu	**Type**	**Revenu**	**Client**	**Réponse**
banlieue	villa	élevé	non	+

   (a) Est-il alors possible de déterminer un arbre de décision ne faisant aucune erreur d'apprentissage ? Justifiez votre réponse.

   (b) Comment intégrer ce nouvel exemple d'apprentissage à l'arbre appris auparavant ? Comment prendre en compte l'impureté des feuilles pour la classification de nouveaux exemples ?

[1] Excepté les dictionnaires de traduction

Lieu	**Type**	**Revenu**	**Client**	**Réponse**
banlieue	villa	élevé	non	−
banlieue	villa	élevé	oui	−
campagne	villa	élevé	non	+
ville	maison	élevé	non	+
ville	maison	faible	non	+
ville	maison	faible	oui	−
campagne	maison	faible	oui	+
banlieue	appartement	élevé	non	−
banlieue	maison	faible	non	+
ville	appartement	faible	non	+
banlieue	appartement	faible	oui	+
campagne	appartement	élevé	oui	+
campagne	villa	faible	non	+
ville	appartement	élevé	oui	−

TAB. 1 – Résultats de la campagne publicitaire

3 Noyaux − 30mn

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 : $(a+b)^n = \sum_{k=0}^{n} C_n^k a^k b^{n-k}$.

4 Méthodes de séchage [Daudin et al., 2001] − 40 mn

On a obtenu les données du tableau 2 (page 2) pour deux méthodes différenes de séchage du maïs. On suppose que les taux de séchage sont normalement distribués et que les écart-types de ces taux sont égaux.

1. Calculez les moyennes $\bar{x}$ et $\bar{y}$ des taux de séchage en utilisant les méthodes avec et sans préchauffage.
2. Calculez les variances empiriques s_x^2 et s_y^2 de ces méthodes.
3. Calculez la variance empirique commune s^2 des deux échantillons. Rappel : utilisez $s^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$, où n_x et n_y sont respectivement le nombre de mesures faites pour l'une et l'autre méthode de séchage (ici $n_x = n_y = 5$).

Avec préchauffage	Sans préchauffage
16	20
12	10
22	21
14	10
19	12

TAB. 2 – Comparaison de deux méthodes de séchage

4. En utilisant un test de Student (*t-test*) au risque 5%, peut-on conclure que l'une des deux méthodes de séchage est plus efficace que l'autre ?
5. Même question en utilisant la méthode des couples (*paired t-test*) ?

Références

[Daudin et al., 2001] Daudin, J.-J., Robin, S., and Vuillet, C. (2001). *Statistique inférentielle – idées, démarches, exemples*. Presses universitaires de Rennes.