1 Classifieurs des k-plus-proches-voisins

Soit $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell)\}$ un ensemble d'apprentissage avec $\mathcal{Y} = \{-1, +1\}$. La classification par méthode des k-plus-proches-voisins ou k-ppv se fait de manière très simple : étant donné un exemple de test $\mathbf{x}$, on affecte à $\mathbf{x}$ la classe la plus souvent représentée parmi les k exemples de $\mathcal{S}$ qui lui sont le plus proches (au sens d'une distance définie sur $\mathcal{X}$, l'espace des instances).

1. Quel problème peut-on rencontrer lorsque k est pair ?
2. Quelle est la complexité de classification d'un exemple par k-ppv ?
3. Proposer une variante de k-ppv prenant en compte les distances des plus proches voisins.

2 Descente de gradient

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

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

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) $\eta = 0.05$.

3 Arbre de décision (1)

En utilisant l'index de Gini ou l'entropie de Shannon, construire l'arbre de décision associé à cet ensemble d'apprentissage [**?**]. Détailler chaque étape de l'apprentissage.

Exemple	Prévision	Température	Humidité	Vent	Tennis
1	soleil	élevée	haute	faible	non
2	soleil	élevée	haute	fort	non
3	nuage	élevée	haute	faible	oui
4	pluie	moyenne	haute	faible	oui
5	pluie	basse	normale	faible	oui
6	pluie	basse	normale	fort	non
7	nuage	basse	normale	fort	oui
8	soleil	moyenne	haute	faible	non
9	soleil	basse	normale	faible	oui
10	pluie	moyenne	normale	faible	oui
11	soleil	moyenne	normale	fort	oui
12	nuage	moyenne	haute	fort	oui
13	nuage	élevée	normale	faible	oui
14	pluie	moyenne	haute	fort	non

4 Arbre de décision (2)

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 c ampagne auprès de 14 familles sont donnés dans le tableau 1 (page 2).

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

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 ?