

Exercice 5. On considère un problème d'apprentissage supervisé à deux classes dont les données d'apprentissage sont de la forme $Z = (x_i, y_i)_{i \in [1..n]}$ avec $\forall i \in [1..n], x_i \in \mathbb{B}^d, y_i \in \{+, -\}$, où d est le nombre d'attributs binaires d'un exemple et $\mathbb{B} = \{\text{YES}, \text{NO}\}$, les valeurs possibles pour les attributs.

Par exemple, le tableau ci-dessous est un échantillon Z de données relatif aux infections à la COVID 19, extrait de IJCRD 2019. La première colonne indique l'identifiant d'un exemple x (qui comporte trois attributs F, T et R) et la dernière colonne son étiquette y (ici I).

ID	F	T	R	Problèmes respiratoires (R)	Infecté (I)
1	NO	YES	NO	YES	-
2	YES	YES	YES	YES	+
3	YES	YES	NO	NO	-
4	YES	NO	YES	YES	+
5	YES	YES	YES	YES	+
6	YES	NO	YES	YES	-
7	YES	NO	NO	YES	+
8	YES	NO	YES	YES	-
9	NO	YES	YES	YES	+
10	YES	YES	YES	NO	+
11	NO	YES	YES	NO	-
12	NO	YES	YES	YES	-
13	NO	YES	YES	YES	-
14	YES	YES	YES	NO	-

1. Rappeler le principe de l'apprentissage supervisé.

2. Dessiner l'arbre de décision obtenu en considérant successivement et dans l'ordre les attributs F, T et R . Commenter. *On rappelle qu'un arbre de décision est un arbre binaire dont les nœuds internes sont étiquetés par les attributs et les feuilles par $\{+, -\}$. Les fils gauches correspondent à une réponse NO et les fils droits à une réponse YES.*

L'entropie d'un ensemble S d'exemples est définie par :

$$H(S) = -\frac{n_+}{n} \log_2 \left(\frac{n_+}{n} \right) - \frac{n_-}{n} \log_2 \left(\frac{n_-}{n} \right)$$

où n_+, n_- et n désignent respectivement le nombre d'éléments de S dont l'étiquette est $+$, le nombre d'éléments de S dont l'étiquette est $-$ et enfin le nombre total d'éléments de S . Dans le cas où $k = 0$, on prend la convention $k \log_2(k) = 0$. Par exemple, l'entropie de l'ensemble de toutes les données Z ci-dessus est 1.00.

Étant donné un attribut A , on définit le gain de A par rapport à S par

$$G(S, A) = H(S) - \frac{n_{A=\text{YES}}}{n} H(S_{A=\text{YES}}) - \frac{n_{A=\text{NO}}}{n} H(S_{A=\text{NO}})$$

où $S_{A=\text{YES}}$ désigne le sous-ensemble des éléments de S dont l'attribut A est YES et $n_{A=\text{YES}}$ désigne son cardinal, de même pour NO et n désigne toujours le cardinal de S . Par exemple, $G(Z, F) = 0.26$, $G(Z, T) = 0.07$ et $G(Z, R) = 0.26$ (les valeurs données sont approchées au centième).

Si l'on considère le sous-ensemble $Z_{F=\text{YES}}$ des individus qui ont eu de la fièvre, et en supprimant l'attribut F , on obtient le sous-tableau ci-dessous. Le gain d'information de l'attribut T est alors $G(Z_{F=\text{YES}}, T) = 0.20$.

ID	T	Problèmes respiratoires (R)	Infecté (I)
2	YES	YES	+
3	YES	NO	-
4	NO	YES	+
5	YES	YES	+
7	NO	YES	+
8	NO	YES	+
10	YES	NO	+
14	YES	NO	-

3. Calculer le gain d'entropie $G(Z_{F=\text{YES}}, R)$ de l'attribut problèmes respiratoires pour le sous-ensemble des individus qui ont eu de la fièvre.

L'algorithme *Iterative Dichotomiser 3 (ID3)* (Algorithme 1) peut être utilisé pour construire un arbre de décision. Pour l'appel initial, il suffit de prendre l'ensemble de tous les exemples pour S_p et pour S , et l'ensemble de tous les attributs pour D .

4. Indiquer comment compléter l'algorithme 1.



Algorithme 1 : Algorithme ID3.

Fonction ID3(S_p, S, D)

Entrées : S_p sous-ensemble des exemples du nœud parent, S sous-ensemble des exemples à considérer, D

sous-ensemble des attributs à considérer

Sorties : Un arbre de décision

début

 si l'ensemble des exemples S est vide alors

 ↳ retourner

 si l'ensemble A des attributs est vide alors

 ↳ retourner

 si tous les exemples de S ont une même étiquette y alors

 ↳ retourner

 sinon

 Soit $A \in D$ l'attribut de plus grand gain $G(S, A)$.

 Construire l'arbre de racine A et de sous-arbre gauche ID3($S, S_{A=\text{NO}}, D \setminus \{A\}$) et de sous-arbre

 droit ID3($S, S_{A=\text{YES}}, D \setminus \{A\}$).

Exercice 6. On considère le schéma de base de données suivant, qui décrit un ensemble de fabricants de matériel informatique, les matériels qu'ils vendent, leurs clients et ce qu'ils achètent leurs clients. Les attributs des clés primaires des six premières relations sont soulignés.

Production(NonFabricant, Modele)

Ordinateur(Modele, Frequence, Ram, Dd, Prix)

Portable(Modele, Frequence, Ram, Dd, Ecran, Prix)

Imprimante(Modele, Couleur, Type, Prix)

Fabricant(Nom, Adresse, NonPatron)

Client(Nom, Non, Prenom)

Achat(NumClient, NonFabricant, Modele, Quantite)

Chaque client possède un numéro unique connu de tous les fabricants. La relation **Production** donne pour chaque fabricant l'ensemble des modèles fabriqués par ce fabricant. Deux fabricants différents peuvent proposer le même matériel. La relation **Ordinateur** donne pour chaque modèle d'ordinateur la vitesse du processeur (en Hz), les tailles de la Ram et du disque dur (en Go) et le prix de l'ordinateur (en €). La relation **Portable**, en plus des attributs précédents, comporte la taille de l'écran (en pouces). La relation **Imprimante** indique pour chaque modèle d'imprimante si elle imprime en couleur (vrai/faux), le type d'impression (laser ou jet d'encre) et le prix (en €). La relation **Fabricant** stocke les noms et adresses de chaque fabricant, ainsi que le nom de son patron. La relation **Client** stocke les noms et prénoms de tous les clients de tous les fabricants. Enfin la relation **Achat** regroupe les quadruplets (client c , fabricant f , modèle m , quantité q) tels que le client de numéro c a acheté q fois le modèle m au fabricant de nom f . On suppose que l'attribut **Quantite** est toujours strictement positif.

1. Proposer une clé primaire pour la relation **Achat** et indiquer ses conséquences en terme de modélisation.
2. Identifier l'ensemble des clés étrangères éventuelles de chaque table.
3. Donner en SQL des requêtes répondant aux questions suivantes :

- (a) Quels sont les numéros de modèles des matériels (ordinateur, portable ou imprimante) fabriqués par l'entreprise de nom Durand ?
- (b) Quels sont les noms et adresses des fabricants produisant des portables dont le disque dur a une capacité d'au moins 500 Go ?
- (c) Quels sont les noms des fabricants qui produisent au moins 10 modèles différents d'imprimantes ?
- (d) Quels sont les numéros des clients n'ayant acheté aucune imprimante ?

