MASTER MIASHS - ***Apprentissage automatique***

Novembre 2019 - Durée 1h30

Présentation des données et du contexte

Les données sont celles que vous avez traitées en projet. Pour rappel, il s'agit d'une image hyperspectrale observée selon 103 bandes spectrales. A partir de cet image, on construit un tableau dont les lignes correspondent aux pixels appartenant à l'une des catégories suivantes :

codage	catégorie	nombre de pixels
1	asphalte	6631
2	prairie	18649
3	gravier	2099
4	arbres	3064
5	Toles peintes	1345
6	sol nu	5029
7	bitume	1330
8	pavés auto-bloquants	3682
9	ombres	947

Le nombre total de pixels auxquels on s'intéresse est de 42776. On dispose de 103 prédicteurs (V1, V2, ..., V103) correspondant aux 103 bandes spectrales ; la variable cible nommée `type de matériau` attribue à chaque pixel une des catégories précédemment énoncées.

L'objectif de l'étude est de retrouver automatiquement le type de matériau pour chaque pixel en fonction de ses valeurs observées pour les 103 bandes spectrales.

Attention : toute réponse devra être justifiée de façon concise.

Arbre de décision

Une première analyse fournit l'arbre de décision représenté figure 1 en annexe.

1. On s'intéresse au noeud correspondant à la condition `"V13 < 863"` qui partitionne la population de pixels en 2 groupes G_1 et G_2 :
   - G_1 = ensemble des pixels tels que `"V13 < 863"`,
   - G_2 = ensemble des pixels tels que `"V13 >= 863"`.

   (a) Quelle est la variable de segmentation?

   (b) Reportez le tableau suivant sur votre copie puis complérer les cases vides en indiquant pour chaque groupe la répartition des pixels selon le type de matériau :

type de matériau	G_1	G_2
1		
2		
3		
4		
5		
6		
7		
8		
9		

(c) En supposant que l'arbre de décision contienne uniquement ce noeud, quelle serait le type de matériau prédit pour un pixel appartenant au groupe G_1? Même question pour un pixel appartenant au groupe G_2.

(d) On s'intéresse aux indices d'impureté de ce noeud :

i. Pour chacun des groupes G_1 et G_2, calculer le taux d'erreur ainsi que l'indice Gini.

ii. Déduire de ce qui précède le taux d'erreur ainsi que l'indice de Gini associé à ce noeud.

2. Pour différents pixels pris dans un échantillon test, on dispose des observations suivantes :

	V13	V14	V55	V69	V85	V101	V103
pixel 1	1234	2671	679	4045	2309	3598	2713
pixel 2	409	660	2101	341	4405	1893	451
pixel 3	2275	5571	4207	187	936	5904	6706
pixel 4	6227	4059	306	6478	4167	198	3045
pixel 5	971	5071	2570	6103	1825	672	156

Indiquer pour chaque pixel le type de matériau estimé par l'arbre.

3. D'après la figure 2 fournie en annexe, donner la taille de l'arbre que vous choisiriez de construire.

4. Déterminer le taux d'erreur global d'apprentissage à partir des résultats fournis par l'arbre de décision (figure 1). Dans ce but, reproduisez sur votre copie le tableau suivant

Type de matériau								
1	2	3	4	5	6	7	8	9

(a) Compléter ce tableau en indiquant dans chaque cellule le nombre de pixels mal estimées (détailler vos calculs pour une seule cellule),

(b) En déduire le taux d'erreur global d'apprentissage pour l'arbre de décision.

5. À partir d'un échantillon test contenant 10694 pixels, on obtient la matrice de confusion suivante :

	Prédictions pour le type de matériau								
	1	2	3	4	5	6	7	8	9
1	1386	170	0	2	0	11	0	108	0
2	0	4315	0	45	0	224	0	13	0
3	91	4	0	0	0	9	0	401	0
4	0	264	0	451	0	14	0	0	0
5	15	0	0	0	0	269	0	85	0
6	0	908	0	0	0	360	0	13	0
7	350	3	0	0	0	0	0	6	0
8	17	10	0	0	0	26	0	863	0
9	1	0	0	260	0	0	0	0	0

En déduire le taux d'erreur global de prédiction pour cet échantillon test ; ce résultat confirme-t-il le taux d'erreur global d'apprentissage calculé précédemment?

Forêt aléatoire

On analyse maintenant ces données en utilisant une forêt aléatoire.

1. Expliquez ce que représente la figure 3 ; que pouvez-vous conclure à partir de ce graphique?
2. À partir de l'échantillon test contenant 10694 pixels, on obtient la matrice de confusion suivante :

	Prédictions pour le type de matériau								
	1	2	3	4	5	6	7	8	9
1	1483	8	17	0	0	21	50	73	0
2	2	4402	0	83	0	93	0	9	0
3	111	2	225	0	0	2	2	178	0
4	0	86	0	685	1	4	0	0	0
5	3	1	0	0	340	0	0	5	0
6	5	589	0	0	0	736	0	14	0
7	114	0	3	0	0	0	194	1	0
8	29	1	61	0	0	7	0	813	0
9	0	0	0	0	0	0	0	0	241

   (a) En déduire le taux d'erreur global de prédiction pour cet échantillon test,

   (b) Si vous aviez le choix entre l'arbre de décision et la forêt aléatoire, quelle méthode utiliseriez-vous?

3. La figure 4 (voir annexe) fournit les prédicteurs de plus grande importance ; cette mesure d'importance est-elle en accord avec l'arbre de décision fourni en annexe?

Annexes

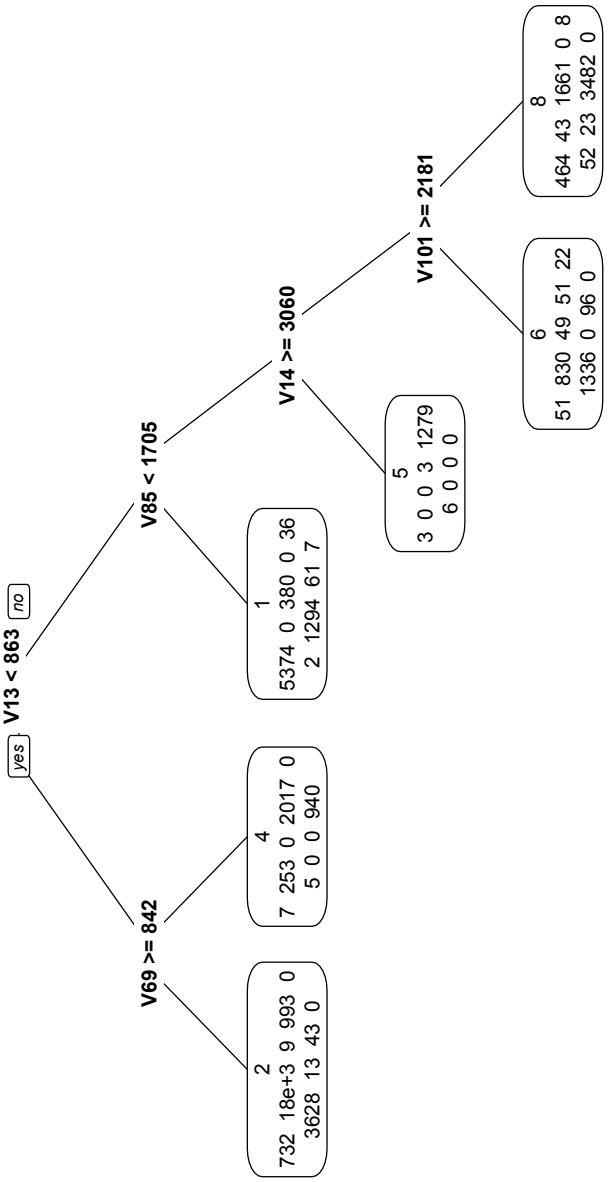

Figure 1: Arbre de décision

~

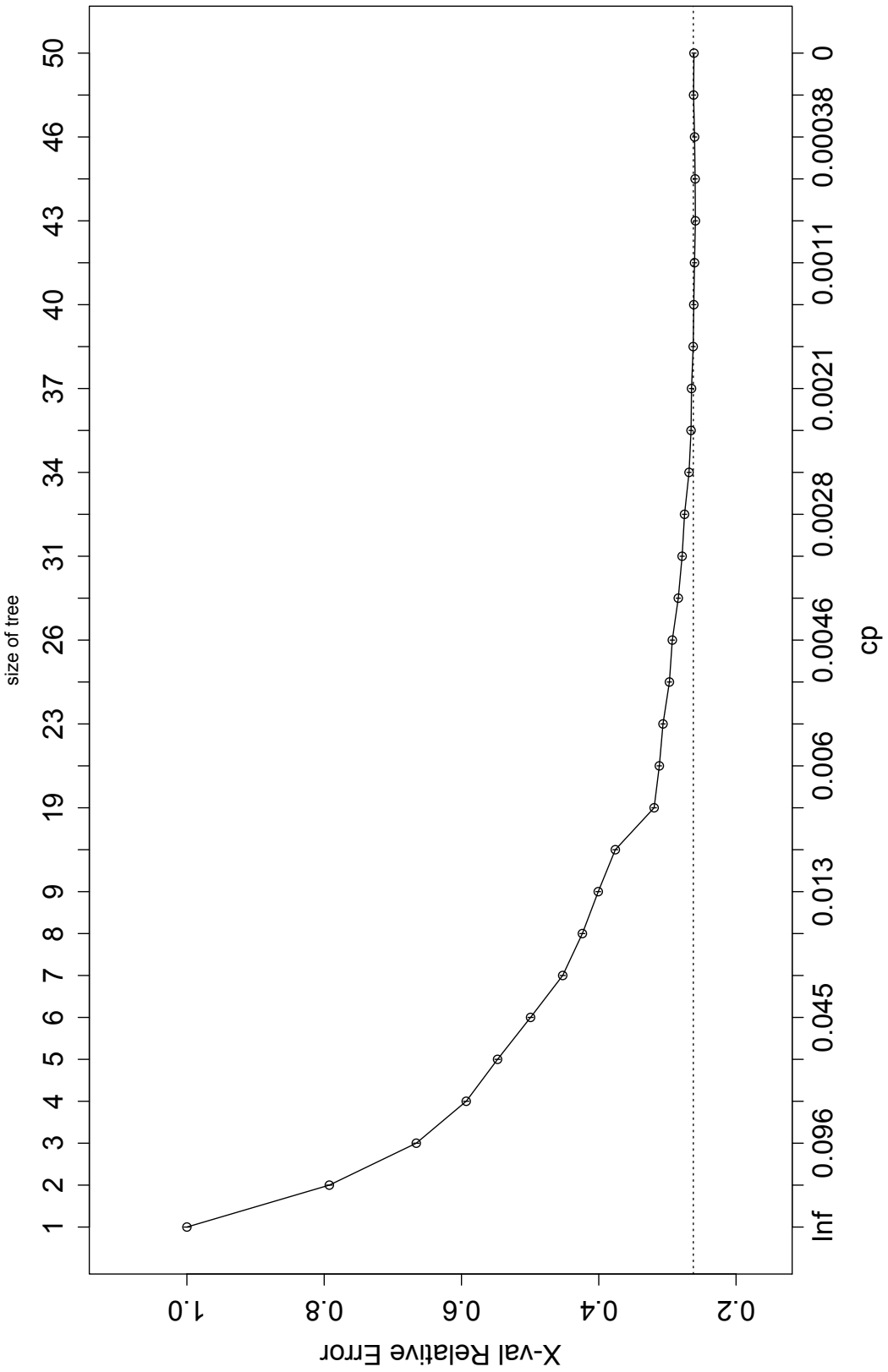
size of tree
1
2
3
4
5
6
7
8
9
19
23
26
31
34
37
40
43
46
50
X-val Relative Error
0.2
0.4
0.6
0.8
1.0
cp
Inf
0.096
0.045
0.013
0.006
0.0046
0.0028
0.0021
0.0011
0.00038
0

Figure 2:

~

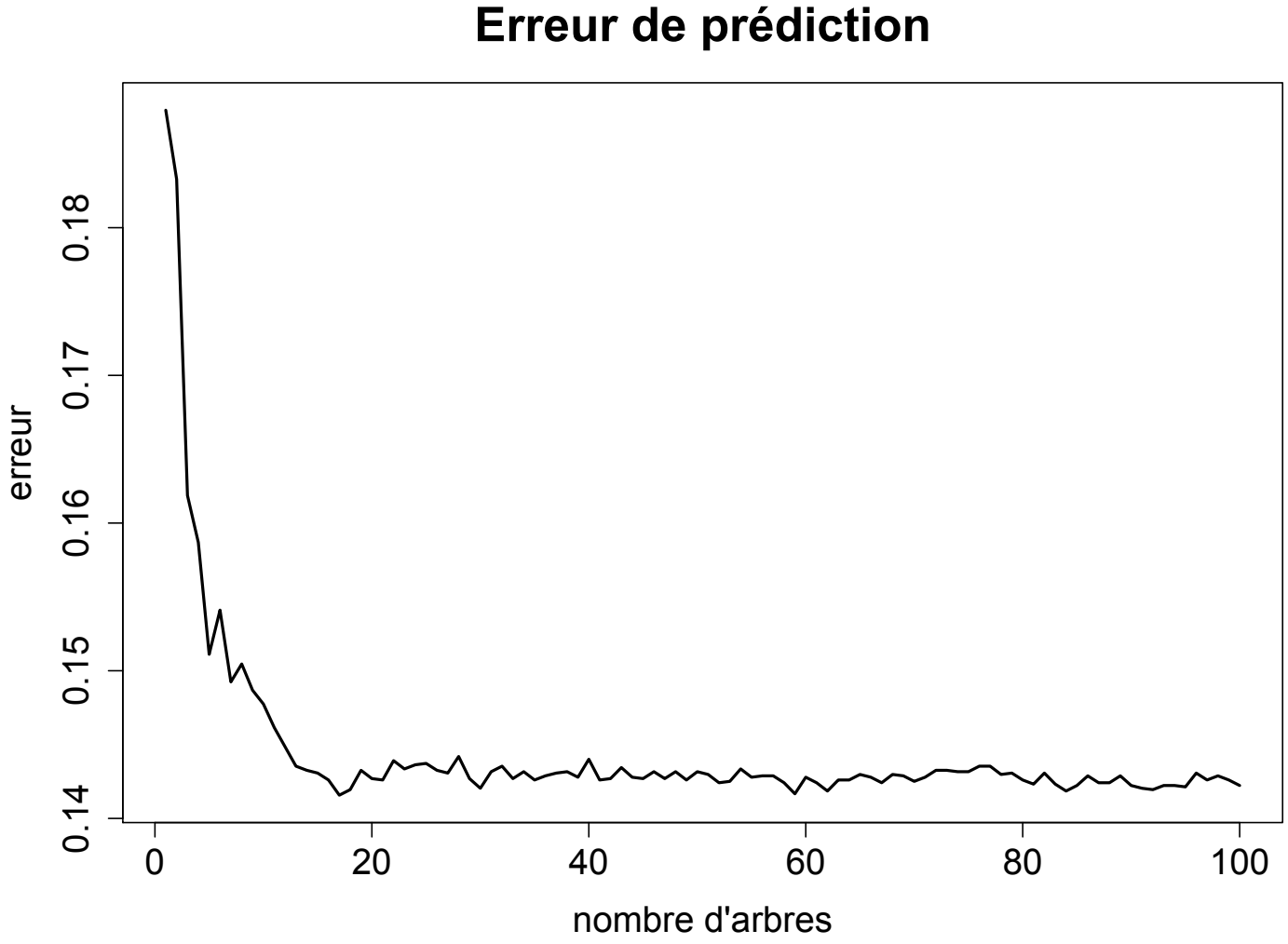
Erreur de prédiction
erreur
0.14
0.15
0.16
0.17
0.18
0
20
40
60
80
100
nombre d'arbres

Figure 3:

~

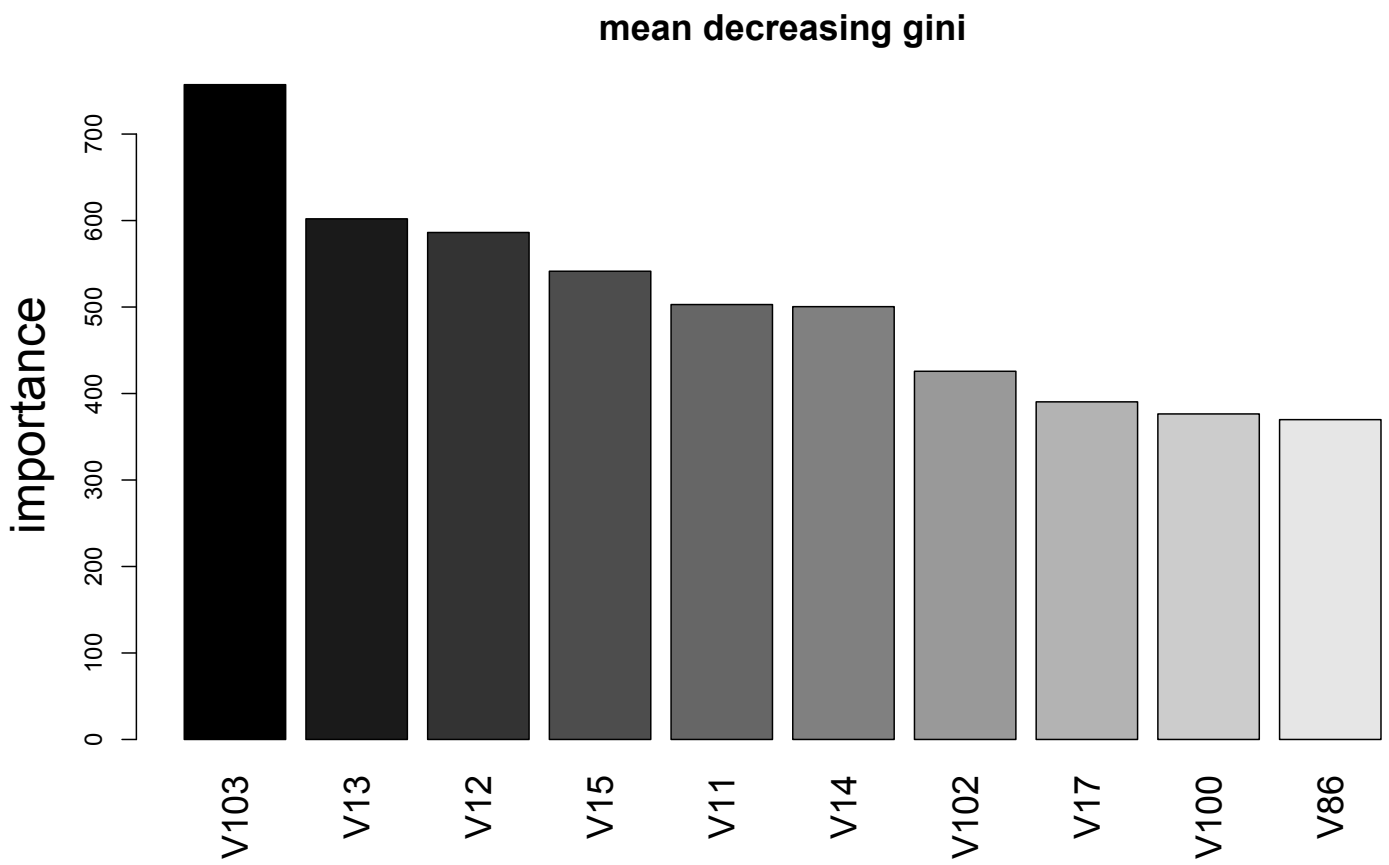

Figure 4: Forêt aléatoire : importance des prédicteurs

~