Examen d'analyse de données – sujet 1

Fabrice Rossi

6 janvier 2017

Les exercices sont indépendants et peuvent être traités dans n'importe quel ordre.

Exercice 1

On étudie un ensemble de 20 observations $(X_i, Y_i)_{1 \leq i \leq 20}$, avec $Y_i \in \{1, 2, 3\}$. On construit sur cet ensemble deux modèles, g_1 et g_2 dont les prévisions sur cet ensemble sont données par le tableau suivant :

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y_i	3	1	2	2	3	1	3	3	2	2	1	1	1	3	2	3	2	3	3	2
$g_1(X_i)$	1	3	2	2	3	1	3	3	2	3	1	1	1	3	2	1	2	1	3	3
$g_2(X_i)$	3	3	2	1	3	1	2	3	2	1	1	1	3	3	2	3	2	3	3	2

Dans ce tableau, chaque colonne correspond à une observation (X_i, Y_i) et précise la valeur de Y_i et les prévisions des deux modèles.

Question 1 Calculer les matrices de confusion empiriques des deux modèles.

Correction

On obtient pour g_1

	1	2	3
1	4	0	3
2	0	5	0
3	1	2	5

et pour g_2

	1	2	3
1	3	2	0
2	0	5	1
3	2	0	7

Question 2 Déterminer le meilleur modèle (entre g_1 et g_2) au sens du risque empirique pour la fonction de perte $l_1(u,v) = |u - v|$.

Correction

Comme on utilise le même nombre d'observations (20) dans chaque cas, on peut se contenter de comparer la somme des pertes empiriques.

On obtient pour g_1 la somme 10 et pour g_2 la somme 7. Le meilleur modèle est donc g_2.

Question 3 Même question pour la fonction de perte l_2 donnée par la table suivante :

$l_2(u,v)$		v		
		1	2	3
u	1	0	2	1
	2	1	0	1
	3	2	1	0

On rappelle que par convention, le premier argument de la fonction de perte est la prévision du modèle.

Correction

On utilise aussi la somme plutôt que la moyenne. On obtient pour g_1 la somme 7 et pour g_2 la somme 9. Le meilleur modèle est donc g_1.

Exercice 2

On étudie un problème à trois variables explicatives binaires $X1$, $X2$ et $X3$ et à une variable à expliquer binaire Y (toutes les variables sont donc à valeurs dans $\{0,1\}$). On suppose que le modèle optimal pour la fonction de perte $l_0(u,v) = \mathbb{I}_{u \neq v}$ est donné par l'arbre de décision de la figure 1. Il s'agit du modèle optimal g^* théorique.

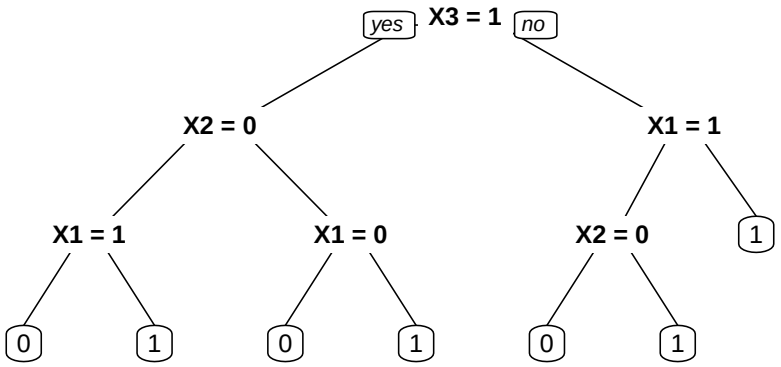

FIGURE 1 – Modèle optimal. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». La valeur indiquée dans chaque feuille donne la classe associée à celle-ci.

On dispose d'un total de 100 observations, dont 63 correspondent $Y = 1$. Les observations sont décrites de façon simplifiée par le tableau suivant :

Valeur de Y	Nombre de X1=1	Nombre de X2=1	Nombre de X3=1
0	25	10	26
1	23	39	13

Chaque ligne correspond à une valeur de Y (la première concerne les observations telles que $Y = 0$, la seconde celles pour lesquelles $Y = 1$). Les trois colonnes des variables explicatives indiquent le nombre d'observations avec la valeur 1 pour la variable concernée et avec simultanément la valeur de Y précisée en ligne. Par exemple, la valeur 25 de la première ligne, colonne $X1$ indique que parmi les 37 observations pour lesquelles $Y = 0$, 25 ont une valeur de 1 pour la variable $X1$. De même la valeur 39 de la deuxième ligne, colonne $X2$ indique que parmi les 63 observations pour lesquelles $Y = 1$, 39 ont une valeur de 1 pour la variable $X2$.

On cherche à construire un classifieur bayésien naïf sur ces données.

Question 1 Donner de façon précise et complète le modèle génératif du classifieur bayésien naïf (CBN) pour le problème étudié. Rappeler en particulier les différentes lois choisies et les hypothèses d'indépendance utilisées.

Correction

Le modèle CBN suppose d'abord que les observations sont indépendantes et identiquement distribuées (c'est l'hypothèse de base en apprentissage). De façon plus spécifique, le modèle suppose que les variables $X1$, $X2$ et $X3$ sont conditionnellement indépendantes, sachant Y. On a donc

$$\mathbb{P}(X1 = x1, X2 = x2, X3 = x3|Y = y) = \mathbb{P}(X1 = x1|Y = y)\mathbb{P}(X2 = x2|Y = y)\mathbb{P}(X3 = x3|Y = y).$$

Les variables sont toutes binaires et suivent donc des lois de Bernoulli. On note π le paramètre de la loi de Y, soit $\mathbb{P}(Y = 1) = \pi$, et θ_{ik} le paramètre de la loi de Xi conditionnellement à $Y = k$, soit donc $\mathbb{P}(Xi = 1|Y = k) = \theta_{ik}$.

Question 2 En utilisant le principe du maximum de vraisemblance, estimer grâce aux informations fournies ci-dessus les paramètres des lois utilisées par le modèle CBN.

Correction

On sait que l'estimation par maximum de vraisemblance (MV) du CBN conduit à réaliser une estimation par MV variable par variable et classe par classe. On sait en outre que l'estimation par MV du paramètre d'une loi de Bernoulli est la fréquence du cas positif. Pour estimer π on utilise donc la fréquence de $Y = 1$ dans les données. Pour estimer θ_{ik} on utilise la fréquence de $Xi = 1$ pour les observations telles que $Y = k$.

On obtient pour $\pi = \frac{63}{100}$. Les autres valeurs sont :

$$\theta_{10} = \frac{25}{37} \qquad \theta_{20} = \frac{10}{37} \qquad \theta_{30} = \frac{26}{37}$$
$$\theta_{11} = \frac{23}{63} \qquad \theta_{21} = \frac{39}{63} \qquad \theta_{31} = \frac{13}{63}$$

Question 3 Exprimer pour le modèle CBN $\log \frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}$ (avec $X = (X1, X2, X3)$) sous forme d'une fonction affine en $x = (x1, x2, x3)$ faisant apparaître les paramètres du modèle.

Correction

On commence par utiliser la règle de Bayes pour écrire :

$$\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \frac{\mathbb{P}(X=x|Y=1)\mathbb{P}(Y=1)}{\mathbb{P}(X=x|Y=0)\mathbb{P}(Y=0)}$$

On applique ensuite l'hypothèse d'indépendance conditionnelle du CBN pour obtenir (cf au dessus) :

$$\mathbb{P}(X=x|Y=y) = \mathbb{P}(X1=x1|Y=y)\mathbb{P}(X2=x2|Y=y)\mathbb{P}(X3=x3|Y=y).$$

Enfin, on utilise la technique de symétrisation des lois de Bernoulli pour écrire

$$\mathbb{P}(Xi=xi|Y=y) = \theta_{ik}^{xi}(1-\theta_{ik})^{1-xi}.$$

En combinant les résultats, on voit qu'on a

$$\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \frac{\pi}{1-\pi}\frac{\prod_{i=1}^{3}\theta_{i1}^{xi}(1-\theta_{i1})^{1-xi}}{\prod_{i=1}^{3}\theta_{i0}^{xi}(1-\theta_{i0})^{1-xi}}.$$

En prenant le log et en regroupant les termes, on obtient

$$\log\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \log\frac{\pi}{1-\pi} + \sum_{i=1}^{3} xi\left(\log\frac{\theta_{i1}}{\theta_{i0}}\right) + \sum_{i=1}^{3}(1-xi)\left(\log\frac{1-\theta_{i1}}{1-\theta_{i0}}\right),$$

ce qui s'écrit

$$\log\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \alpha_0 + \sum_{i=1}^{3}\alpha_i x_i,$$

avec

$$\alpha_0 = \log\frac{\pi}{1-\pi} + \sum_{i=1}^{3}\log\frac{1-\theta_{i1}}{1-\theta_{i0}},$$
$$\alpha_i = \log\frac{\theta_{i1}(1-\theta_{i0})}{\theta_{i0}(1-\theta_{i1})}.$$

Question 4 Calculer sous forme d'une table la fonction correspondant au modèle optimal (pour la fonction de perte l_0) en supposant que les données sont bien distribuées selon

un modèle CBN et en remplaçant les paramètres réels par les estimations obtenues à la question précédente.

Correction

Il suffit de calculer les ratios $\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}$ pour chaque valeur possible pour x. Les variables étant binaires, il s'agit des huit vecteurs binaires à 3 coordonnées, de $(0,0,0)$ à $(1,1,1)$. On obtient la table suivante :

X1	X2	X3	g(X)
0	0	0	1
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

Les matrices de confusion de l'arbre de la figure 1 et du modèle optimal calculé à la question 4 sont données ci-dessous (pour les 100 observations) :

	0	1
0	33	1
1	4	62

Arbre

	0	1
0	17	12
1	20	51

CBN

Dans ces matrices, les prévisions des modèles sont en ligne, les valeurs réelles en colonne.

Question 5 Pourrait-on calculer ces matrices à partir des informations fournies dans l'énoncé ?

Correction

Non, il faudrait avoir accès aux valeurs complètes des observations, éléments dont on ne dispose pas dans l'énoncé.

Question 6 Représenter la fonction de décision de l'arbre optimal sous forme d'une table similaire à celle construite à la question 4.

Correction

On obtient la table suivante :

X1	X2	X3	g(X)
0	0	0	1
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	1

Question 7 Sachant que l'arbre est le modèle optimal, pouvait-on attendre des résultats satisfaisants lors de l'utilisation du modèle CBN ? Réponse à justifier soigneusement.

Correction

On constate que le CBN donne de bien moins bons résultats (sur l'ensemble d'apprentissage) que l'arbre optimal. On pouvait s'y attendre en théorie car le CBN fait une hypothèse d'indépendance entre les variables au sein d'une classe, alors qu'un arbre utilise au contraire les dépendances entre variable pour prendre ses décisions.

Pour confirmer ce raisonnement théorique, on pourrait chercher un exemple de comportement de type XOR dans la table (on fixe une variable et on étudie la décision avec les deux variables libres).

Exercice 3

On étudie un ensemble de 310 patients d'un service hospitalier. Chaque patient est décrit par six variables numériques ($X1$ à $X6$) et par une variable nominale Y prenant les trois valeurs $\{DH, NO, SL\}$. La valeur NO désigne un « patient » sain (référence pour l'étude). Les deux autres modalités correspondent à deux pathologies.

Dans un premier temps, l'analyste applique la méthode CART aux données en utilisant l'intégralité des observations comme ensemble d'apprentissage. La figure 2 représente l'évolution du nombre d'erreurs de classement en fonction du nombre de feuilles de l'arbre, sur l'ensemble d'apprentissage et estimé par une validation croisée à 10 blocs.

Question 1 L'analyste décide de retenir un arbre à 5 feuilles. Justifier son choix.

Correction

Il s'agit du meilleur arbre au sens du risque évalué par une procédure de validation croisée. Cette procédure étant statistiquement valide, l'estimation proposée est considérée comme suffisamment représentative du vrai risque des arbres étudiés pour qu'on puisse l'utiliser pour choisir l'arbre à retenir.

Question 2 L'arbre élagué est donné par la figure 3. Déduire de la figure la matrice de confusion de l'arbre à 5 feuilles sur les données d'apprentissage.

Correction

On obtient :

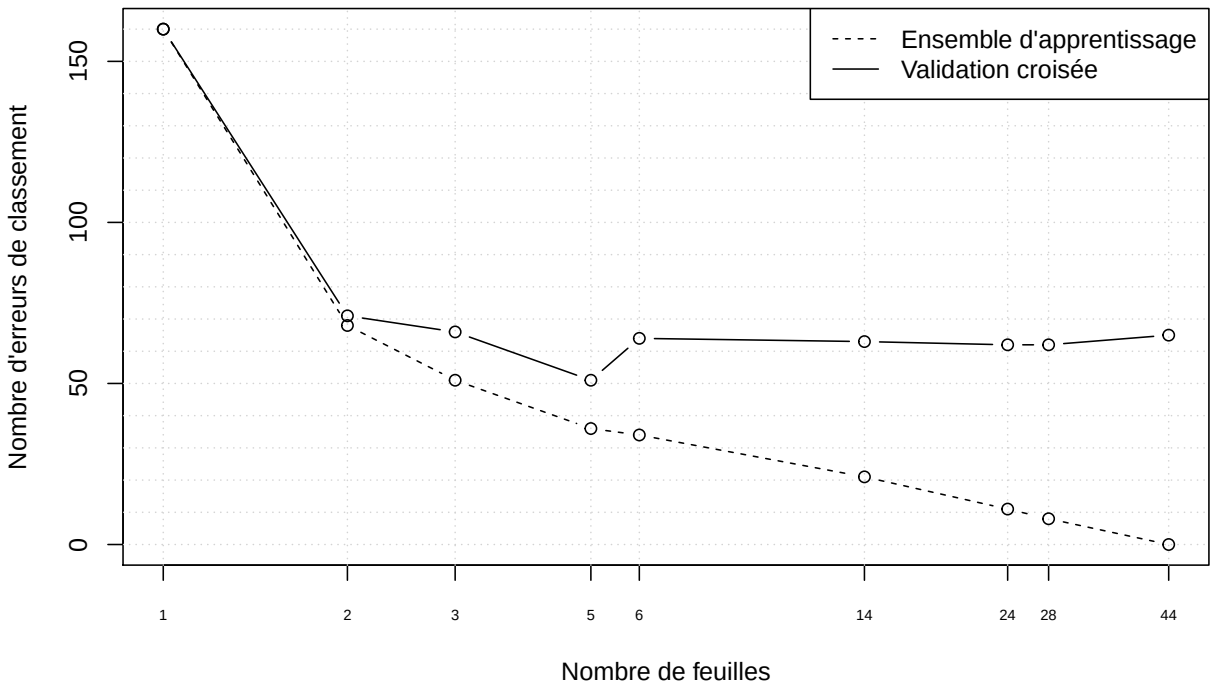

FIGURE 2 – Évolution du nombre d'erreurs de classement en fonction du nombre de feuilles dans l'arbre, sur l'ensemble d'apprentissage et estimé par validation croisée. L'axe des x utilise une échelle logarithmique.

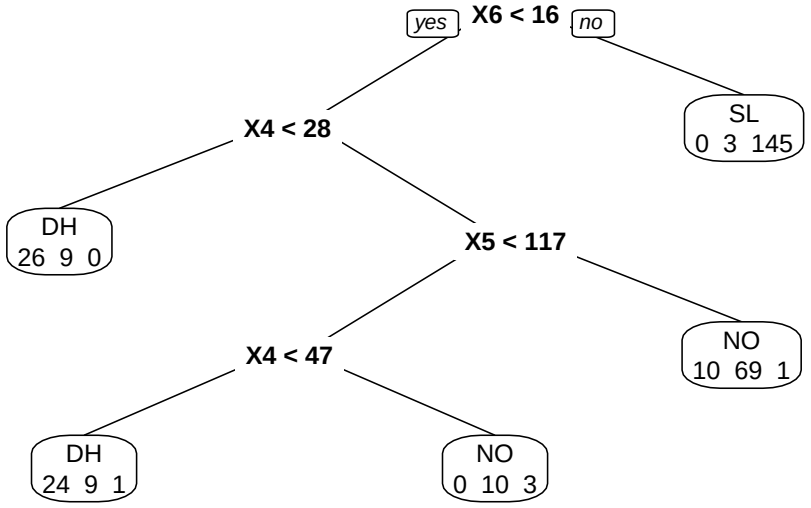

FIGURE 3 – Arbre de classification optimal. La branche de gauche de chaque nœud correspond toujours à la réponse « oui » à la question du nœud, l'autre branche à la réponse « non ». Les lettres indiquées sur la première ligne de chaque feuille donnent la classe associée à la feuille. Les nombres de la deuxième ligne d'une feuille indiquent les effectifs des trois classes dans la feuille, dans l'ordre DH, NO et SL.

	DH	NO	SL
DH	50	18	1
NO	10	79	4
SL	0	3	145

Question 3 L'analyste souhaite estimer le taux d'erreurs que l'arbre à 5 feuilles aura sur de nouvelles données. Peut-elle le faire à partir des éléments présentés jusqu'à présent ?

Correction

En toute rigueur, non. En effet, l'estimation du risque par la validation croisée a déjà été utilisé pour sélectionner le meilleur modèle. Il y a donc potentiellement une surestimation des performances de cet arbre élagué. Cependant, on peut dire que les performances observées en validation croisée donnent une idée des meilleures performances qu'on peut espérer pour cet arbre sur de nouvelles données.

Notons que les performances sur l'ensemble d'apprentissage telles que données par la matrice de confusion ne sont pas celles sur lesquelles il faudrait se baser (on voit notamment sur la figure 2 l'écart de taux d'erreur).

L'analyste décide de partitioner aléatoirement l'ensemble des données en deux sous-ensembles, A et B, d'effectifs approximativement égaux.

Question 4 L'analyste impose une (des) contrainte(s) sur la partition : laquelle (lesquelles) ?

Correction

On constate sur l'arbre, par exemple, que les classes ne sont pas équilibrées. Il est donc naturel de respecter ce déséquilibre dans les partitionnements. On impose donc à la partition en A et B d'avoir des effectifs approximativement égaux classe par classe (stratégie de stratification).

En appliquant une méthode de validation croisée respectant la (les) même(s) contrainte(s) que la partition, l'analyste obtient un arbre « optimal » avec 5 feuilles. La matrice de confusion de l'arbre obtenu appliqué à l'ensemble B est

	DH	NO	SL
DH	21	15	1
NO	8	32	2
SL	1	3	72

où les prévisions sont en ligne.

Question 5 Comparer les résultats avec ceux obtenus sur l'ensemble complet.

Correction

On constate que les performances de deux approches sont compatibles. En effet, sur l'ensemble d'apprentissage, l'arbre élagué fait 36 erreurs de classement pour 310 observations. Avec la nouvelle procédure, on observe 30 erreurs pour 155 observations, mais les performances du premier arbre sont surestimées par l'évalution sur l'ensemble

d'apprentissage. Par validation croisée, on avait estimé le nombre d'erreur à 51, ce qui est compatible avec les résultats de la second approche. Comme annoncé dans la réponse à la question 3, il semble tout de même que la validation croisée ait sous-estimé le taux d'erreur sur un nouvel ensemble.

Question 6 Quelles sont les performances attendues pour l'arbre obtenu sur l'ensemble A si on l'appliquait à un nouvel ensemble d'observations issu de patients similaires à ceux étudiés ici ?

Correction

Comme l'ensemble B est indépendant de l'ensemble A et comme toute la procédure de construction de l'arbre a été réalisée uniquement sur A, on s'attend à ce que les performances sur B correspondent bien à ce qui se passerait sur un nouvel ensemble.

Exercice 4

Soit $\mathcal{G}$ un espace vectoriel de fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ de dimension m. On rappelle que la dimension de Vapnik-Chervonenkis ($VC - dim$) de l'ensemble $\mathcal{F}$ de fonctions de $\mathbb{R}^d$ dans $\{0,1\}$ défini par

$$\mathcal{F} = \{f : \mathbb{R}^d \to \{0,1\} \mid \exists g \in \mathcal{G}, \forall x \in \mathbb{R}^d, f(x) = \mathbb{I}_{g(x)\geq 0}\},$$

est telle que $VC - dim(\mathcal{F}) \leq m$.

On s'intéresse à la classe de fonctions $\mathcal{H}$ définie par

$$\mathcal{H} = \{h : \mathbb{R}^d \to \{0,1\} \mid \exists a \in \mathbb{R}^d, b \in \mathbb{R}, \forall x \in \mathbb{R}^d, f(x) = \mathbb{I}_{\|x-a\|^2 \leq b^2}\}.$$

En termes moins formels, il s'agit des fonctions indicatrices des boules fermées de $\mathbb{R}^d$ (a est le centre de la boule, b est son rayon, $h(x)$ vaut 1 si et seulement si x est dans la boule). En utilisant le résultat rappelé ci-dessus, on cherche à montrer que $VC - dim(\mathcal{H}) \leq d + 2$.

Question 1 Pour $a \in \mathbb{R}^d$ et $b \in \mathbb{R}$ fixés, écrire $b^2 - \|x - a\|^2$ sous la forme d'une combinaison linéaire de fonctions de $\mathbb{R}^d$ dans $\mathbb{R}$ indépendantes de a et b : les coefficients peuvent dépendre de a et b, mais pas l'expression des fonctions.

Correction

Il suffit de développer l'expression $\|x - a\|^2$. On a en effet

$$\begin{aligned} b^2 - \|x - a\|^2 &= b^2 - \|x - a\|^2, \\ &= (b^2 - \|a\|^2) + 2x.a - \|x\|^2, \\ &= (b^2 - \|a\|^2) - \|x\|^2 + 2\sum_{k=1}^{d} a_k x_k. \end{aligned}$$

On pose alors $f_0(x) = 1$, $f_k(x) = x_k$ et $f_{d+1}(x) = \|x\|^2$, ce qui montre que

$$b^2 - \|x - a\|^2 = \sum_{k=0}^{d+1} \lambda_k f_k(x),$$

avec $\lambda_0 = b^2 - \|a\|^2$, $\lambda_k = 2a_k$ et $\lambda_{d+1} = -1$.

Question 2 Montrer que les fonctions obtenues sont linéairement indépendantes. On pourra procéder par l'absurde.

Correction

Soit $d+2$ coefficients $\gamma_0, \ldots, \gamma_{d+1}$ tels que $\sum_{k=0}^{d+1} \gamma_k f_k = 0$. En évaluant l'équation en $x = 0$, on déduit que $\gamma_0 = 0$. Pour chaque $1 \leq l \leq d$, on considère ensuite le vecteur $x^{(l)}$ défini par $x_k^{(l)} = \delta_{lk}\epsilon$. On a alors $f_k(x^{(l)}) = \delta_{lk}\epsilon$ et $f_{d+1}(x^{(kl)}) = \epsilon^2$. L'équation devient alors (pour $\epsilon \neq 0$)

$$\begin{aligned} 0 &= \sum_{k=0}^{d+1} \gamma_k f_k(x^{(l)})) \\ &= \gamma_l \epsilon + \gamma_{d+1}\epsilon^2 \\ &= \gamma_l + \gamma_{d+1}\epsilon \end{aligned}$$

En prenant $\epsilon > 0$, on constate que γ_l doit avoir le signe opposé de celui de γ_{d+1}. Mais en prenant $\epsilon < 0$, on aboutit à la conclusion contraire et donc $\gamma_l = 0$. L'équation se réduit alors à $\gamma_{d+1}\|x\|^2 = 0$ et donc $\gamma_{d+1} = 0$. Les coefficients sont donc tous nuls, ce qui permet de conclure à l'indépendance linéaire des fonctions.

Question 3 Conclure.

Correction

Soit $\mathcal{G}$ l'espace vectoriel engendré par les f_k. On vient de montrer que la dimension de $\mathcal{G}$ est $d+2$. Donc la $VC-dim$ de l'ensemble $\mathcal{F}$ défini comme dans l'énoncé est inférieure ou égale à $d+2$. Or, il est clair que l'ensemble $\mathcal{H}$ est un sous-ensemble de $\mathcal{F}$ (mais les ensembles ne sont pas égaux). Or, si $U \subset V$, alors $VC-dim(U) \leq VC-dim(V)$. En effet, il est clair que si $U \subset V$, alors pour tout n, le coefficient de pulvérisation pour n de U est plus petit ou égal à celui de V. De ce fait, si U pulvérise un ensemble, alors V le fait, ce qui permet de conclure. En appliquant cette propriété générale $\mathcal{F}$ et $\mathcal{H}$ on déduit que la $VC-dim$ de ce dernier est inférieure ou égale à $d+2$.