

Chapitre 2

Chaînes de Markov cachées discrètes

Les chaînes de Markov cachées jouent un rôle important dans les applications les plus divers en économie, finances, biologie, santé, informatique, réseaux, traitement du signal, ... C'est un modèle parmi les plus simples modélisant des dépendances stochastiques entre les variables aléatoires constituant une suite (X_1, \dots, X_n) (qui seront considérées comme inobservées) à valeurs discrètes finies et une suite (Y_1, \dots, Y_n) , qui seront considérées comme observées. L'intérêt de la modélisation par des chaînes de Markov cachées est qu'elle permet de faire des calculs analytiques des quantités d'intérêt – qui sont celles qui permettent de rechercher les réalisations cachées (X_1, \dots, X_n) à partir des réalisations observées de (Y_1, \dots, Y_n) – pour des grandes valeurs de n . Malgré leur simplicité ces modèles présentent une grande robustesse et peuvent donner d'excellents résultats.

1. Classification bayésienne

On considère une variable aléatoire Y prenant ses valeurs dans R et on suppose que sa loi dépend d'un paramètre discret appartenant à un ensemble $\Omega = \{\omega_1, \dots, \omega_k\}$. Le problème est d'estimer le paramètre $\omega \in \Omega$ à partir de l'observation $Y = y$. L'ensemble Ω sera dit ensemble des "classes", et tout estimateur $\hat{s} : R \rightarrow \Omega$ sera dit "stratégie de classification".

Si les lois de Y admettent des densités $p(\cdot|\omega_1), \dots, p(\cdot|\omega_k)$ par rapport à la mesure de Lebesgue on peut utiliser l'estimateur du maximum de vraisemblance (EMV) :

$$[\hat{s}_{MV}(y) = \omega_i] \Leftrightarrow [p(y|\omega_i) = \sup_{1 \leq l \leq k} p(y|\omega_l)] \quad (1.1)$$

Supposons que nous nous trouvons devant un problème de classification de plusieurs nombres réels y_1, \dots, y_n , qui sont des réalisations de la variable Y , et que nous connaissons la fréquence d'apparition des classes. Par exemple, on classe les individus en classe "hommes" et la classe "femme", uniquement à partir de leur poids, et on sait "a priori" (ce qui signifie ici "avant l'observation"), que la population que nous devons classer contient deux tiers d'hommes et un tiers de femmes. Une telle connaissance "a priori" peut être modélisée par une probabilité (dite "a priori") sur $\Omega = \{\omega_1, \dots, \omega_k\}$: on la notera encore avec p , en précisant davantage s'il y a un risque de confusion. Cette probabilité peut alors être considérée comme la loi d'une variable aléatoire X prenant ses valeurs dans $\Omega = \{\omega_1, \omega_2, \dots, \omega_k\}$, et les $p(\cdot|x = \omega_i)$ apparaissent comme les lois de Y conditionnelles à X .

Finalement, la loi p sur $\Omega = \{\omega_1, \dots, \omega_k\}$ et les lois $p(\cdot|\omega_i)$ sur R définissent une probabilité sur $\Omega \times R$ et, par conséquent, la probabilité conditionnelle $p(\cdot|y)$ (sachant $y \in R$) sur Ω (dite "a posteriori").

De manière très générale, on parle d'une approche "bayésienne" si une connaissance "a priori" sur les paramètres est modélisée par une loi de probabilité sur l'ensemble des paramètres.

Exemple de calcul

On peut considérer que la probabilité définie sur $\Omega \times R$ par la probabilité $p(x)$ sur Ω et les densités conditionnelles $p(y|x)$ sur R est définie par une application $h : \Omega \times R \rightarrow R^+$ donnée par

$$h(x, y) = p(x)p(y|x), \quad (1.2)$$

qui est une densité de ladite probabilité par rapport à la mesure $\nu \otimes \mu$, où ν est la mesure de dénombrement (ou comptage) sur Ω (la mesure de dénombrement est celle qui associe à chaque sous-ensemble son cardinal), et μ la mesure de Lebesgue sur R . Pour tous les calculs on doit alors se souvenir que "intégrer" par rapport à x revient à sommer par rapport aux éléments de Ω , et "intégrer" par rapport à y revient à intégrer au sens usuel (intégrale classique de Lebesgue) sur R .

Pour calculer $p(\cdot|y)$, la probabilité sur Ω conditionnelle à l'observation $Y = y$, on applique les règles générales de calcul des lois conditionnelles : ayant la densité du couple donnée par (1.2), la densité conditionnelle, qui est la densité de la loi de X conditionnelle à l'observation $Y = y$, est le quotient de la densité du couple par la densité marginale, ce qui s'écrit :

$$p(x|y) = \frac{h(x, y)}{\int_{\Omega} h(x, y) d\nu(x)} = \frac{h(x, y)}{\sum_{x \in \Omega} h(x, y)} = \frac{p(x)p(y|x)}{\sum_{x \in \Omega} p(x)p(y|x)} \quad (1.3)$$

Intuitivement, la différence entre la probabilité a priori p et la probabilité « a posteriori » $p(\cdot|y)$ sur Ω illustre l'apport de l'information (sur l'identité de la classe non observable) contenue dans l'observation y (« a priori » signifie « avant » l'observation, et « a posteriori » signifie « après » l'observation). On retrouve le fait que si les variables sont indépendantes, l'observation de l'une d'entre elles n'apporte aucune connaissance sur le comportement de l'autre et donc ces deux probabilités sont égales.

Considérons une probabilité sur $\Omega \times R$, qui est une loi d'un couple de variables aléatoires (X, Y) . Ainsi $(x, y) \in \Omega \times R$ étant une réalisation de (X, Y) , le problème de la classification devient celui de l'estimation de la réalisation inobservable de la variable X à partir de la variable observable Y .

Considérons une stratégie de classification \hat{s}

$$\hat{s} : R \rightarrow \Omega$$

Pour chaque réalisation $(X, Y) = (x, y)$, \hat{s} peut donner la bonne réponse, $\hat{s}(y) = x$, ou se tromper, $\hat{s}(y) \neq x$. Supposons que les différentes erreurs ne sont pas de gravité équivalente. On le modélise en définissant une application $L : \Omega \times \Omega \rightarrow R^+$ dite "fonction de perte":

$$L(\omega_i, \omega_j) = \begin{cases} 0 & \text{si } \omega_i = \omega_j \\ \lambda_{ij} & \text{si } \omega_i \neq \omega_j \end{cases} \quad (1.4)$$

λ_{ij} modélisant la gravité de l'erreur "on a choisi ω_i alors que la vraie valeur est ω_j ".

Insistons sur le fait que la « perte » modélisée par $L : \Omega \times \Omega \rightarrow R^+$ ne fait pas partie de la modélisation probabiliste considérée. Par ailleurs, à une erreur donnée, deux utilisateurs peuvent avoir des intérêts différents, et donc les pertes qu'ils associent à une même erreur peuvent être différentes. Ainsi que nous allons le voir dans la suite, la possibilité de l'utilisation des fonctions de perte différentes introduit une grande généralité - et une grande souplesse - des modèles probabilistes utilisés à des fins de classification des données.

A stratégie \hat{s} et fonction de perte L données, comment mesurer la qualité de \hat{s} ? Supposons que l'on a n observations indépendantes y_1, \dots, y_n , chacune correspondant à une classe inconnue, à classer. En notant x_1, \dots, x_n les classes correspondantes, la perte globale est $L(\hat{s}(y_1), x_1) + \dots + L(\hat{s}(y_n), x_n)$. On cherche à minimiser cette perte globale, ce qui revient à minimiser son quotient par n . Par la loi des grands nombres, ce dernier tend vers $E[L(\hat{s}(Y), X)]$:

$$\frac{L(\hat{s}(y_1), x_1) + \dots + L(\hat{s}(y_n), x_n)}{n} \xrightarrow{n \rightarrow +\infty} E[L(\hat{s}(Y), X)] \quad (1.5)$$

On constate qu'à « long terme », la qualité d'une stratégie \hat{s} est mesurée par $E[L(\hat{s}(Y), X)]$, qui appelée "perte moyenne".

Insistons sur le fait que, dans (1.5), la suite $\frac{L(\hat{s}(y_1), x_1) + \dots + L(\hat{s}(y_n), x_n)}{n}$ est aléatoire - et donc, en toute rigueur, on doit écrire $\frac{L(\hat{s}(Y_1), X_1) + \dots + L(\hat{s}(Y_n), X_n)}{n}$ - et le nombre $E[L(\hat{s}(Y), X)]$ ne l'est pas.

Par définition, la stratégie bayésienne \hat{s}_B est celle parmi toutes les stratégies pour laquelle la perte moyenne est minimale :

$$E[L(\hat{s}_B(Y), X)] = \min_{\hat{s}} E[L(\hat{s}(Y), X)] \quad (1.6)$$

La qualité de \hat{s}_B est ainsi appréhendée via la loi des grands nombres et on ne peut rien dire pour une seule (ou un petit nombre) observation.

Remarque 1

Tout en étant optimale, \hat{s}_B peut être inefficace. En effet, $E[L(\hat{s}_B(Y), X)]$ peut être grande, ce qui signifie que les variables observées sont « peu liées » aux variables cachées, ou encore, en adoptant le langage du traitement du signal, que les données sont très « bruitées ».

Réciproquement, dans le cas des données « peu bruitées », une mauvaise méthode de classification peut donner de très bons résultats. Nous devons ainsi être très prudents quand à l'appréciation de la qualité de la méthode utilisée à partir de celle des résultats obtenus.

Soit une fonction de perte $L(\omega_i, \omega_j) = \begin{cases} 0 & \text{si } \omega_i = \omega_j \\ \lambda_{ij} & \text{si } \omega_i \neq \omega_j \end{cases}$. Montrons que la stratégie bayésienne associée \hat{s}_B est définie par

$$[\hat{s}_B(y) = \omega_i] \Leftrightarrow [\forall 1 \leq j \leq k, \sum_{m=1}^k \lambda_{im} p(\omega_m|y) \leq \sum_{m=1}^k \lambda_{jm} p(\omega_m|y)] \quad (1.7)$$

Rappelons que pour deux variables aléatoires réelles U, V et une fonction quelconque Ψ on a (formule de Fubini):

$$E[\Psi(U, V)] = E[E[\Psi(U, V)|U]] = E[E[\Psi(U, V)|V]] \quad (1.8)$$

En l'appliquant à $E[L(\hat{s}(Y), X)]$ on peut écrire :

$$E[L(\hat{s}(Y), X)] = E[E[L(\hat{s}(Y), X)|Y]] \quad (1.9)$$

Posons $E[E[L(\hat{s}(Y), X)|Y]] = \varphi(Y)$ et calculons $\varphi(y)$. Nous avons :

$$\varphi(y) = E[L(\hat{s}(Y), X)|Y = y] = \sum_{j=1}^k L(\hat{s}(y), \omega_j) p(\omega_j|y) \quad (1.10)$$

Déterminons $\hat{s}(y) = \omega_i$ qui minimise $\varphi(y)$. Sachant que $L(\omega_i, \omega_j) = \begin{cases} 0 & \text{si } \omega_i = \omega_j \\ \lambda_{ij} & \text{si } \omega_i \neq \omega_j \end{cases}$

l'élément ω_i minimisant $\varphi(y)$ sera celui qui minimise $\sum_{j=1}^k \lambda_{ij} p(\omega_j|y)$, ce qui donne (1.7).

Notons que \hat{s} ainsi déterminée minimise bien $E[L(\hat{s}(Y), X)]$ car on a

$$E[L(\hat{s}(Y), X)] = E[E[L(\hat{s}(Y), X)|Y]] = \int_R \varphi(y) p(y) dy$$

et donc la minimisation de φ en tout point assure la minimisation de son intégrale.

Notons que $p(y)$ est donnée par la densité $p(y) = \sum_{i=1}^k p(\omega_i) p(y|\omega_i)$. Une telle densité est dite densité "mélange".

Remarque 2

Pour calculer la perte moyenne (qui est minimale pour la stratégie bayésienne) associée à la stratégie \hat{s} et la fonction de perte L , on utilise toujours la formule (1.8) en conditionnant par X :

$$E[L(\hat{s}(Y), X)] = E[E[L(\hat{s}(Y), X)|X]] = \sum_{i=1}^k E[L(\hat{s}(Y), X)|X = \omega_i] p(\omega_i) = \sum_{i=1}^k \int_R [L(\hat{s}(y), \omega_i)] p(y|\omega_i) dy p(\omega_i)$$

Nous disposons ainsi de la stratégie qui assure, à long terme, d'avoir une perte minimale et, de plus, il est possible de calculer sa valeur.

Remarque 3

Ainsi la stratégie bayésienne dépend des λ_{ij} que l'on choisit de façon subjectif. Si on souhaite détecter une classe donnée avec une précision ε , on peut calculer les coefficients λ_{ij} de façon à ce que la stratégie bayésienne correspondante vérifie cette condition. Ce type de possibilités montre la puissance de la modélisation en question.

Exemple 4

Soit $\Omega = \{\omega_1, \omega_2\}$ et L définie par

$$L(\omega_i, \omega_j) = \begin{cases} 0 & \text{si } \omega_i = \omega_j \\ 1 & \text{si } \omega_i \neq \omega_j \end{cases} \quad (1.11)$$

$L(\hat{s}(y), \omega)$ désigne alors la valeur, au point (ω, y) , de la fonction indicatrice du sous-ensemble de $\Omega \times Y$ sur lequel \hat{s} se trompe et donc $E[L(\hat{s}(Y), X)]$ représente la probabilité pour que \hat{s} se trompe. Ainsi dans ce cas la stratégie bayésienne \hat{s}_B définie par

$$\hat{s}_B(y) = \begin{cases} \omega_1 & \text{si } p(\omega_1|y) \geq p(\omega_2|y) \\ \omega_2 & \text{si } p(\omega_1|y) < p(\omega_2|y) \end{cases}, \quad (1.12)$$

qui est un cas particulier de (1.7), est celle pour laquelle la probabilité de se tromper est minimale. Sachant qu'en vertu de la loi des grands nombres la probabilité d'un événement peut être vue comme la fréquence de son apparition lorsque le phénomène se reproduit un grand nombre de fois de façon indépendante, la stratégie définie ci-dessus est celle qui produira, lorsqu'on l'utilisera dans un grand nombre de cas indépendants, la plus petite proportion d'erreurs.

Ainsi \hat{s}_B consiste, dans ce cas, à associer à chaque $y \in \mathbf{Y}$ l'élément de Ω dont la probabilité a posteriori, i.e., conditionnelle à $Y = y$, est maximale. Cette règle de décision est aussi appelée celle du "maximum de vraisemblance a posteriori".

Notons que les probabilités « a posteriori » de (1.12) peuvent être remplacées par les densités $h(x, y) = p(x)p(y|x)$ (voir A2). La stratégie \hat{s}_B peut donc également s'écrire

$$\hat{s}_B(y) = \begin{cases} \omega_1 & \text{si } h(x = \omega_1, y) \geq h(x = \omega_2, y) \\ \omega_2 & \text{si } h(x = \omega_1, y) < h(x = \omega_2, y) \end{cases} \quad (1.13)$$

les fonctions $p(\omega_i)p(y|\omega_i)$ étant dites « fonctions discriminantes ».

2. Restaurations bayésiennes de Markov cachés

2.1 Introduction

Une chaîne de Markov cachée est un processus à temps discret doublement stochastique, ou encore composé de deux processus $X = (X_n)_{n \in \mathbb{N}}$ et $Y = (Y_n)_{n \in \mathbb{N}}$. Le processus X est une chaîne de Markov et nous supposons ici que chaque Y_n prend ses valeurs dans l'ensemble des nombres réels R . L'appellation "cachée" signifie que les réalisations de X sont inobservables. Le problème général est alors celui de l'estimation de la réalisation de X à partir de la réalisation observée de Y , ou encore celui de la "restauration" de Y . Le terme de restauration se justifie dans l'optique du vocabulaire du traitement du signal : en adoptant ce dernier, le processus Y peut être considéré comme une version "bruitée" du processus X . La notion du "bruit" est cependant à considérer dans un sens très général. A titre d'exemple, en traitement de la parole un discours peut être considéré comme une suite de phonèmes : le n ième phonème est ainsi la réalisation de X_n . Le bruit modélise ici le fait qu'un phonème particulier est prononcé différemment selon les locuteurs : la mesure représentant une prononciation particulière est ainsi la réalisation de Y_n . L'objectif du modèle est ici de permettre la conception des méthodes de transcription des discours indépendantes du discours considéré (aspect aléatoire de X), mais aussi indépendantes de la personne qui le prononce (aspect aléatoire de Y conditionnellement à X). Un autre exemple est celui d'une image numérique satellite où de l'eau et de la forêt sont présentes. On souhaite établir une carte, ce qui revient à associer à chaque pixel un élément dans l'ensemble $\Omega = \{\omega_1, \omega_2\} = \{eau, forêt\}$. Sur le pixel n

la réalisation invisible de X_n est ainsi "eau" ou "forêt" et la réalisation observée de Y_n est le niveau de gris (un nombre) de l'image numérique considérée. Le bruit modélise ici la "variabilité naturelle" de l'eau et de la forêt : les deux classes ne produisent pas nécessairement deux mesures uniques. D'autres bruits (transmission, acquisition, ...) peuvent éventuellement s'ajouter au bruit "variabilité naturelle". Le bruit global est donc modélisé par les lois de Y_n conditionnelles à X_n .

Notons que les problèmes traités par les chaînes de Markov cachées peuvent également être traités par les méthodes du paragraphe précédent. En effet, le problème général, à savoir retrouver un phénomène discret que l'on ne peut pas observer, à partir des observations "continues", est le même. La différence est que dans le cas de l'échantillon traité précédemment chaque X_n est estimé à partir du seul Y_n et l'information contenue dans les autres Y_i est perdue. Cela revient à considérer que X_n est indépendant (conditionnellement à Y_n) des autres Y_i , ce qui est, dans certains cas, une approximation très grossière de la réalité.

Nous reprenons ci-dessous, dans le cadre des modèles de Markov cachés, la démarche générale utilisée dans le chapitre précédent : description du modèle et étude des méthodes bayésiennes de classification.

2.2 Représentations graphiques des dépendances

Il est souvent pratique, pour avoir une idée intuitive des dépendances entre les variables aléatoires, de considérer des graphes de dépendances. Considérons une famille finie de variables $U = (U_s)_{s \in S}$. On appellera « graphe » tout couple $G = (S, \Gamma)$, où $\Gamma \subset S^2$. Les éléments de Γ sont appelés « arrêtes ». Ainsi deux points de S sont reliés par une arrête ou pas. On dira qu'il existe un chemin entre un point $s \in S$ et un point $t \in S$ si l'on peut passer de s à t en suivant des arrêtes (en passant donc, éventuellement, par d'autres points). Un graphe apporte des renseignements sur les différentes dépendances de la manière suivante :

Définition

Soit S un ensemble fini, $G = (S, \Gamma)$ graphe, et $U = (U_s)_{s \in S}$ un ensemble des variables aléatoires. On dira que la loi de U est « markovienne par rapport au graphe G » si on a la propriété suivante.

Soient $A, B, C \subset S$.

- (i) Si pour tout $a \in A$ et tout $c \in C$ il n'existe pas de chemin entre a à c , alors $U_A = (U_s)_{s \in A}$ et $U_C = (U_s)_{s \in C}$ sont indépendantes ;
- (ii) Si pour tout $a \in A$ et tout $c \in C$ tout chemin éventuel reliant a à c passe par un point de B (on dit que « B sépare A et C »), alors $U_A = (U_s)_{s \in A}$ et $U_C = (U_s)_{s \in C}$ sont indépendantes conditionnellement à $U_B = (U_s)_{s \in B}$.

Notons (ii) est en particulier valable pour les singletons, ce qui donne : pour tous points $s, t, v \in S$, s'il est impossible de passer de s à t sans passer par v , alors U_s et U_t sont indépendantes conditionnellement à U_v .

L'absence d'arrêtes nous renseigne ainsi sur l'indépendance (éventuellement conditionnelle) ; cependant, il est important de noter que la présence d'une arrête n'implique pas la dépendance.

Notons en particulier :

- s'il n'existe pas de chemin entre s et t alors les variables U_s et U_t sont indépendantes ;
- s'il existe des chemin entre s et t , alors les variables U_s et U_t **peuvent être dépendantes ou indépendantes.**

Notons également que si la loi de $U = (U_s)_{s \in S}$ est markovienne par rapport à un graphe $G = (S, \Gamma)$, alors elle est également markovienne par rapport à tout graphe $G' = (S, \Gamma')$, où Γ' est obtenu à partir de Γ par l'ajout d'arrêtes : $\Gamma \subset \Gamma'$. Une loi P_U de $U = (U_s)_{s \in S}$ est ainsi markovienne par rapport à plusieurs graphes. En particulier, n'importe quelle loi est markovienne par rapport au graphe $G = (S, \Gamma)$ où Γ est l'ensemble de toutes les arrêtes.

Pour une loi P_U de $U = (U_s)_{s \in S}$ donnée il est intéressant de chercher un graphe $G = (S, \Gamma)$ dit « **minimal** » qui est tel que **lorsque l'on retire une arrête de Γ la loi de P_U n'est plus markovienne par rapport au nouveau graphe.** Bien entendu, le graphe minimal est le plus informatif car, comme précisé ci-dessus, seule l'absence d'arrêtes permet d'affirmer l'indépendance.

Soit $U = (U_s)_{s \in S}$ dont la loi est markovienne par rapport à un graphe $G = (S, \Gamma)$. Nous avons deux règles suivantes concernant le conditionnement et la marginalisation.

Règle 1 (R1)

Soit A, B une partition de S . Alors la loi de $U_B = (U_s)_{s \in B}$ conditionnelle à $U_A = (U_s)_{s \in A}$ est markovienne par rapport au graphe $G' = (B, \Gamma')$, où Γ' est obtenu à partir de Γ en supprimant toutes les arrêtes **ayant un point de A comme extrémité.** On pourra retenir cette règle en se souvenant que le conditionnement « brise » les arrêtes ;

Règle 2 (R2)

Soit $B \subset S$. Alors la loi de $U_B = (U_s)_{s \in B}$ est markovienne par rapport au graphe $G'' = (B, \Gamma'')$, avec Γ'' obtenu à partir de Γ de la manière suivante. **On retire les points constituant $S - B$ « un par un ». A chaque fois lorsque l'on retire un point on supprime les arrêtes dont il était une des extrémités, et on met une arrête entre tous les couples des points dont chacun était lié par une arrête avec le point retiré.**

Des exemples de graphes obtenus par conditionnement et par marginalisation à partir d'un graphe présenté à la Figure 1.1 sont données sur les Figures 1.2-1.5.

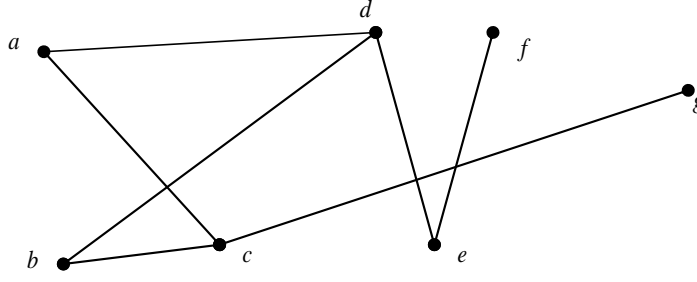


Figure 1. Exemple de graphe des dépendances d'une loi de $U_S = (U_s)_{s \in S}$, avec $S = \{a, b, c, d, e, f, g\}$.

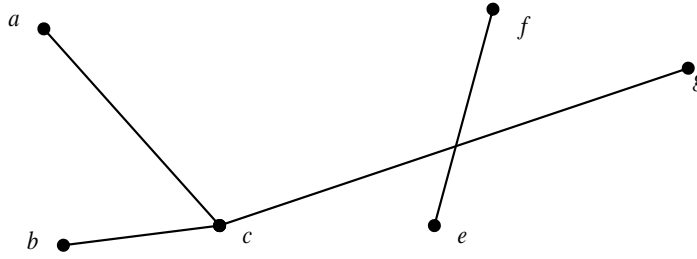


Figure 2. Le graphe des dépendances, obtenu à partir du graphe de la Figure 1, de la loi de $U_B = (U_s)_{s \in B}$ conditionnelle à $U_A = (U_s)_{s \in A}$, avec $B = \{a, b, c, e, f, g\}$ et $A = \{d\}$.

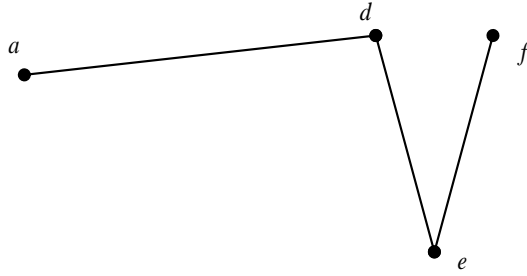


Figure 2. Le graphe des dépendances, obtenu à partir du graphe de la Figure 1, de la loi de $U_B = (U_s)_{s \in B}$ conditionnelle à $U_A = (U_s)_{s \in A}$, avec $B = \{a, d, e, f\}$ et $A = \{b, c, g\}$.

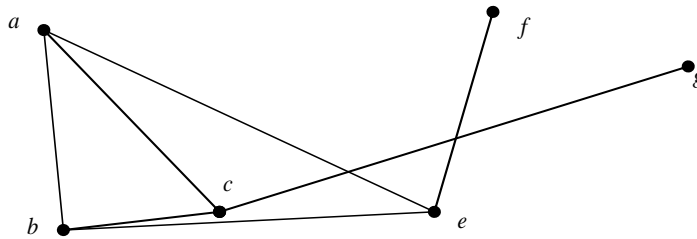


Figure 3. Le graphe des dépendances, obtenu à partir du graphe de la Figure 1, de la loi de $U_B = (U_s)_{s \in B}$ avec $B = \{a, b, c, e, f, g\}$.

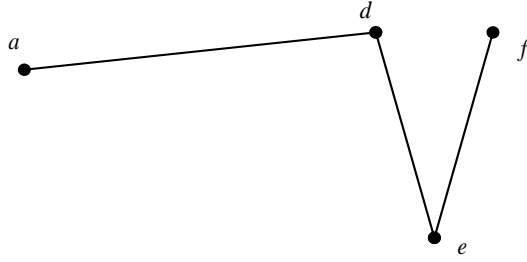


Figure 4. Le graphe des dépendances, obtenu à partir du graphe de la Figure 1, de la loi de $U_B = (U_s)_{s \in B}$ avec $B = \{a, d, e, f\}$.

Remarque

Les graphes exposés ci-dessus sont parfois dits « non-orientés », par opposition aux « graphes orientés », qui utilisent des flèches. Attentions, les deux notions sont différentes. Un graphe orienté peut apporter des précisions sur la manière dont la loi a été définie ; en particulier, en faisant le produit entre la loi de l'ensemble des variables d'où partent les flèches (variables « de départ ») par la loi de l'ensemble où aboutissent les flèches (variables « d'arrivée ») conditionnelles aux variables de départ. Cela peut avoir de l'importance car dans un tel cas lorsque l'on marginalise par rapport à un ensemble d'arrivée on ne modifie pas le graphe de l'ensemble de départ. Le cas simple de trois variables est illustré à la figure 1.5. En marginalisant le graphe non orienté (a) on obtient le graphe (b), on obtient le graphe (b) et on ne peut pas conclure à la dépendance ou non des variable A, B. En marginalisant le graphe orienté (c) on obtient le graphe (d), et on peut conclure à l'indépendance des variable A, B. Les variable A, B, C peuvent être remplacées par des ensembles de variables.

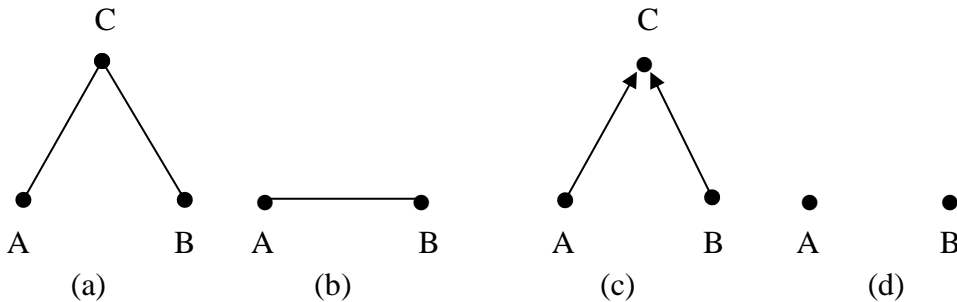


Figure 5 (a) : graphe non orienté du triplet (A, B, C), (b) graphe de (A, B) après marginalisation de (a) ; (c) : graphe orienté du triplet (A, B, C), (d) graphe de (A, B) après marginalisation de (c).

2.3 Le modèle Chaîne de Markov cachées (CMC)

On considère une suite doublement stochastique $(X_1, Y_1, \dots, X_N, Y_N)$, N étant fixé mais quelconque. Pour simplifier, on va noter $X = (X_1, \dots, X_N)$ et $Y = (Y_1, \dots, Y_N)$. Dans une CMC,

la loi de (X, Y) est définie par la loi de X , qui est supposée être une chaîne de Markov, et la loi de Y conditionnelle à X . La loi $p(x)$ de X vérifie donc :

$$p(x_n | x_1, \dots, x_{n-1}) = p(x_n | x_{n-1}), \quad (2.1)$$

pour tout $2 \leq n \leq N$. $p(x)$ est alors déterminée par la loi de X_1 , dite loi initiale, et la suite de matrices de transition $p(x_n | x_{n-1})$. On a

$$p(x) = p(x_1, \dots, x_N) = p(x_1)p(x_2 | x_1) \dots p(x_N | x_{N-1}) \quad (2.2)$$

La loi de X étant définie, il reste à définir les lois de Y conditionnelles à X . On pose les hypothèses suivantes :

- (H1) les variables aléatoires (Y_n) sont indépendantes conditionnellement à X ;
- (H2) la loi de chaque (Y_n) conditionnelle à X est égale à sa loi conditionnelle à X_n .

Avec ces deux hypothèses, nous avons

$$p(y | x) = p(y_1 | x_1) \dots p(y_N | x_N) \quad (2.3)$$

(2.2) et (2.3) impliquent

$$p(x, y) = p(x_1)p(x_2 | x_1) \dots p(x_N | x_{N-1})p(y_1 | x_1) \dots p(y_N | x_N) \quad (2.4)$$

Ainsi que nous allons le voir plus loin, les chaînes de Markov cachées possèdent la propriété suivante, qui est cruciale pour la mise en place des estimations bayésiennes de la chaîne cachée : **la loi de X a posteriori, i.e. conditionnelle à $Y = y$, est de Markov.**

Considérons **la probabilités « progressive » (en anglais « forward ») $\alpha_n(x_n) = p(x_n, y_1, \dots, y_n)$** , qui peut calculée de manière récursive par:

- $\alpha_1(x_1) = p(x_1, y_1)$;
- $\alpha_{n+1}(x_{n+1}) = [\sum_{x_n' \in \Omega} \alpha_n(x_n')p(x_{n+1} | x_n')]p(y_{n+1} | x_{n+1})$ pour $1 \leq n \leq N-1$;

(2.6)

et la probabilité **« rétrograde » (en anglais « backward ») $\beta_n(x_n) = p(y_{n+1}, \dots, y_N | x_n)$** , qui peut calculée de manière récursive par:

- $\beta_N(x_N) = 1$;
- $\beta_n(x_n) = \sum_{x_{n+1}' \in \Omega} \beta_{n+1}(x_{n+1}')p(x_{n+1}' | x_n)p(y_{n+1} | x_{n+1}')$ pour $1 \leq n \leq N-1$.

(2.7)

Nous avons vu que la loi de X a posteriori est markovienne. Montrons le résultat suivant :

Proposition 2.1

(i) Les lois marginales a posteriori $p(x_n|y)$ sont données par

$$p(x_n|y) = \frac{\alpha_n(x_n)\beta_n(x_n)}{\sum_{x_n \in \Omega} \alpha_n(x_n)\beta_n(x_n)} \quad (2.8)$$

(ii) La loi de X a posteriori est markovienne, avec les matrices de transition données par :

$$p(x_{n+1}|x_n, y) = \frac{p(x_{n+1}|x_n)p(y_{n+1}|x_{n+1})\beta_{n+1}(x_{n+1})}{\beta_n(x_n)}, \quad (2.9)$$

Preuve.

(i) $p(y_1, \dots, y_{i-1}, x_i, y_i, y_{i+1}, \dots, y_n) = p(y_1, \dots, y_{i-1}, x_i, y_i) p(y_{i+1}, \dots, y_n | y_1, \dots, y_{i-1}, x_i, y_i) =$
 $= p(y_1, \dots, y_{i-1}, x_i, y_i) p(y_{i+1}, \dots, y_n | x_i) = \alpha_i(x_i) \beta_i(x_i),$

(ii) Selon le graphe d'une CMC les variables (Y_1, \dots, Y_n) et (Y_{n+1}, \dots, Y_N) sont indépendantes conditionnellement à X_n (voir Figure 6, avec $n=3$); il en résulte que

$p(x_{n+1}|x_n, y) = p(x_{n+1}|x_n, y_{n+1}, \dots, y_N)$. Donc

$$\begin{aligned} p(x_{n+1}|x_n, y) \beta_n(x_n) &= p(x_{n+1}|x_n, y_{n+1}, \dots, y_N) p(y_{n+1}, \dots, y_N | x_n) = p(x_{n+1}, y_{n+1}, \dots, y_N | x_n) = \\ &= p(x_{n+1}|x_n) p(y_{n+1}|x_{n+1}, x_n) p(y_{n+2}, \dots, y_N | x_{n+1}, x_n, y_{n+1}) = \\ &= p(x_{n+1}|x_n) p(y_{n+1}|x_{n+1}) p(y_{n+2}, \dots, y_N | x_{n+1}) = p(x_{n+1}|x_n) p(y_{n+1}|x_{n+1}) \beta_{n+1}(x_{n+1}) \end{aligned}$$

Notons que la matrice de transition $p(x_{n+1}|x_n, y)$ de la distribution markovienne de X a posteriori dépend des observations (y_{n+1}, \dots, y_N) et est indépendante des observations (y_1, \dots, y_n) .

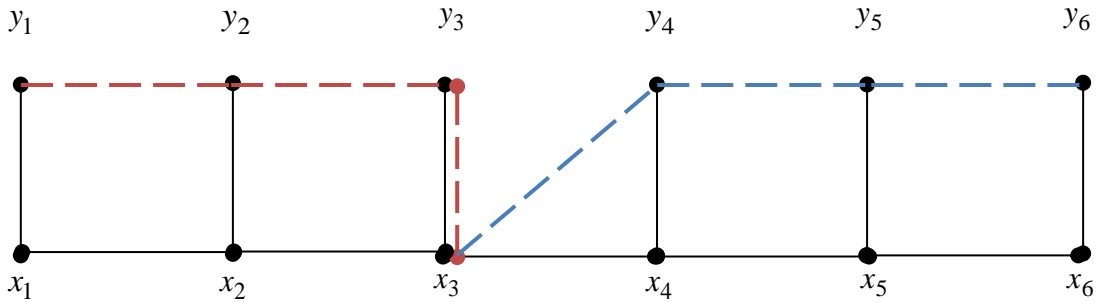


Figure 6. Graphe des dépendances d'une chaîne de Markov cachée.

Nous constatons que dans une CMC de longueur N et nombre de classes k la loi de $X = (X_1, \dots, X_N)$ est donnée par $k + k(k-1)(N-1)$ paramètres (contre k^N dans le cas général), et le calcul de $p(x_n|y)$ demande $2k(N-1)$ additions (contre k^{N-1} dans le cas général). On obtient ainsi un modèle simplifié relativement brutalement, dans lequel les calculs

des quantités d'intérêt sont faisables dans des temps raisonnables. Les CMC seront étendues à des modèles plus généraux (modèles de Markov « couples » et « triplets »), dans lesquels les mêmes calculs restent faisables, dans le chapitre 4.

Finalement, en résumé : Dans une CMC $X = (X_1, \dots, X_N)$, $Y = (Y_1, \dots, Y_N)$, les lois $p(x_n|y)$ et $p(x_{n+1}|x_n, y)$ sont calculables avec une complexité linéaire en N .

2.4 Restauration des chaînes de Markov cachées

Soit $(X, Y) = (X_1, Y_1, \dots, X_N, Y_N)$ une chaîne de Markov cachée, avec les X_n à valeurs dans $\Omega = \{\omega_1, \dots, \omega_k\}$ et les Y_n à valeurs dans R . On s'intéresse à l'estimation - ou « restauration » - de $X = (X_1, \dots, X_N) = x = (x_1, \dots, x_N)$ à partir de $Y = (Y_1, \dots, Y_N) = y = (y_1, \dots, y_N)$. On se place dans le cas où N est trop grand pour que l'on puisse envisager d'utiliser la loi $p(x, y)$ de (X, Y) dans sa forme générale.

On souhaite utiliser les méthodes bayésiennes, ce qui demande une définition préalable d'une fonction de perte $L: \Omega^N \times \Omega^N \rightarrow R^+$. Considérons L simple de la forme suivante

$$L(x^1, x^2) = \sum_{n=1}^N 1_{[x_n^1 \neq x_n^2]}, \quad (2.10)$$

qui consiste à considérer que la perte est le nombre d'éléments mal classés. Nous verrons d'autres exemples de fonctions de perte par la suite.

Proposition 2.2

La stratégie bayésienne \hat{s}_B^L correspondant à la fonction de perte (2.10) est donnée par :

$$[\hat{s}_B^L(y) = (\hat{x}_1, \dots, \hat{x}_N)] \Leftrightarrow [\forall n = 1, \dots, N, \quad p(\hat{x}_n|y) = \sup_{x_n \in \Omega} p(x_n|y)] \quad (2.11)$$

Preuve

De manière analogue à la démarche considérée dans le cas des données indépendantes au chapitre précédent, évaluons, pour une stratégie quelconque \hat{s} , $\varphi(y) = E[L[\hat{s}(y), X]|Y = y]$.

L'objectif est alors de trouver $\hat{s}_B^L(y)$ correspondant pour laquelle cette quantité est minimale. Posons $\hat{s}(y) = (\hat{x}_1(y), \dots, \hat{x}_N(y))$. Nous avons

$$\begin{aligned} \varphi(y) &= E[L[\hat{s}(y), X] | Y = y] = E\left[\sum_{n=1}^N 1_{[\hat{x}_n(y) \neq X_n]} | Y = y\right] = \sum_{n=1}^N E[1_{[\hat{x}_n(y) \neq X_n]} | Y = y] = \\ &= \sum_{n=1}^N P(\hat{x}_n(y) \neq X_n | Y = y) = \sum_{n=1}^N [1 - P(X_n = \hat{x}_n(y) | Y = y)] = \sum_{n=1}^N [1 - p(\hat{x}_n(y) | Y = y)] \end{aligned} \quad (2.12)$$

Les termes dans la dernière somme étant positifs elle est minimale si chacun est minimal, ce qui signifie que chaque terme $p(\hat{x}_n(y) | Y = y)$ est maximal. (2.11) est bien une stratégie bayésienne.

Une deuxième méthode bayésienne souvent utilisée est celle correspondant à la fonction de perte

$$L(x^1, x^2) = 1_{[x^1 \neq x^2]} \quad (2.13)$$

La solution, qui est un cas particulier de la méthode étudiée dans le chapitre 1, s'écrit

$$[\hat{s}_B^L(y) = (\hat{x}_1, \dots, \hat{x}_N)] \Leftrightarrow [p(\hat{x} | y) = \max_{x \in \Omega^N} p(x | y)]. \quad (2.14)$$

La méthode (2.14), appelée simplement « maximum a posteriori » (MAP), est calculable de la manière suivante (algorithme de Viterbi) :

Proposition 2.3

Soit $(X, Y) = (X_1, Y_1, \dots, X_N, Y_N)$ une chaîne de Markov cachée, avec les X_n à valeurs dans $\Omega = \{\omega_1, \dots, \omega_k\}$ et les Y_n à valeurs dans R . Considérons la suite $\hat{x}_{1:1}^1(x_2), \dots, \hat{x}_{1:N-1}^{N-1}(x_N)$, avec

$$p(\hat{x}_{1:n-1}^{n-1}(x_n) | y_{1:n}) = \max_{x_{1:n-1} \in \Omega^n} p(x_{1:n-1}, x_n | y_{1:n}) \quad (2.15)$$

Pour chaque $x_n \in \Omega$, $\hat{x}_{1:n-1}^{n-1}(x_n)$ est ainsi le chemin le plus probable (conditionnellement à $y_{1:n}$) conduisant à x_n , et $p(\hat{x}_{1:n-1}^{n-1}(x_n) | y_{1:n})$ est sa probabilité. La solution $\hat{x}^{MAP} = \hat{s}_B^L(y)$ de (2.14), égale à

$$\hat{x}^{MAP} = \hat{x}_{1:N}^N = \max_{x_N \in \Omega} p(\hat{x}_{1:N-1}^{N-1}(x_N) | y_{1:N}), \quad (2.16)$$

est calculée séquentiellement par :

$$(i) \text{ Calculer } \hat{x}_{1:1}^1(x_2) = \arg \max_{x_1 \in \Omega} p(x_1, x_2 | y_{1:2}) ; \quad (2.17)$$

$$(ii) \text{ Pour } n = 2, \dots, N-1, x_{n+1}, \text{ calculer } \hat{x}_{1:n}^n(x_{n+1}) \text{ avec} \quad (2.18)$$

$$\hat{x}_{1:n}^n(x_{n+1}) = \arg \max_{x_n \in \Omega} p(\hat{x}_{1:n-1}^{n-1}(x_n) | y_{1:n}) p(x_{n+1} | x_n),$$

$$(iii) \text{ Calculer } \hat{x}^{MAP} = \hat{x}_{1:N}^N \text{ avec (2.16)}$$

Preuve.

Nous avons $p(x_{1:n+1} | y_{1:n+1}) = \frac{p(x_{1:n+1}, y_{1:n+1})}{p(y_{1:n+1})} = \frac{p(x_{1:n}, y_{1:n}) p(x_{1:n+1}, y_{1:n+1} | x_{1:n}, y_{1:n})}{p(y_{1:n+1})} =$
 $\frac{p(x_{1:n}, y_{1:n}) p(x_{n+1}, y_{n+1} | x_n, y_n)}{p(y_{1:n+1})} = \frac{p(x_{1:n}, y_{1:n}) p(x_{n+1} | x_n) p(y_{n+1} | x_{n+1})}{p(y_{1:n+1})}$. Pour maximiser $p(x_{1:n+1} | y_{1:n+1})$ à x_{n+1} fixé on maximise donc, par rapport à x_n , le produit $p(\hat{x}_{1:n-1}^{n-1}(x_n) | y_{1:n}) p(x_{n+1} | x_n)$, d'où (ii). (iii) est immédiat.

La calculabilité de MPM et MAP, même pour des tailles très importantes des chaînes (plusieurs millions), est à l'origine du succès des chaînes de Markov cachées.

Remarques

(i) Une fonction de perte est liée à la notion de « ressemblance », ou de « dissemblance », entre deux séquences. Définir un bon critère de ressemblance est un problème en soi, qui peut être très compliqué. Les fonctions de perte (2.11) et (2.14) semblent naturelles ; cependant, elles sont facilement critiquables. Par exemple, considérons deux séquences à deux classes « noir » et « blanc ». La première est (noir, blanc, noir, blanc, ...) et la deuxième (blanc, noir, blanc, noir, ...). Dans certaines applications, comme la segmentation d'images, on peut admettre que ces deux séquences sont les mêmes ; pourtant, la perte (2.11) est maximale. De manière similaire la perte (2.14) est la même dans le cas où deux séquences diffèrent par un sel point que dans celui où elles diffèrent sur tous les points.

(ii) La fonction de perte (2.11) peut être généralisée en posant

$$L(x^1, x^2) = \sum_{i=1}^n L_n(x_n^1, x_n^2), \text{ avec} \quad (2.19)$$

$$L_n(x_n^1 = \omega_i, x_n^2 = \omega_j) = \lambda_{ij}^n \quad (2.20)$$

La solution bayésienne est alors similaire à celle donnée par (2.7) dans le chapitre 1 :

$$[\hat{s}_B^L(y) = (\hat{x}_1, \dots, \hat{x}_N)] \Leftrightarrow$$

$$[\forall 1 \leq n \leq N, \forall 1 \leq j \leq k, \sum_{m=1}^k \lambda_{im}^n p(\hat{x}_n | y) \leq \sum_{m=1}^k \lambda_{jm}^n p(\hat{x}_n | y)] \quad (2.21)$$

Exemple

L'apport de la markovianité peut être spectaculaire. Les champs de Markov cachés (modèle différent des chaînes, mieux adapté aux images) permet d'estimer un champs observé, où l'œil humain ne distingue rien, avec une erreur raisonnable. Dans le même temps, bien qu'optimale, la méthode « point par point » (sans markovianité) donne des résultats inexploitable (Figure 7).


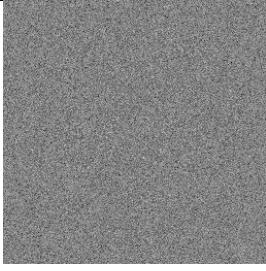


			
Image des classes $X = x$	Image observée $Y = y$	Segmentation « pixel par pixel »	Segmentation par champs de Markov

Figure 1. Image originale, la version bruitée (les deux bruits non gaussiens avec les mêmes moyennes et les mêmes mêmes variances), segmentation optimale par la méthode « pixel par pixel » et une méthode fondée sur le modèle par champ de Markov caché.

3. Estimation des paramètres des CMC

3.1. Paramètres des méthodes « locales ».

- Estimation avec échantillon d'apprentissage.

Considérons le cas de deux classes et supposons que les lois de Y conditionnelles à $X = \omega_1$, $X = \omega_2$ sont **gaussiennes**. Notons f_1, f_2 les densités correspondantes. Le paramètre θ a **six** composantes: $\pi_1 = P[X = \omega_1]$, $\pi_2 = P[X = \omega_2]$ (loi "a priori"), $\mu_1, \mu_2, \sigma_1, \sigma_2$: les moyennes et écarts type définissant f_1, f_2 . On dispose d'un échantillon x_1, x_2, \dots, x_n de réalisations de X ("échantillon d'apprentissage"). Les réalisations de Y étant toujours observables, on estime alors les paramètres θ à partir de $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Notons $x^1 = (x_1^1, \dots, x_m^1)$ les x_i valant ω_1 et $x^2 = (x_1^2, \dots, x_k^2)$ les x_i valant ω_2 . La loi a priori $\pi = (\pi_1, \pi_2)$ peut alors être estimée par les fréquences ($\frac{m}{n}$, $\frac{k}{n}$ respectivement) et μ_1, σ_1 (respectivement

μ_2, σ_2) par la moyenne et l'écart type empiriques à partir de $x^1 = (x_1^1, \dots, x_m^1)$ (respectivement $x^2 = (x_1^2, \dots, x_k^2)$).

- Estimation sans échantillon d'apprentissage.

(i) Algorithme SEM

Reprenons le cas de deux classes $\Omega = \{\omega_1, \omega_2\}$. On souhaite donc estimer $\theta = (\pi_1, \pi_2, m_1, m_2, \sigma_1^2, \sigma_2^2)$ à partir de $(Y_1, Y_2, \dots, Y_n) = (y_1, y_2, \dots, y_n)$, chaque Y_i suivant la loi mélange de densité $f_\theta(y_i) = \pi_1 f_1(y_i) + \pi_2 f_2(y_i)$, où f_1, f_2 sont les densités gaussiennes $N(m_1, \sigma_1^2)$, $N(m_2, \sigma_2^2)$. On peut retenir l'idée générale de l'algorithme SEM en se souvenant que si x_1, x_2, \dots, x_n était disponibles l'estimation du paramètre $\theta = (\pi_1, \pi_2, m_1, m_2, \sigma_1^2, \sigma_2^2)$ serait facile (voir l'estimation avec l'échantillon d'apprentissage ci-dessus). x_1, x_2, \dots, x_n n'étant pas disponible, on les "fabrique" de façon artificielle. Cette "fabrication" est faite au moyen de simulation. Le déroulement de l'algorithme est le suivant :

(i) on se donne la valeur initiale

$$\theta_0 = (\pi_1^0, \pi_2^0, m_1^0, m_2^0, \sigma_1^0, \sigma_2^0) \quad (3.1)$$

des paramètres ;

(ii) on calcul de θ_{k+1} à partir de θ_k et (y_1, y_2, \dots, y_n) de la manière suivante :

- on calcule, pour chaque $1 \leq i \leq n$ et en utilisant θ_k , la loi a posteriori sur $\Omega = \{\omega_1, \omega_2\}$ (la loi a posteriori de X_i est la loi de X_i de conditionnelle à $Y_i = y_i$, donnée par

$$\pi_1^{y_i} = \frac{\pi_1 f_1(y_i)}{\pi_1 f_1(y_i) + \pi_2 f_2(y_i)}, \quad \pi_2^{y_i} = \frac{\pi_2 f_2(y_i)}{\pi_1 f_1(y_i) + \pi_2 f_2(y_i)} :$$

$$\pi^{y_i, \theta_k} = (\pi_1^{y_i, \theta_k}, \pi_2^{y_i, \theta_k}) \quad (3.2)$$

- pour chaque $1 \leq i \leq n$ on effectue un tirage dans Λ selon la probabilité $\pi^{y_i, \theta_k} = (\pi_1^{y_i, \theta_k}, \pi_2^{y_i, \theta_k})$. On obtient ainsi une suite x_1^k, \dots, x_n^k .

- la nouvelle valeur θ_{k+1} des paramètres est donnée par les estimateurs empiriques appliqués à x_1^k, \dots, x_n^k et (y_1, y_2, \dots, y_n) .

On note que la suite aléatoire $(\theta_k)_{k \in \mathbb{N}}$ reste "indéfiniment aléatoire" (il n'y a pas de convergence vers une valeur fixe). Le comportement théorique de cette suite est difficile à étudier, cependant, les applications pratiques du SEM en imagerie ont donné des résultats tout à fait concluants. Dans la pratique, on se donne un critère de "stabilité" de la suite et on retient comme valeur estimée des paramètres la moyenne d'un certain nombre des dernières valeurs de la suite.

(ii) Algorithme EM

L'algorithme EM, qui est l'abréviation de "Expectation-Maximization" (qui signifie en anglais espérance-maximisation) est une méthode générale d'estimation des paramètres dans le cas des "données manquantes". De façon générale, considérons un couple de variables (X, Y) , dont la loi dépend d'un paramètre θ . et est donnée par une densité $f_\theta(X, Y)$. Le problème est d'estimer θ à partir de Y seul. Si X était observable, on pourrait envisager l'application du

maximum de vraisemblance, qui consiste en recherche de θ maximisant $\text{Log}[f_\theta(x, y)]$. X n'étant pas observable, on remplace la variable aléatoire $\text{Log}[f_\theta(X, Y)]$ par

$$E_\theta[\text{Log}[f_\theta(X, Y) | Y = y]] \quad (3.3)$$

On arrive alors à une méthode itérative: la valeur suivante du paramètre θ_{k+1} est donnée à partir de sa valeur courante θ_k et $Y = y$ par :

$$E_{\theta_k}[\text{Log}[f_{\theta_{k+1}}(X, Y) | Y = y]] = \max_{\theta} E_{\theta_k}[\text{Log}[f_{\theta}(X, Y) | Y = y]] \quad (3.4)$$

Il existe alors deux phases dans chaque itération (calcul de $E_{\theta_k}[\text{Log}[f_{\theta}(X, Y) | Y = y]]$ et sa maximisation, d'où l'appellation de la méthode).

Dans notre cas particulier nous avons :

$$f_\theta(y_1, \dots, y_n) = \prod_{i=1}^n (\pi_1 f_1(y_i) + \pi_2 f_2(y_i)) \quad (3.5)$$

avec $\theta = (\pi_1, \pi_2, m_1, m_2, \sigma_1^2, \sigma_2^2)$. Les deux phases sont calculables explicitement et on arrive à la méthode itérative suivante:

- (i) on se donne la valeur initiale $\theta_0 = (\pi_1^0, \pi_2^0, m_1^0, m_2^0, \sigma_1^0, \sigma_2^0)$ des paramètres ;
- (ii) on calcul θ_{k+1} à partir de θ_k et (y_1, y_2, \dots, y_n) de la manière suivante :

- on calcule, pour chaque $1 \leq i \leq n$ et en utilisant θ_k , la loi a posteriori $\pi^{y_i, \theta_k} = (\pi_1^{y_i, \theta_k}, \pi_2^{y_i, \theta_k})$ sur $\Omega = \{\omega_1, \omega_2\}$:

- la nouvelle valeur θ_{k+1} des paramètres est donnée par

$$\begin{aligned} \pi_1^{k+1} &= \frac{1}{n} (\pi_1^{y_1, \theta_k} + \dots + \pi_1^{y_n, \theta_k}) \\ m_1^{k+1} &= \frac{y_1 \pi_1^{y_1, \theta_k} + \dots + y_n \pi_1^{y_n, \theta_k}}{\pi_1^{y_1, \theta_k} + \dots + \pi_1^{y_n, \theta_k}} \\ (\sigma_1^{k+1})^2 &= \frac{(y_1 - m_1^{k+1})^2 \pi_1^{y_1, \theta_k} + \dots + (y_n - m_1^{k+1})^2 \pi_1^{y_n, \theta_k}}{\pi_1^{y_1, \theta_k} + \dots + \pi_1^{y_n, \theta_k}} \end{aligned} \quad (3.7)$$

et les formules analogues pour $\pi_2^{k+1}, m_2^{k+1}, \sigma_2^{k+1}$.

On peut retenir cet algorithme en se souvenant que la probabilité a priori est re-estimée par la moyenne des probabilités a posteriori, les moyennes sont re-estimées par les moyennes des observations "pondérées par les probabilités a posteriori", et les variances sont re-estimées par les variances empiriques en considérant des observations "pondérées par les probabilités a posteriori".

3.2. Paramètres des chaînes de Markov cachés.

Soit (X, Y) une chaîne de Markov cachée, que l'on supposera dans ce paragraphe stationnaire.

Les paramètres – pour m classes – sont donnés par $(\pi_{ij})_{1 \leq i, j \leq m}$ et $(m_i, \sigma_i^2)_{1 \leq i \leq m}$.

(i) Algorithme SEM

Soit $(X, Y) = (X_1, Y_1, \dots, X_N, Y_N)$ une CMC. L'algorithme SEM s'applique naturellement dans le contexte des CMC: d'une part, il existe des estimateurs de tous les paramètres à partir des données complètes $(X, Y) = (x, y)$ et, d'autre part, il est possible de simuler les réalisations de X selon sa loi conditionnelle à $Y = y$ (sa loi a posteriori). **En effet, la loi de X a posteriori est une loi markovienne autorisant les simulations rapides des réalisations.**

Afin de simplifier les écritures considérons le cas de deux classes. Nous avons $\theta = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}, m_1, m_2, \sigma_1^2, \sigma_2^2)$. La procédure se déroule de la façon suivante :

1. Initialisation des paramètres à estimer ;

2. A chaque itération k :

- Simulation d'une réalisation x^k de X selon la loi a posteriori basée sur les paramètres courants θ^k (rappelons que cette loi est markovienne, avec $p(x_1|y)$ et $p(x_{n+1}|x_n, y)$ calculées à partir des probabilités « forward, « backward », et (2.6) : on simule donc $X_1 = x_1^k$ selon $p(x_1|y)$, et les $X_n = x_n^k$ suivants en utilisant les $p(x_{n+1}|x_n, y)$;

- Re-estimation des paramètres (π_{ij}) par

$$\pi_{ij}^{k+1} = \frac{1}{N-1} \sum_{n=1}^{N-1} 1_{[x_n^k=i, x_{n+1}^k=j]} \quad (2.1)$$

$$m_i^{k+1} = \frac{\sum_{n=1}^N y_n 1_{[x_n^k=i]}}{\sum_{n=1}^N 1_{[x_n^k=i]}} \quad (2.2)$$

$$\sigma_i^{2,k+1} = \frac{\sum_{n=1}^N (y_n - m_i^{k+1})^2 1_{[x_n^k=i]}}{\sum_{n=1}^N 1_{[x_n^k=i]}} \quad (2.3)$$

(ii) Algorithme EM

Le principe général (3.4) du EM donne dans le cas des CMC :

- la nouvelle valeur θ_{k+1} des paramètres est donnée par

$$\begin{aligned} \pi_{ij}^{k+1} &= \frac{1}{N-1} (p_{ij,1}^k + \dots + p_{ij,N-1}^k) \\ m_1^{k+1} &= \frac{y_1 p_{i,1}^k + \dots + y_N p_{i,N}^k}{p_{i,1}^k + \dots + p_{i,N}^k} \end{aligned} \quad (2.4)$$

$$(\sigma_1^{k+1})^2 = \frac{(y_1 - m_1^{k+1})^2 p_{i,1}^k + \dots + (y_N - m_N^{k+1})^2 p_{i,N}^k}{p_{i,1}^k + \dots + p_{i,N}^k}$$

Les probabilités $p_{ij,n}^k$ sont définies par

$$p_{ij,n}^k = p(x_n = \omega_i, x_{n+1} = \omega_j | Y = y) \quad (2.5)$$

On montre

$$p_{ij,n}^k = \frac{\alpha_n(x_n) p(x_{n+1} = \omega_j | x_n = \omega_i, y) p(y_{n+1} | x_{n+1} = \omega_j) \beta_{n+1}(x_{n+1})}{\sum_{(\omega_i, \omega_j)} \alpha_n(x_n) p(x_{n+1} = \omega_j | x_n = \omega_i, y) p(y_{n+1} | x_{n+1} = \omega_j) \beta_{n+1}(x_{n+1})} \quad (2.6)$$

Algorithme EM se déroule de la façon suivante :

- Initialisation $\theta^0 = (\pi_{ij}^0, m_i^0, \sigma_i^{2,0})$ des paramètres pour $1 \leq i, j \leq m$
- A chaque itération k :
 - Étape "E" :
 - Calcul des probabilités $\alpha_n^k(x_n)$, et $\beta_n^k(x_n)$ avec $\theta^k = (\pi_{ij}^k, m_i^k, \sigma_i^{2,k})$;
 - Calcul de $p_{ij,n}^k$ à partir de $\alpha_n^k(x_n)$, et $\beta_n^k(x_n)$;
 - Étape "M": Calcul des paramètres $\theta^{k+1} = (\pi_{ij}^{k+1}, m_i^{k+1}, \sigma_i^{2,k+1})$ par (2.4)

Remarque 1

Le choix de l'initialisation des paramètres du modèle peut avoir une forte influence sur la rapidité de la convergence de l'algorithme EM. De plus il existe un risque de convergence vers un maximum local éloigné du maximum global de la fonction de vraisemblance de l'observation \mathcal{Y} . Il en résulte que dans certains cas la qualité de l'estimation peut être très moyenne. Cependant, EM s'avère le plus souvent très performant, surtout lorsque le bruit est gaussien.

Remarque 2

On dispose donc de méthodes, fondées sur les chaînes de Markov cachées (CMC), de « traitements non supervisés » : à partir des seules observations on peut d'abord estimer les paramètres, ensuite rechercher les données cachées. Les CMC s'avèrent robustes dans les problèmes mono-dimensionnels. Cependant, on peut les utiliser dans n'importe quel ensemble de variables dépendantes entre elles, e, définissant au préalable un parcours « mono-dimensionnel » dans l'ensemble en question.

Exercices

Exercice 1.

On considère que le passage d'un véhicule sur l'autoroute produit, en un point donné, une pollution dont la mesure (nombre réel) suit une loi normale de moyenne m et de variance σ^2 . Sachant que le nombre de véhicules passant entre les instants t et $t + \Delta t$ suit la loi de Poisson de paramètre $\lambda \Delta t$, quelle est la pollution moyenne produite entre t et $t + \Delta t$?

Exercice 2

On tire n individus dans une population d'humains et on note x_1, \dots, x_n leurs poids, et y_1, \dots, y_n leurs tailles. On suppose donc que le poids et la taille de l'individu $1 \leq i \leq n$ est une réalisation d'un vecteur aléatoire (X_i, Y_i) , que l'on supposera Gaussien.

1. Les variables X_i, Y_i sont-elles indépendantes ?
2. Quels sont les paramètres définissant la loi de (X_i, Y_i) ?
3. Pour un individu j dont on n'observe que le poids x_j , comment est modélisée l'information que x_j procure sur sa taille y_j ?
4. On voudrait proposer une "stratégie \hat{s} " de "prédiction" de la taille à partir du poids $\hat{y}_j = \hat{s}(x_j)$, telle que à la longue, l'erreur $(\hat{y}_1 - y_1)^2 + \dots + (\hat{y}_m - y_m)^2$ soit minimale. Comment faire ? (on appliquera d'abord la loi des grands nombres, ensuite la formule de l'espérance conditionnelle).
5. Transposer le problème en celui de la prédiction du cours d'une action en bourse.

Exercice 3

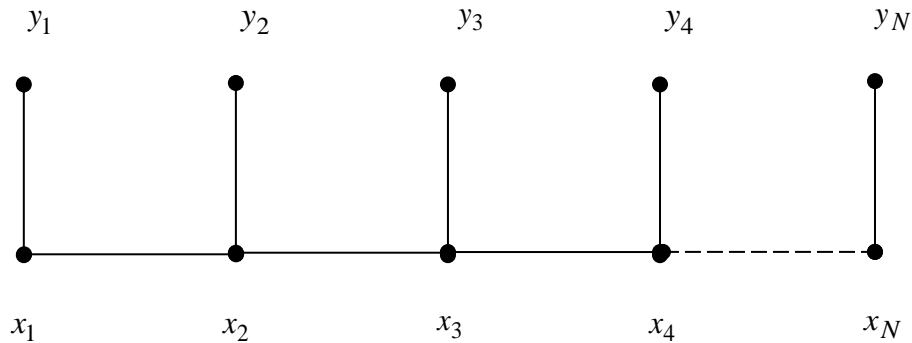
Un élève prépare un examen de statistique et, après y avoir consacré une durée de temps x , il obtient la note y . En normalisant, on suppose que ces valeurs sont dans $[0,1]$. Le lien entre x et y n'étant pas déterministe, on le modélise en considérant un vecteur aléatoire (X, Y) à valeurs dans $[0,1] \times [0,1]$. On supposera que sa loi admet pour densité $f_{(X,Y)}(x, y) = -2x^2 + 4x^2y + 2x$.

1. Montrer que les densités des lois marginales sont $f_X(x) = 2x$, $f_Y(y) = \frac{4}{3}y + \frac{1}{3}$.
Montrer que la durée moyenne du temps consacré à la préparation (l'espérance de X) de l'examen est de $\frac{2}{3}$, et la note moyenne (l'espérance de Y) est de $\frac{11}{18}$;
2. Montrer que la note obtenue dépend, au sens probabiliste, de la durée du travail (les variables X et Y ne sont pas indépendantes) ;
3. Un élève ayant consacré la durée $X = x$ au travail souhaite évaluer, avant l'examen, sa note. Proposez une prédiction utilisant la notion de l'espérance conditionnelle. En quoi cette prédiction est-elle optimale ? Peut-on dire que le travail paie ?

4. Quel est le temps minimal qu'un élève doit consacrer à la préparation de l'examen pour que la probabilité d'obtenir la moyenne (ou plus) soit supérieure ou égale à $\frac{2}{3}$?

Exercice 4

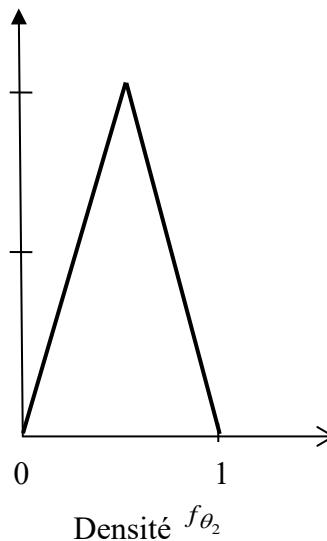
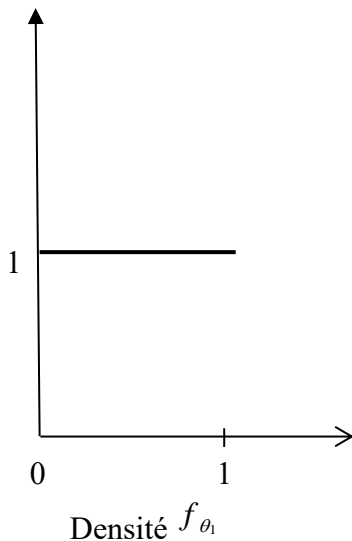
Montrer que le graphe non orienté minimal des dépendances d'une chaîne de Markov cachée est de la forme



Indication : montrer que les axiomes des CMC sont vérifiés ; ensuite que le graphe est minimal (on ne peut pas d'enlever d'arrêtes).

Exercice 5

On considère (X, Y) couple de variables aléatoires à valeurs dans $\{\theta_1, \theta_2\} \times [0, 1]$. La loi de X est donnée par $\pi_{\theta_1} = \pi_{\theta_2} = 0.5$ et les lois de Y conditionnelles à $X = \theta_1$ et $X = \theta_2$ sont données par les densités $f_{\theta_1}, f_{\theta_2}$ respectivement.



$$L(\theta_i, \theta_j) = \begin{cases} 0 & \text{si } \theta_i = \theta_j \\ 1 & \text{si } \theta_i \neq \theta_j \end{cases}$$

Par ailleurs, on considère la fonction de perte

1. Donner la stratégie bayésienne correspondante (on donnera la partition de $[0, 1]$ à partir des formules définissant la stratégie bayésienne sans chercher à les démontrer)

2. Calculer la probabilité de se tromper. Comment s'interprète cette probabilité en termes de taux d'individus mal classés? Est-il possible d'imaginer une stratégie permettant d'améliorer ce taux?

Exercice 6

On considère une chaîne de Markov cachée (X, Y) , avec $X = (X_1, \dots, X_n)$ caché et $Y = (Y_1, \dots, Y_n)$ observé, dont la loi est donc donnée par

$$p(x, y) = p(x_1)p(x_2|x_1)\dots p(x_n|x_{n-1})p(y_1|x_1)\dots p(y_n|x_n). \quad (1)$$

Les X_i prennent leurs valeurs dans $\Omega = \{\omega_1, \omega_2\}$ et les Y_i sont à valeurs réelles. Le graphe des dépendances (pour $n = 6$) est présenté à la Figure 1.

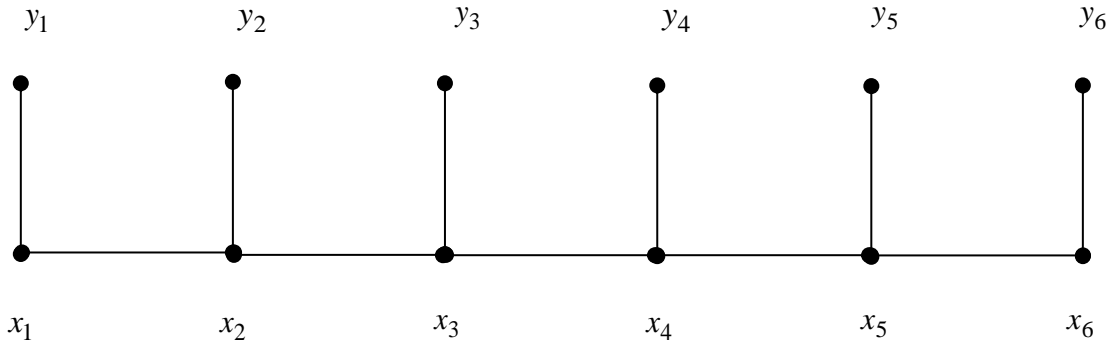


Figure 1. Graphe des dépendances d'une chaîne de Markov cachée.

1. On suppose que les instants $1, 2, \dots, 6$ correspondent aux jours qui passent, le jour 3 étant le jour présent. Ainsi Y_1, Y_2 sont les observations passées, Y_3 est l'observation du jour, et Y_4, Y_5, Y_6 ne sont pas disponibles. Dessiner le graphe des dépendances correspondant à l'ensemble des variables $X_1, X_2, X_3, X_4, X_5, X_6, Y_1, Y_2, Y_3$. Montrez, à partir du graphe, que

$$p(x_6|y_1, y_2, y_3) = \sum_{x_3} p(x_6|x_3)p(x_3|y_1, y_2, y_3). \quad (3)$$

2. On souhaite « prédire » X_6 (l'estimer à partir de Y_1, Y_2, Y_3) par la méthode

$$L(\omega_i, \omega_j) = \begin{cases} 0 & \text{si } \omega_i = \omega_j \\ 1 & \text{si } \omega_i \neq \omega_j \end{cases}$$

bayésienne correspondante à la fonction de perte

à partir des probabilités $p(x_6|y_1, y_2, y_3)$. Expliquer brièvement comment elles sont calculables à partir des probabilités locales dans (1).

Exercice 7

On considère deux vecteurs aléatoires $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_n)$. Chaque X_i est à valeurs dans l'ensemble de classes $\Omega = \{1, 2, 3\}$, et chaque Y_i est à valeurs dans R . On observe $Y = y$ et on souhaite "estimer", en utilisant les techniques Bayésiennes de classification, la réalisation invisible x de X . On se propose d'adopter une modélisation par chaînes de Markov cachées stationnaire, avec les lois $p(y_i|x_i)$ gaussiennes.

1. Préciser les paramètres nécessaires à la mise en œuvre de la méthode MPM;
2. En supposant $n = 30$, on souhaite estimer les paramètres par la méthode SEM. A l'itération k , le tirage dans $\Omega = \{1, 2, 3\}$ a donné :

$$x^k = (x_1^k, \dots, x_{30}^k) = (1, 1, 3, 3, 2, 3, 3, 1, 1, 1, 2, 2, 2, 2, 1, 1, 1, 3, 2, 2, 2, 2, 2, 1, 1, 3, 2, 1)$$

quelle est la valeur suivante des paramètres?

Solutions des Exercices

Exercice 1.

Notons N le nombre (aléatoire) de véhicules passés entre t et $t + \Delta t$, et Y la pollution dégagée dans le même intervalle de temps. Si N était connu le problème serait simple ; on va donc essayer d'appliquer la formule (2.10), avec $\phi(Y, N) = Y$, en conditionnant par N . Nous avons

$$E[Y] = E[E[Y|N]] = \sum_{n \in \mathbb{N}} E[Y|N=n] P_N[n]$$

Sachant que $P_N[n] = \frac{(\lambda \Delta t)^n}{n!} e^{-\lambda \Delta t}$ (loi de Poisson, de moyenne $E[N] = \lambda \Delta t$) et $E[Y|N=n] = nm$ (la moyenne de la pollution dégagée par n véhicules est n fois la moyenne de celle dégagée par 1 véhicule), on a :

$$E[Y] = m \sum_{n \in \mathbb{N}} n P_N[n] = m E[N] = m \lambda \Delta t$$

Exercice 2.

1. En principe non car il n'est pas vrai que le poids "n'a rien à voir" avec la taille, du moins lorsqu'il s'agit des humains.

2. La loi d'un vecteur Gaussien à deux composantes est donnée par 5 paramètres: deux moyennes $m_X = E[X]$, $m_Y = E[Y]$, deux variances, $\sigma_X^2 = E[(X - m_X)^2]$, $\sigma_Y^2 = E[(Y - m_Y)^2]$ et une covariance $c_{XY} = E[(X - m_X)(Y - m_Y)]$.

3. L'information que le poids x_j d'un individu procure sur la taille y_j est contenue dans la loi de Y_j conditionnelle à $X_j = x_j$. Dans le cas Gaussien, cette loi est une loi Gaussienne de

$$\text{moyenne } m_Y^{X=x} = m_Y + \frac{c_{XY}}{\sigma_X^2} (x - m_X) \quad \text{et de variance } (\sigma_Y^{X=x})^2 = \sigma_Y^2 - \frac{(c_{XY})^2}{\sigma_X^2}.$$

4. D'après la loi des grands nombres

$$\frac{(\hat{Y}_1 - Y_1)^2 + \dots + (\hat{Y}_m - Y_m)^2}{m} \xrightarrow{m \rightarrow +\infty} E[(\hat{Y} - Y)^2] = E[(\hat{s}(X) - Y)^2]$$

Donc, à long terme, la stratégie qui minimise la perte est celle qui minimise $E[(\hat{s}(X) - Y)^2]$. D'après (2.10) nous avons $E[(\hat{s}(X) - Y)^2] = E[E[(\hat{s}(X) - Y)^2|X]]$. À $X = x$ fixé on a $E[(\hat{s}(x) - Y)^2|X = x]$, qui est une parabole en $\hat{s}(x)$: le minimum est en $\hat{s}(x) = E[Y|X = x]$ (c'est une propriété très importante de l'espérance conditionnelle, qui est donc la meilleure approximation, au sens de l'erreur quadratique moyenne, d'une variable aléatoire par une fonction d'une autre variable).

Par ailleurs, l'espérance conditionnelle est également l'espérance selon la loi conditionnelle, qui

$$\text{est ici une Gaussienne de moyenne } m_Y^{X=x} = m_Y + \frac{c_{XY}}{\sigma_X^2} (x - m_X) \quad \text{et de variance } (\sigma_Y^{X=x})^2 = \sigma_Y^2 - \frac{(c_{XY})^2}{\sigma_X^2} \quad (\text{voir le point précédent}). \text{ Nous avons donc}$$

$$\hat{s}_{OPTIMALE}(x) = m_Y + \frac{c_{XY}}{\sigma_Y^2} (x - m_X)$$

Nous pouvons donc affirmer que pour n assez grand, il n'existe pas de méthode, probabiliste ou pas, qui donne une erreur $(\hat{y}_1 - y_1)^2 + \dots (\hat{y}_m - y_m)^2$ plus petite que celle donnée par l'utilisation de $\hat{s}_{OPTIMALE}$ donnée ci-dessus.

Notons que l'utilisation d'un autre critère d'erreur, comme par exemple la valeur absolue à la place de carré, donnerait une méthode optimale différente.

5. Si x est le cours d'aujourd'hui et y le cours de demain, on peut faire de la prédiction par la formule du point précédent dès que l'on connaît la loi gaussienne du couple (X, Y) .

Exercice 3

$$1. \quad f_X(x) = \int_0^1 (-2x^2 + 4x^2y + 2x)dy = 2x, \quad f_Y(y) = \int_0^1 (-2x^2 + 4x^2y + 2x)dx = \frac{4}{3}y + \frac{1}{3};$$

$$E[X] = \int_0^1 xf_X(x)dx = \int_0^1 2x^2dx = \frac{2}{3}; \quad E[Y] = \int_0^1 yf_Y(y)dy = \int_0^1 y(\frac{4}{3}y + \frac{1}{3})dy = \frac{11}{18}.$$

2. Si les variables étaient indépendantes on aurait $f_{(X,Y)}(x,y) = f_{(X)}(x)f_{(Y)}(y)$. Or $f_{(X)}(x)f_{(Y)}(y) = 2x(\frac{4}{3}y + \frac{1}{3}) \neq f_{(X,Y)}(x,y)$

3. L'espérance conditionnelle est la prédiction minimisant l'erreur quadratique moyenne. La loi de Y conditionnelle à $X = x$ est donnée par la densité

$$f_{(Y/X=x)}(y) = \frac{f_{(X,Y)}(x,y)}{f_{(X)}(x)} = \frac{-2x^2 + 4x^2y + 2x}{2x} = 2xy - x + 1.$$

L'espérance conditionnelle est

alors l'espérance selon la loi conditionnelle $f_{(Y/X=x)}$, ce qui donne

$$E[Y/X=x] = \int_0^1 yf_{(Y/X=x)}(y)dy = \int_0^1 y(2xy - x + 1)dy = \int_0^1 (2xy^2 - xy + y)dy = \frac{2x}{3} - \frac{x}{2} + \frac{1}{2} = \frac{x}{6} + \frac{1}{2}$$

C'est une fonction croissante de x , ce qui signifie que le travail paie « en moyenne » (notons cependant qu'un élève peut travailler davantage qu'un autre et obtenir une plus mauvaise note quand même).

4. Après avoir travaillé pendant le temps x la probabilité d'avoir la moyenne est

$$\int_{\frac{1}{2}}^1 f_{(Y/X=x)}(y)dy = \int_{\frac{1}{2}}^1 (2xy - x + 1)dy = \frac{x}{4} + \frac{1}{2}$$

Qui est une fonction croissante de x . On a $\frac{x}{4} + \frac{1}{2} = \frac{2}{3}$ pour $x = \frac{2}{3}$, donc cette probabilité est supérieure ou égale $\frac{2}{3}$ pour x supérieure ou égale à $\frac{2}{3}$.

Exercice 4

On constate, en appliquant les règles de conditionnement et de marginalisation, qu'une distribution compatible avec le graphe vérifie les trois hypothèses définissant une CMC (X de

Markov, $p(y|x) = \prod_{n=1}^N p(y_n|x)$, $p(y_n|x) = p(y_n|x_n)$ pour $n = 1, \dots, N$). Par ailleurs, c'est un graphe minimal. En effet, on ne peut enlever ni un segment vertical ni un segment horizontal.

Exercice 5

1. La stratégie bayésienne correspondante à la fonction de perte en question est donnée par :

$$\hat{s}_B(y) = \begin{cases} \theta_1 & \text{si } \Pi_1 f_1(y) \geq \Pi_2 f_2(y) \\ \theta_2 & \text{si } \Pi_1 f_1(y) \leq \Pi_2 f_2(y) \end{cases}$$

Étant donné que $\Pi_1 = \Pi_2$, l'allure des courbes f_1, f_2 permet d'écrire :

$$\hat{s}_B(y) = \begin{cases} \theta_1 & \text{si } f_1(y) \geq f_2(y) \\ \theta_2 & \text{si } f_1(y) \leq f_2(y) \end{cases} = \begin{cases} \theta_1 & \text{si } y \in [0, \frac{1}{4}] \cup [\frac{3}{4}, 1] \\ \theta_2 & \text{si } y \in [\frac{1}{4}, \frac{3}{4}] \end{cases}$$

2.

$$\begin{aligned} P[\hat{s}_B \text{ setrompe}] &= P[\hat{s}_B(Y) = \theta_1 \text{ et } X = \theta_2] + P[\hat{s}_B(Y) = \theta_2 \text{ et } X = \theta_1] = \\ &= P[\hat{s}_B(Y) = \theta_1 / X = \theta_2]P[X = \theta_2] + P[\hat{s}_B(Y) = \theta_2 / X = \theta_1]P[X = \theta_1] = \\ &= \left(\frac{1}{4}\right)\frac{1}{2} + \left(\frac{3}{4} - \frac{1}{4}\right)\frac{1}{2} = \frac{3}{8} \end{aligned}$$

Par la loi des grands nombres le taux d'individus mal classés τ_n tend vers la probabilité de se tromper, en effet

$$\tau_n = \frac{L(\hat{s}_B(Y_1), X_1) + \dots + L(\hat{s}_B(Y_n), X_n)}{n} \xrightarrow{n \rightarrow +\infty} E[L(\hat{s}_B(Y), X)] = P[\hat{s}_B \text{ setrompe}]$$

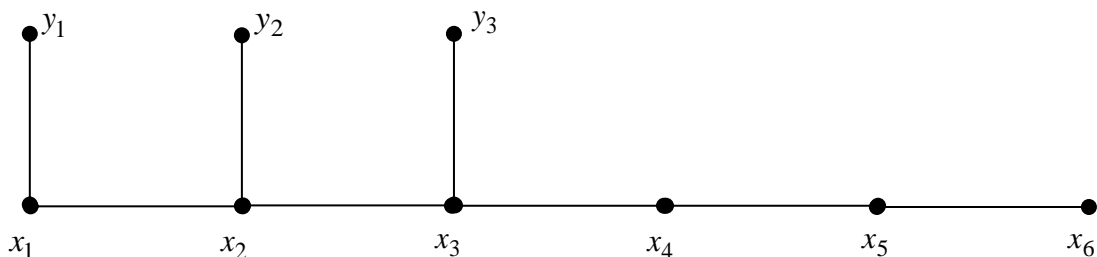
et donc

$$\tau_n \xrightarrow{n \rightarrow +\infty} P[\hat{s}_B \text{ setrompe}]$$

Selon la théorie la stratégie bayésienne minimise $E[L(\hat{s}_B(Y), X)]$, donc, dans ce cas précis, la probabilité de se tromper. Il est en conséquence impossible de trouver une autre méthode de classification procurant un taux plus petit (du moins lorsque le nombre des classifications justifie l'application de la loi des grands nombres).

Exercice 6

1. En appliquant les règles de marginalisation dans les graphes on trouve



Selon le graphe X_6 et (Y_1, Y_2, Y_3) sont indépendants conditionnellement à X_3 ; on peut donc écrire

$$p(x_6|y_1, y_2, y_3) = \sum_{x_3} p(x_6, x_3|y_1, y_2, y_3) = \sum_{x_3} p(x_3|y_1, y_2, y_3) p(x_6|x_3, y_1, y_2, y_3) = \sum_{x_3} p(x_3|y_1, y_2, y_3) p(x_6|x_3).$$

2. Pour la fonction de perte considérée la stratégie bayésienne est

$$\hat{s}(y_1, y_2, y_3) = \begin{cases} \omega_1 & \text{si } p(x_6 = \omega_1|y_1, y_2, y_3) \geq p(x_6 = \omega_2|y_1, y_2, y_3) \\ \omega_2 & \text{si } p(x_6 = \omega_1|y_1, y_2, y_3) < p(x_6 = \omega_2|y_1, y_2, y_3) \end{cases}.$$

La probabilité $p(x_3|y_1, y_2, y_3)$ est une marginale a posteriori : elle est calculable par la procédure « forward-backward ». Par ailleurs, (X_3, X_4, X_5, X_6) est une chaîne de Markov, la matrice de transition de X_3 à X_6 , qui donne $p(x_6|x_3)$, est le produit des matrices de transition de X_3 à X_4 , de X_4 à X_5 , et de X_5 à X_6 . Ayant $p(x_6|x_3)$ et $p(x_3|y_1, y_2, y_3)$ on utilise le résultat de 1.

On note que le modèle permet de « prédire » X_6 . De manière générale, on peut prédire X_{n+k} à partir de (Y_1, \dots, Y_n) , même pour grands (plusieurs millions). Cependant, (Y_1, \dots, Y_n) et sont de moins en moins « dépendant », et donc la prédiction de moins en moins précise.

Exercice 7

1. Neuf paramètres $\pi_{ij} = p(x_1 = i, x_2 = j)$ et six paramètres (m_i, σ_i^2) des trois gaussiennes $p(y_1|x_1 = i)$.

2. On fait comme si x^k était une réalisation de X et on applique les estimateurs classiques : les paramètres suivants sont

$$\pi_{11}^{k+1} = \frac{8}{29}, \pi_{12}^{k+1} = \frac{1}{29}, \pi_{13}^{k+1} = \frac{3}{29}, \dots$$

$$m_1^{k+1} = \frac{y_1 + y_2 + y_8 + y_9 + y_{10} + y_{15} + y_{16} + y_{17} + y_{18} + y_{25} + y_{26} + y_{27} + y_{30}}{13}$$

$$m_2^{k+1} = \frac{y_5 + y_{11} + y_{12} + y_{13} + y_{14} + y_{20} + y_{21} + y_{22} + y_{23} + y_{24} + y_{29}}{11}$$

$$m_3^{k+1} = \frac{y_3 + y_4 + y_6 + y_7 + y_{19} + y_{28}}{6}$$

$$\sigma_1^{2,k+1} = \frac{(y_1 - m_1^{k+1})^2 + (y_2 - m_1^{k+1})^2 + (y_8 - m_1^{k+1})^2 + \dots + (y_{30} - m_1^{k+1})^2}{13}$$

$$\sigma_2^{2,k+1} = \frac{(y_5 - m_2^{k+1})^2 + (y_{11} - m_2^{k+1})^2 + \dots + (y_{29} - m_2^{k+1})^2}{11}$$

$$\sigma_3^{2,k+1} = \frac{(y_3 - m_3^{k+1})^2 + (y_4 - m_3^{k+1})^2 + \dots + (y_{28} - m_3^{k+1})^2}{6}$$