Exercice 1 *(4 pts)* Soit (X, Y) un couple de variables aléatoires réelles et soit $((x_1, y_1), \ldots, (x_n, y_n))$ un échantillon de n observations. Chacune des situations présentées dans la Figure 1 représente le nuage de données d'un échantillon de taille $n = 500$. Pour chaque situation, donner une valeur approchée du cœfficient de corrélation linéaire empirique r et justifier votre réponse.

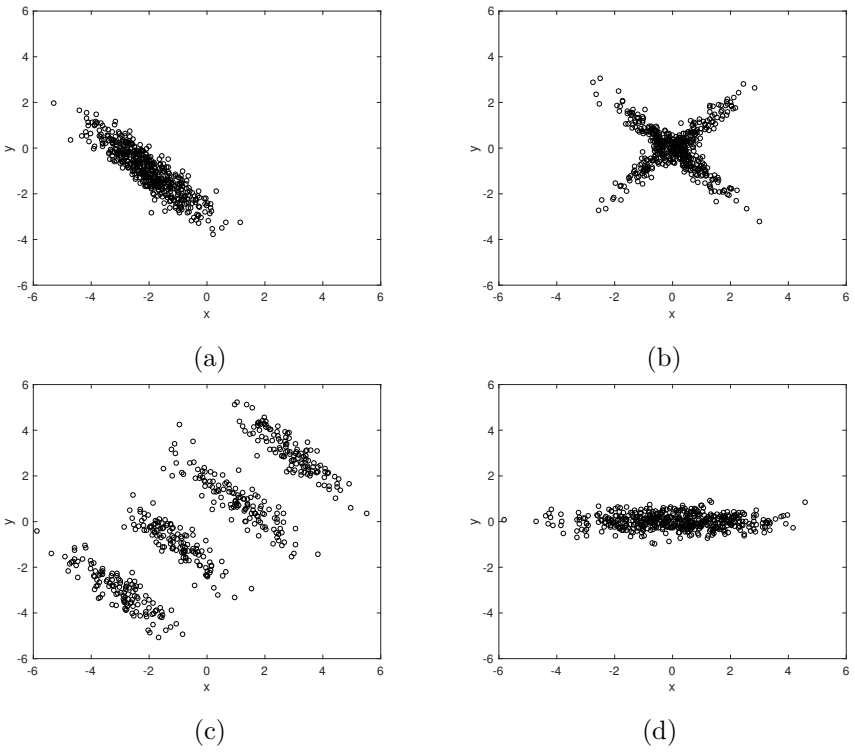


FIGURE 1 – Nuages de données (∘)

Solution 1 Voir TD ((a) forte relation linéaire (anti-corrélation), r proche de -0.8/.7 ; (b) r proche de 0 car pas de lien linéaire ; (c) r proche de 0.7 (paradoxe de Simpson) ; (d) r proche de 0 : quasiment pas de lien entre les deux variables.

Exercice 2 *(4 pts)* On considère un échantillon aléatoire indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ du couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est un vecteur composé de $p = 2$ prédicteurs réels ($\boldsymbol{X} \in \mathbb{R}^2$) et $Y \in [\![0, 1]\!]$ une variable à prédire. On dispose d'un échantillon observé d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et on cherche à prédire les classes de nouvelles observations $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m)$ sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère l'analyse discriminante où la densité de la classe $k \in [\![0, 1]\!]$ est définie par : $f(\boldsymbol{x}_i | Y_i = k; \boldsymbol{\theta}) = \frac{1}{2\pi |\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top {\boldsymbol{\Sigma}_k}^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)$. La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori :

$$\widehat{y}_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}),$$

$\boldsymbol{\theta}$ étant le vecteur paramètre du modèle. On note par $\pi_k = \mathbb{P}(Y_i = k)$, la probabilité a priori de la classe k. On suppose que $\pi_0 = 0.5$, $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 = \mathbf{g}$, $\boldsymbol{\Sigma}_0 = \mathbf{I}$ et $\boldsymbol{\Sigma}_1 = \lambda \boldsymbol{\Sigma}_0$, où $\lambda > 1$ et $\mathbf{I}$ est la matrice identité.

1. Montrer que $Y = 1$ si et seulement si $||\boldsymbol{x} - \mathbf{g}||_2^2 \geq r$, en déterminant r en fonction de λ.

2. A quoi correspond la frontière de décision dans ce cas ?

Solution 2

1. On a $Y = 1$ si et seulement si :

$$\begin{aligned}
\ln \frac{\mathbb{P}(y=1|\mathbf{x})}{\mathbb{P}(y=0|\mathbf{x})} &= \ln \frac{\pi_1}{\pi_0} - \frac{1}{2} \ln \frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_0|} - \frac{1}{2} \left\{ (\mathbf{x}-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}_1^{-1} (\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_0)^T \mathbf{\Sigma}_0^{-1} (\mathbf{x}-\boldsymbol{\mu}_0) \right\} \geq 0 \\
&= -\frac{1}{2} \ln(\lambda^2) - \frac{1}{2} \left\{ (\mathbf{x}-\boldsymbol{g})^T \frac{1}{\lambda} (\mathbf{x}-\boldsymbol{g}) - (\mathbf{x}-\boldsymbol{g})^T (\mathbf{x}-\boldsymbol{g}) \right\} \geq 0 \\
&= -\ln(\lambda) - \frac{1}{2} \left\{ ||\boldsymbol{x}-\mathbf{g}||_2^2 \left(\frac{1}{\lambda} - 1 \right) \right\} \geq 0 \\
&= \frac{1}{2} ||\boldsymbol{x}-\mathbf{g}||_2^2 \left(\frac{\lambda - 1}{\lambda} \right) \geq \ln(\lambda) \\
&= ||\boldsymbol{x}-\mathbf{g}||_2^2 \geq 2\ln(\lambda) \left(\frac{\lambda}{\lambda - 1} \right) \qquad (1)
\end{aligned}$$

   Au finale on a donc : $Y = 1$ si et seulement si $||\boldsymbol{x}-\mathbf{g}||_2^2 \geq r$ avec $r = 2\ln(\lambda)\left(\frac{\lambda}{\lambda-1}\right)$.

2. D'après la question précédent, la frontière de décision est définie par l'équation :

$$||\boldsymbol{x}-\mathbf{g}||_2^2 = 2\ln(\lambda) \left(\frac{\lambda}{\lambda - 1} \right)$$

   qui est celle d'un cercle de centre $\mathbf{g}$ et de rayon $\sqrt{2\ln(\lambda)\left(\frac{\lambda}{\lambda-1}\right)}$.

Exercice 3 *(4 pts)* On considère le cadre de l'exercice précédent et on se propose maintenant d'utiliser un autre modèle de prédiction, celui de régression logistique non-linéaire suivant défini par

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(\phi(\boldsymbol{x}; \boldsymbol{\theta}))}{1 + \exp(\phi(\boldsymbol{x}; \boldsymbol{\theta}))} \qquad (2)$$

où $\phi(\boldsymbol{x}; \boldsymbol{\theta})$ étant une transformation non-linéaire (polynomiale) de $\boldsymbol{x}$ et $\boldsymbol{\theta}$ le vecteur paramètre du modèle. La prédiction avec ce modèle consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori, i.e. :

$$\widehat{y}_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k|\boldsymbol{X}_i = \boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}),$$

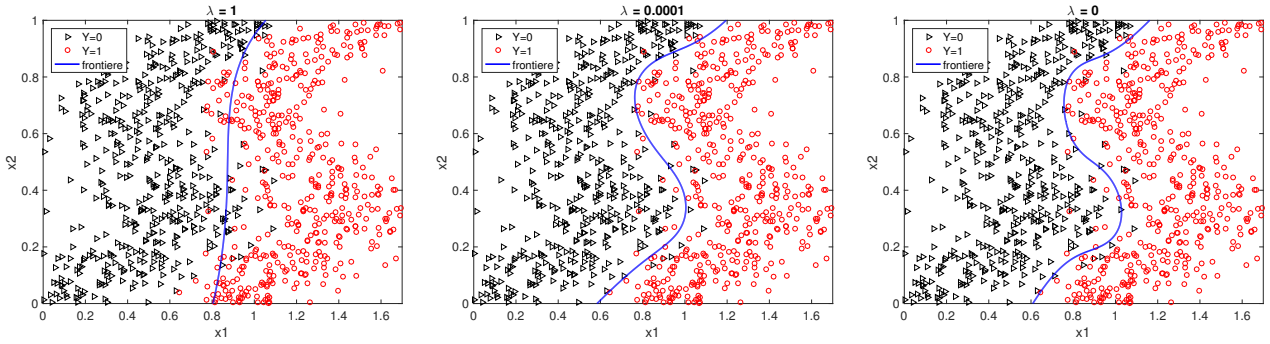
où $\widehat{\boldsymbol{\theta}}$ étant le vecteur paramètre du modèle appris en minimisant le risque quadratique régularisé suivant :

$$\ell_\lambda(\boldsymbol{\theta}) = -\ln L(\boldsymbol{\theta}) + \lambda \Omega(\boldsymbol{\theta}) \qquad (3)$$

où $\ln L(\boldsymbol{\theta})$ est la fonction de log-vraisemblance classique et $\Omega(\boldsymbol{\theta})$ une fonction de pénalité sur $\boldsymbol{\theta}$ de niveau λ dont l'objectif est d'assurer un compromis entre complexité du modèle et ajustement aux données.

1. Rappeler le nom et le principe de l'algorithme d'estimation des paramètres en régression logistique.
2. Rappeler la décomposition biais-variance du risque quadratique.
3. Chacune des figures 2-(a, b, c) montre un échantillon d'apprentissage et un modèle de prédiction estimé $\widehat{\boldsymbol{\theta}}$ en minimisant (3), représenté par la frontière de décision, pour la même régularisation $\Omega(\boldsymbol{\theta})$ et différentes valeurs de λ. Discuter la qualité prédictive de chacun des trois modèles.

Solution 3



(a) (b) (c)

FIGURE 2 – Nuage de données (○,▷) et modèle de prédiction (—)

1. Voir cours/td
2. Voir cours/td
3. On peut voir sur le graphique (a), pour lequel le niveau de pénalisation λ pour apprendre le modèle, est le plus fort parmi les trois situations, le modèle prédictif est un peu simpliste ; il s'agit d'un sous-apprentissage. Le modèle du graphique (c) correspond quant à lui à l'estimateur du maximum de vraisemblance non-régularisé ($\lambda = 0$), et on peut voir qu'il sur-apprend un peu et devient spécifique aux données d'apprentissage. Le meilleur modèle est celui du graphique (b) et correspond au meilleur compromis entre l'attache aux données et la complexité du modèle. D'un point de vue décomposition biais-variance du risque quadratique, on a :
   - modèle (a) : grand biais mais petite variance,
   - modèle (c) : petit biais mais grande variance,
   - modèle (b) : biais et variance tous les deux petits.

Exercice 4 *(8 pts)* On considère un échantillon aléatoire indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ du couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est un vecteur de p prédicteurs binaires ($\boldsymbol{X} \in [\![0,1]\!]^p$) et $Y \in [\![0,1]\!]$ une variable à prédire représentant la classe de $\boldsymbol{X}$. On dispose d'un échantillon observé d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ et on cherche à prédire les classes de nouvelles observations $(\boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_m)$ sur la base d'un modèle probabiliste appris sur les données d'apprentissage. On considère la classifieur Bayésien naïf, une méthode probabiliste de prédiction qui repose sur l'hypothèse simplificatrice forte suivante :

$\mathcal{H}$: Pour chaque classe Y_i, les variables $X_{ij}, j = 1, \ldots, p$ de chaque individu $\boldsymbol{X}_i$, sont indépendantes.

La prédiction s'effectue par la règle du maximum a posteriori (MAP) qui consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe y_i maximisant la probabilité a posteriori :

$$\widehat{y}_i = \arg \max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}), \tag{4}$$

$\widehat{\boldsymbol{\Psi}}$ étant l'estimateur du vecteur paramètre $\boldsymbol{\Psi}$ du modèle estimé sur l'ensemble d'apprentissage.

1. Les variables binaires X_{ij} de chaque vecteur $\boldsymbol{X}_i$ de la même classe $Y_i = k$ sont de loi Bernoulli de

paramètre $0 < \theta_{kj} < 1$, i.e. :

$$\mathbb{P}(X_{ij} = x_{ij}|Y_i = k; \theta_{kj}) = {\theta_{kj}}^{x_{ij}}(1-\theta_{kj})^{1-x_{ij}} \text{ où } \theta_{kj} = \mathbb{P}(X_{ij} = 1|Y_i = k; \theta_{kj}).$$

En déduire d'après $\mathcal{H}$ la loi conditionnelle du vecteur $\boldsymbol{X}_i$, notée $\mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i|Y_i = k; \boldsymbol{\theta}_k)$.

2. On note par $\alpha = \mathbb{P}(Y_i = 1)$, la probabilité a priori de la classe 1. On a alors $\boldsymbol{\Psi} = (\alpha, \boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_0^\top)^\top$ que l'on cherche à estimer la méthode du maximum de vraisemblance. Donner l'expression de la fonction de vraisemblance

$$L(\boldsymbol{\Psi}) = \mathbb{P}\big((\boldsymbol{X}_1 = \boldsymbol{x}_1, Y_1 = y_1), \ldots, (\boldsymbol{X}_n = \boldsymbol{x}_n, Y_n = y_n); \boldsymbol{\Psi}\big). \tag{5}$$

3. En déduire celle de la fonction de log-vraisemblance $\ln L(\boldsymbol{\Psi})$.

4. Montrer en maximisant la log-vraisemblance que les estimateurs du maximum de vraisemblance sont donnés par :

$$\widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} Y_i; \quad \widehat{\theta}_{1j} = \frac{\sum_{i=1}^n Y_i X_{ij}}{\sum_{i=1}^n Y_i}; \quad \widehat{\theta}_{0j} = \frac{\sum_{i=1}^n (1-Y_i) X_{ij}}{\sum_{i=1}^n (1-Y_i)}.$$

soit en formulation vectorielle pour les $\boldsymbol{\theta}$ par :

$$\widehat{\boldsymbol{\theta}}_1 = \frac{\sum_{i=1}^n Y_i \boldsymbol{X}_i}{\sum_{i=1}^n Y_i}; \quad \widehat{\boldsymbol{\theta}}_0 = \frac{\sum_{i=1}^n (1-Y_i) \boldsymbol{X}_i}{\sum_{i=1}^n (1-Y_i)}.$$

5. On considère le jeu de données de la Table 1. Prédire en appliquant la règle (4) la valeur de Y pour le dernier vecteur.

$\boldsymbol{X}$		Y
0	1	1
1	0	1
1	0	0
0	1	0
0	1	1
1	1	?

TABLE 1 – Jeu de données

Solution 4

1. Pour cela il suffit d'exploiter l'hypothèse $\mathcal{H}$, qui correspond à une hypothèse d'indépendance conditionnelle des variables. Par cette indépendance conditionnelle on peut donc écrire : étant le paramètre de la loi de Bernoulli de la variable binaire $x_{ij}, j = 1 \ldots, d$. Cette loi conditionnelle est la loi de Bernoulli multivariée pour la classe k

$$\begin{aligned} \mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i|Y_i = k; \boldsymbol{\theta}_k) &= \prod_{j=1}^{p} \mathbb{P}(X_{ij} = x_{ij}|Y_i = k; \theta_{kj}) \\ &= \prod_{j=1}^{p} {\theta_{kj}}^{x_{ij}}(1-\theta_{kj})^{1-x_{ij}}. \end{aligned} \tag{6}$$

2. L'échantillon étant indépendant, la fonction de vraisemblance est donc donnée par :

$$\begin{aligned} L(\boldsymbol{\Psi}) &= \mathbb{P}\big((\boldsymbol{X}_1 = \boldsymbol{x}_1, Y_1 = y_1), \ldots, (\boldsymbol{X}_n = \boldsymbol{x}_n, Y_n = y_n); \boldsymbol{\Psi}\big) \\ &= \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i, Y_i = y_i; \boldsymbol{\Psi}) \\ &= \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i | Y_i = y_i; \boldsymbol{\Psi}) \mathbb{P}(Y_i = y_i). \end{aligned} \tag{7}$$

Comme Y est binaire, on peut donc écrire

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \big[\mathbb{P}(Y_i = 1)\mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i | Y_i = 1; \boldsymbol{\theta}_1)\big]^{Y_i} \big[\mathbb{P}(Y_i = 0)\mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i | Y_i = 0; \boldsymbol{\theta}_0)\big]^{1-Y_i},$$

où $\mathbb{P}(Y_i = 1) = 1 - \mathbb{P}(Y_i = 0) = \alpha$. Ce qui donne, d'après la loi conditionnelle (6) :

$$L(\boldsymbol{\Psi}) = \prod_{i=1}^{n} \big[\alpha \prod_{j=1}^{p} {\theta_{1j}}^{x_{ij}} (1-\theta_{1j})^{1-x_{ij}}\big]^{Y_i} \big[(1-\alpha) \prod_{j=1}^{p} {\theta_{0j}}^{x_{ij}} (1-\theta_{0j})^{1-x_{ij}}\big]^{1-Y_i}. \tag{8}$$

3. La fonction de log-vraisemblance $\ln L(\boldsymbol{\Psi})$ est donc donnée par

$$\begin{aligned} \ln L(\boldsymbol{\Psi}) &= \sum_{i=1}^{n} \ln \Big\{ \big[\alpha \prod_{j=1}^{p} {\theta_{1j}}^{x_{ij}} (1-\theta_{1j})^{1-x_{ij}}\big]^{Y_i} \big[(1-\alpha) \prod_{j=1}^{p} {\theta_{0j}}^{x_{ij}} (1-\theta_{0j})^{1-x_{ij}}\big]^{1-Y_i} \Big\} \\ &= \sum_{i=1}^{n} \Big\{ Y_i \ln\alpha + Y_i \sum_{j=1}^{d} [x_{ij} \ln\theta_{1j} + (1-x_{ij}) \ln(1-\theta_{1j})] \\ &\quad + (1-Y_i)\ln(1-\alpha) + (1-Y_i) \sum_{j=1}^{d} \big[x_{ij} \ln\theta_{0j} + (1-x_{ij}) \ln(1-\theta_{0j})\big] \Big\}. \end{aligned} \tag{9}$$

4. La dérivée partielle de la log-vraisemblance par rapport à α est donnée par

$$\frac{\partial \ln L(\boldsymbol{\Psi})}{\partial \alpha} = \sum_{i=1}^{n} \left(\frac{y_i}{\alpha} - \frac{1-y_i}{1-\alpha} \right) = \frac{1}{\alpha(1-\alpha)} \sum_{i=1}^{n} (Y_i - \alpha).$$

En l'annulant par rapport à α on trouve l'EMV de α :

$$\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Celle par rapport à θ_{1j} est donnée par

$$\begin{aligned} \frac{\partial \ln L(\boldsymbol{\Psi})}{\partial \theta_{1j}} &= \frac{\partial \Big\{ \sum_{i=1}^{n} y_i \sum_{j=1}^{d} \big[x_{ij} \ln\theta_{1j} + (1-x_{ij}) \ln(1-\theta_{1j})\big] \Big\}}{\partial \theta_{1j}} \\ &= \sum_{i=1}^{n} y_i \left(\frac{x_{ij}}{\theta_{1j}} - \frac{1-x_{ij}}{1-\theta_{1j}} \right) \\ &= \sum_{i=1}^{n} y_i \left(\frac{x_{ij}(1-\theta_{1j})}{\theta_{1j}(1-\theta_{1j})} - \frac{\theta_{1j}(1-x_{ij})}{\theta_{1j}(1-\theta_{1j})} \right) \\ &= \frac{\sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} y_i \theta_{1j}}{\theta_{1j}(1-\theta_{1j})}. \end{aligned}$$

En prenant la valeur θ_{1j} qui annule cette dérivée, on trouve l'EMV de θ_{1j} :

$$\widehat{\theta}_{1j} = \frac{\sum_{i=1}^{n} Y_i X_{ij}}{\sum_{i=1}^{n} Y_i},$$

qui correspond à la moyenne des descripteurs X_{ij} de la classe 1, et l'expression vectorielle correspondante est

$$\widehat{\boldsymbol{\theta}}_1 = \frac{1}{\sum_{i=1}^n Y_i} \sum_{i=1}^n Y_i \boldsymbol{X}_i,$$

ce qui correspond au vecteur moyen de tous les vecteurs $\mathbf{x}_i$ de la classe 1.

De façon similaire, on trouve l'EMV de $\boldsymbol{\theta}_0$:

$$\widehat{\theta}_{0j} = \frac{1}{\sum_{i=1}^n (1-Y_i)} \sum_{i=1}^n (1-Y_i) X_{ij},$$

ce qui correspond au à la moyenne de tous les descripteurs Xx_{ij} de la classe 0 et on a

$$\widehat{\boldsymbol{\theta}}_0 = \frac{1}{\sum_{i=1}^n (1-Y_i)} \sum_{i=1}^n (1-Y_i) \boldsymbol{X}_i$$

qui correspond à la moyenne des individus de la classe 0.

5. La prédiction consiste à affecter l'individu $\boldsymbol{x}_i$ à la classe $\widehat{y}_i$ maximisant la probabilité a posteriori :

$$\begin{aligned}
\widehat{y}_i &= \arg\max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \widehat{\boldsymbol{\Psi}}) \\
&= \arg\max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k) \mathbb{P}(\boldsymbol{X}_i = \boldsymbol{x}_i | Y_i = k; \widehat{\boldsymbol{\Psi}}) \\
&= \arg\max_{k \in [\![0,1]\!]} \mathbb{P}(Y_i = k) \prod_{j=1}^p \widehat{\theta}_{kj}^{x_{ij}} (1-\widehat{\theta}_{kj})^{1-x_{ij}} \qquad (10)
\end{aligned}$$

$\widehat{\boldsymbol{\Psi}}$ étant l'EMV du vecteur paramètre $\boldsymbol{\Psi}$. Afin de pouvoir appliquer cette règle, nous devons calculer tout d'abord $\widehat{\boldsymbol{\Psi}} = (\widehat{\alpha}, \widehat{\boldsymbol{\theta}}_1^\top, \widehat{\boldsymbol{\theta}}_0^\top)^\top$. Selon les données de la Table 1, on trouve :

$$\widehat{\alpha} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{3}{5};$$

$$\widehat{\theta}_{11} = \frac{\sum_{i=1}^n Y_i X_{ij}}{\sum_{i=1}^n Y_i} = \frac{1}{3}; \quad \widehat{\theta}_{12} = \frac{\sum_{i=1}^n Y_i X_{ij}}{\sum_{i=1}^n Y_i} = \frac{2}{3};$$
$$\widehat{\theta}_{01} = \frac{\sum_{i=1}^n (1-Y_i) X_{ij}}{\sum_{i=1}^n (1-Y_i)} = \frac{1}{2}; \quad \widehat{\theta}_{02} = \frac{\sum_{i=1}^n (1-Y_i) X_{ij}}{\sum_{i=1}^n (1-Y_i)} = \frac{1}{2}.$$

Avec $\boldsymbol{x}_i = (1,1)^\top$, pour la classe $Y_i = 1$, on a

$$\widehat{\alpha} \prod_{j=1}^p \widehat{\theta}_{kj}^{x_{ij}} (1-\widehat{\theta}_{kj})^{1-x_{ij}} = \frac{3}{5} \left(\frac{1}{3}\right)^1 \left(1-\frac{1}{3}\right)^0 \left(\frac{2}{3}\right)^1 \left(1-\frac{2}{3}\right)^0 = \frac{2}{15}$$

et pour la classe $Y_i = 0$ on a

$$(1-\widehat{\alpha}) \prod_{j=1}^p \widehat{\theta}_{kj}^{x_{ij}} (1-\widehat{\theta}_{kj})^{1-x_{ij}} = \frac{2}{5} \left(\frac{1}{2}\right)^1 \left(1-\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^1 \left(1-\frac{1}{3}\right)^0 = \frac{1}{10}.$$

On peut donc en déduire que $\widehat{y}_i = 1$.