CHAPITRE 9

Classification bayésienne naïve

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Automne 2022

Au chapitre 3, nous avons vu la régression logistique permettant la classification de la variable réponse en deux catégories. Dans ce chapitre, nous verrons la classification bayésienne naïve, une méthode de classification basée directement sur le théorème de Bayes. Contrairement aux méthodes de régression, les variables explicatives sont considérées aléatoires dans ce modèle. Bien qu'elle soit qualifiée de *simple* par de nombreux auteurs, la classification bayésienne naïve est très performante. Elle constitue d'ailleurs la méthode privilégiée pour filtrer les pourriels des messages électroniques. La classification bayésienne naïve se généralise aisément aux variables réponses à plus de deux catégories.

À la fin du chapitre, vous devriez être en mesure de :

— Comprendre les implications de l'hypothèse d'indépendance conditionnelle.
— Écrire la fonction de vraisemblance du modèle bayésien naïf.
— Calculer les lois prédictives permettant la classification.

Exemple 1

Un filtre anti-pourriel pour les messages électroniques authentiques d'un employé de la compagnie Enron sera développé. Les messages électroniques de 158 employés de la compagnie Enron ont été récupérés par la *Federal Energy Regulatory Commission* pendant la commission d'enquête qui a eu lieu après l'effondrement de la compagnie. Les messages d'un seul employé sont utilisés car le filtre est spécifique à l'utilisateur. Le jeu de données complet est disponible ici et une version prétraitée ici.

9.1 Le modèle marginal

Le but de ce chapitre est de classifier une nouvelle observation en ayant auparavant *appris* d'un échantillon aléatoire. Cette section consiste à présenter le modèle de classification de base en n'utilisant aucune variable explicative.

9.1.1 Modèle statistique

Considérons le problème de la classification des messages électroniques en courriels et pourriels. Posons la variable aléatoire suivante :

$$Y_i = \begin{cases} 0 & \text{si le message } i \text{ est un pourriel}\,; \\ 1 & \text{si le message } i \text{ est un courriel.} \end{cases}$$

Alors on a que $Y_i \sim \mathcal{B}ernoulli(\theta)$ où θ correspond à la probabilité que le message électronique soit un courriel. Le paramètre θ est inconnu et devra être estimé avec un échantillon aléatoire.

Supposons que l'on obtienne un échantillon aléatoire de n messages électroniques et que l'on a identifié lesquels étaient des courriels et des pourriels. On observe donc une réalisation $\boldsymbol{y}$ composée de 0 et de 1 du vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. La vraisemblance du paramètre θ est donnée par l'équation suivante :

$$\begin{aligned} f_{(\boldsymbol{Y}|\theta)}(\boldsymbol{y}) &= \prod_{i=1}^{n} \theta^{y_i} \, (1-\theta)^{1-y_i}; \\ &= \theta^{n_1} \, (1-\theta)^{n_0}; \end{aligned}$$

où $n_1 = \sum_{i=1}^{n} y_i$ correspond au nombre de courriels et $n_0 = n - n_1$ correspond au nombre de pourriels.

9.1.2 Estimation bayésienne

Le paramètre θ peut être estimé en utilisant le théorème de Bayes. L'inférence bayésienne est utilisée notamment pour la calcul de la loi prédictive et pour la mise en jour en temps réel de la connaissance de θ. Une loi *a priori* pour θ est donc nécessaire. Supposons que l'on utilise la loi *a priori* conjuguée suivante pour encoder l'information *a priori* :

$$f_\theta(\theta) = \mathcal{B}eta(\theta \mid \alpha, \beta);$$

alors nous montrerons en classe que la loi *a posteriori* correspondante est la suivante :

$$f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta) = \mathcal{B}eta(\theta \mid \alpha + n_1, \beta + n_0).$$

Exemple 2

Supposons que la loi uniforme sur l'intervalle $(0, 1)$ est utilisée comme loi *a priori* pour la probabilité de courriel θ. Cette loi correspond à la loi $\mathcal{B}eta(1, 1)$. La loi *a posteriori* est donc la loi suivante :

$$f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta) = \mathcal{B}eta(\theta \mid 1 + n_1, 1 + n_0).$$

L'espérance de cette loi *a posteriori* est $\mathbb{E}(\theta|\boldsymbol{Y} = \boldsymbol{y}) = \frac{1+n_1}{2+n}$.

Dans le cas de l'exemple 1 des courriels de l'employé d'Enron, l'échantillon d'entraînement est composé de $n_1 = 2448$ courriels et $n_0 = 1000$ pourriels. Si on suppose que cela correspond à un écahntillon aléatoire, alors

$$f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta) = \mathcal{B}eta(\theta \mid 2449, 1001).$$

Cette densité est illustrée à la figure 9.1. On constate que la proportion de courriels se situe autour de 71%.

Exercice 1

Si la loi non-informative et impropre suivante :

$$f_\theta(\theta) \propto (1 - \theta)^{-1}\, \theta^{-1}$$

est utilisée comme loi *a priori* pour θ, montrez que la loi *a posteriori* correspondante est la suivante :

$$f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) = \mathcal{B}eta(\theta \mid n_1, n_0) \text{ pour } n_0 > 0 \text{ et } n_1 > 0.$$

9.1.3 Loi prédictive

Bien que la loi *a posteriori* de θ soit essentielle pour les calculs bayésiens, ce n'est généralement pas la quantité d'intérêt. La plupart du temps, on cherche plutôt à classifier un nouveau message électronique $\tilde{Y}$. Les probabilités prédictives que le message soit un courriel, $\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y})$, ou un pourriel, $\mathbb{P}(\tilde{Y} = 0 \mid \boldsymbol{Y} = \boldsymbol{y})$, constituent les quantités de prédilection pour effectuer la classification. La probabilité prédictive que le message soit

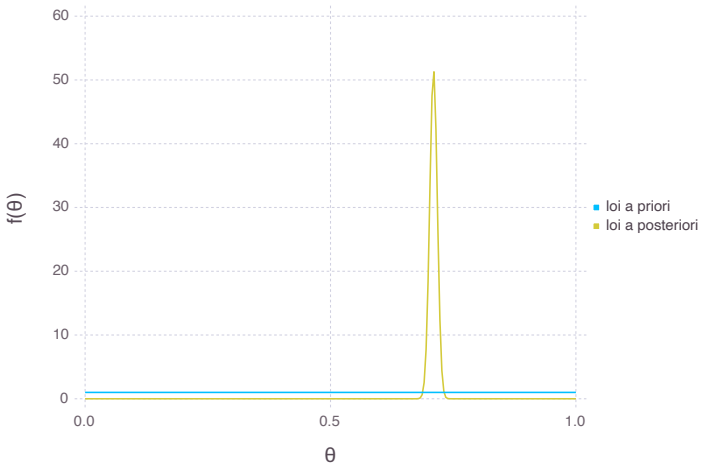

FIGURE 9.1 – Lois *a priori* et *a posteriori* de l'exemple 2.

un courriel se calcule en intégrant l'incertitutde que l'on a sur θ :

$$\begin{aligned}
\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}) &= \int_0^1 f_{(\tilde{Y}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta})}(1) \times f_{(\theta|\boldsymbol{Y}=\boldsymbol{y})}(\theta)\ d\theta \\
&= \int_0^1 \theta \times \mathcal{B}eta(\theta \mid \alpha + n_1,\ \beta + n_0)\ d\theta \\
&= \mathbb{E}(\theta|\boldsymbol{Y} = \boldsymbol{y}) \\
&= \frac{\alpha + n_1}{\alpha + \beta + n}.
\end{aligned}$$

De façon analogue, on peut calculer la probabilité prédictive que le message soit un pourriel :

$$\mathbb{P}(\tilde{Y} = 0 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{\beta + n_0}{\alpha + \beta + n}.$$

Le nouveau message électronique $\tilde{Y}$ sera classé comme courriel si la probabilité $\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y})$ est suffisamment grande.

Exercice 2

Montrez que si on utilise la loi impropre de l'exercice 1, les probabilités prédictives correspondent aux expressions suivantes :

$$\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{n_1}{n},$$
$$\mathbb{P}(\tilde{Y} = 0 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{n_0}{n}.$$

Autrement dit, la probabilité prédictive qu'un nouveau message soit un courriel correspond à la proportion empirique des courriels parmi les messages de l'échantillon d'entraînement. De façon analogue, la probabilité prédictive qu'un nouveau message soit un pourriel correspond à la proportion empirique des pourriels parmi les messages de l'échantillon d'entraînement.

Exemple 3

Dans le cas de l'exemple 2, la probabilité prédictive qu'un nouveau message soit un courriel est égale à :

$$\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}) = \frac{1 + 1000}{2 + 3448} \approx 0.71.$$

La règle de décision correspondante consiste à classer tous les nouveaux messages en courriel puisque $\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}) > \mathbb{P}(\tilde{Y} = 0 \mid \boldsymbol{Y} = \boldsymbol{y})$.

Sur l'échantillon de validation composé de 1224 courriels et de 500 pourriels, on obtient alors les résultats suivants :

		Réalité	
		Pourriel	Courriel
Prédiction	Pourriel	0	0
	Courriel	500	1224

Le score F_1 correspondant et de 0.83.

L'utilisation d'une loi *a priori* impropre peut être problématique par exemple dans le cas où l'on n'observe aucun pourriel dans l'échantillon d'entraînement, *i.e.* $n_0 = 0$. Dans ce cas, la la loi la *a posteriori* de θ est dégénérée. Cela aura pour conséquence d'attribuer à 0 la probabilité que les nouveaux messages entrants soit des pourriels.

Dans le cas où la loi propre mais non informative de l'exemple 1 est utilisée, si on n'observe exemple aucun pourriel dans l'échantillon d'entraînement, alors la probabilité qu'un nouveau message soit un pourriel sera de $\frac{1}{n+2}$. L'utilisation d'une loi *a priori* propre protège en quelque sorte du *sur-apprentissage*.

9.2 Inclusion d'une variable explicative

Une variable explicative est maintenant introduite dans le modèle pour améliorer la classification des messages électroniques. L'information apportée par cette variable est introduite en utilisant le théorème de Bayes, d'où la nomenclature *classification bayésienne.*

Exemple 4

Supposons la variable explicative suivante pour classer les messages électroniques en courriels et pourriels :

$$X_1 = \begin{cases} 1 & \text{si le mot } \textit{http} \text{ se trouve dans le message ;} \\ 0 & \text{sinon.} \end{cases}$$

Dans ce cas, X_1 peut être modélisée par la loi de Bernoulli.

Contrairement aux méthodes de régression, on considère que la variable explicative est une variable aléatoire. On suppose que la variable explicative possède une distribution différente en fonction de la classe de Y. Alors, la loi conditionnelle $f_{(X_1|Y=0,\boldsymbol{\theta}_{01})}(x_1)$ modélise la distribution de paramètres $\boldsymbol{\theta}_{01}$ de la variable explicative pour la classe $Y = 0$ et la loi $f_{(X_1|Y=1,\boldsymbol{\theta}_{11})}(x_1)$ modélise la distribution de paramètres $\boldsymbol{\theta}_{11}$ de la variable explicative pour la classe $Y = 1$.

Remarque. *La plupart du temps, on considère la même famille de distribution pour les différentes classes, il n'y a que les paramètres qui sont différents. D'ailleurs,* $\boldsymbol{\theta}_{yp}$ *dénote le vecteur des paramètres de la distribution de la classe* y *pour la* p^e *variable explicative.*

9.2.1 Modèle statistique

Dans le cas de l'exemple 4, les lois conditionnelles à la classe de la variable explicative X_1 sont définies ainsi :

$$f_{(X_1|Y=0,\theta_{01})}(x_1) = \mathcal{B}ernoulli(x_1 \mid \theta_{01});$$
$$f_{(X_1|Y=1,\theta_{11})}(x_1) = \mathcal{B}ernoulli(x_1 \mid \theta_{11}),$$

où θ_{01} correspond à la probabilité que le mot *http* se retrouve dans les pourriels et θ_{11} correspond à la probabilité que le mot *http* se retrouve dans les courriels. La vraisemblance du modèle, c'est-à-dire la distribution conjointe des observations conditionnellement aux

paramètres, s'obtient à l'aide de la règle de multiplication :

$$f_{\{(\boldsymbol{Y},\boldsymbol{X}_1)|\theta,\theta_{01},\theta_{11}\}}(\boldsymbol{y},\boldsymbol{x}_1) = \prod_{i=1}^{n} f_{(X_{i1}|Y_i=y_i,\theta_{01},\theta_{11})}(x_{i1}) \times f_{(Y_i|\theta)}(y_i)$$
$$= \left\{ \prod_{\{i:y_i=0\}} f_{(X_{i1}|\theta_{01})}(x_{i1}) \right\} \times \left\{ \prod_{\{i:y_i=1\}} f_{(X_{i1}|\theta_{11})}(x_{i1}) \right\} \times \left\{ \prod_{i=1}^{n} f_{(Y_i|\theta)}(y_i) \right\}$$

Dans le cas de l'exemple 2 où la variable explicative est distribuée selon la loi de Bernoulli, on a que

$$f_{\{(\boldsymbol{Y},\boldsymbol{X}_1)|\theta,\theta_{01},\theta_{11}\}}(\boldsymbol{y},\boldsymbol{x}_1)$$
$$= \left\{ \prod_{\{i:y_i=0\}} \theta_{01}^{x_{i1}}(1-\theta_{01})^{1-x_{i1}} \right\} \times \left\{ \prod_{\{i:y_i=1\}} \theta_{11}^{x_{i1}}(1-\theta_{11})^{1-x_{i1}} \right\} \times \left\{ \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \right\}$$

$$= \theta_{01}^{n_{01}}(1-\theta_{01})^{n_0-n_{01}} \times \theta_{11}^{n_1}(1-\theta_{11})^{n_1-n_{11}} \times \theta^{n_1}(1-\theta)^{n_0},$$

où

n_{01} **:** le nombre de pourriels où le mot *http* apparaît ;

n_{11} **:** le nombre de courriels où le mot *http* apparaît.

Exercice 3

Si on considère la loi *a priori* suivante pour les paramètres :

$$f_{(\theta,\theta_{01},\theta_{11})}(\theta,\theta_{01},\theta_{11}) = \mathcal{B}eta(\theta \mid \alpha,\beta) \times \mathcal{B}eta(\theta_{01} \mid \alpha_{01},\beta_{01}) \times \mathcal{B}eta(\theta_{11} \mid \alpha_{11},\beta_{11}),$$

alors montrez que la loi *a posteriori* s'exprime sous la forme suivante :

$$\begin{aligned} f_{\{(\theta,\theta_{01},\theta_{11})|\boldsymbol{Y}=\boldsymbol{y}\}}(\theta,\theta_{01},\theta_{11}) &= \mathcal{B}eta(\theta_{01} \mid \alpha_{01}+n_{01},\ \beta_{01}+n_0-n_{01}) \\ &\times \mathcal{B}eta(\theta_{11} \mid \alpha_{11}+n_{11},\ \beta_{11}+n_1-n_{11}) \\ &\times \mathcal{B}eta(\theta \mid \alpha+n_1,\ \beta+n_0). \end{aligned}$$

Exemple 5

Posons $\alpha_{01} = \beta_{01} = \alpha_{11} = \beta_{11} = 1$. Dans l'ensemble d'entraînement, le mot *http* est présent dans 321 pourriels et 89 courriels. Selon l'exercice 3, on a donc que $n_{01} = 321$ et $n_{11} = 89$. Les lois *a posteriori* marginales correspondantes sont les suivantes :

$$\begin{aligned} f_{(\theta_{01}|\boldsymbol{Y}=\boldsymbol{y})}(\theta_{01}) &= \mathcal{B}eta(\theta_{01} \mid 1+321, 1+1000-321) \\ f_{(\theta_{11}|\boldsymbol{Y}=\boldsymbol{y})}(\theta_{11}) &= \mathcal{B}eta(\theta_{11} \mid 1+89, 1+2448-89). \end{aligned}$$

Remarque. *Si la variable explicative ne s'exprime pas sous la forme d'une loi de Bernoulli, la procédure de décomposition de la variable explicative demeure valide pour le modèle bayésien naïf et ce,même si la variable explicative est continue.*

9.2.2 Loi prédictive

Supposons maintenant que l'on reçoive un nouveau message $\tilde{Y}$ contenant le mot *http*, *i.e.* $\tilde{X}_1 = 1$. On voudrait classifier ce message en courriel ou en pourriel sachant que $\tilde{X}_1 = 1$. On peut alors calculer la probabilité prédictive que le message soit un pourriel :

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = \tilde{x}_1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1). \tag{9.1}$$

Remarque. *Afin de ne pas alourdir la notation de cette section, le conditionnement sur l'ensemble d'entraînement* $(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)$ *est omis pour le reste du chapitre. Il faut garder en tête que nous avons* appris *de l'échantillon d'entraînement, même si on ne l'écrit pas explicitement avec la notation conditionnelle.*

La probabilité prédictive exprimée à l'équation (9.1) peut être calculée en utilisant le théorème de Bayes :

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = 1) = \frac{\mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 0) \times \mathbb{P}(\tilde{Y} = 0)}{\mathbb{P}(\tilde{X}_1 = 1)}.$$

Nous calculerons séparément chacun des trois termes à droite de cette dernière égalité. Calculons d'abord la probabilité prédictive de retrouver le mot *http* dans un pourriel :

$$\begin{aligned}
&\mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 0) \\
&\quad = \int_0^1 \int_0^1 \int_0^1 \mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 0, \boldsymbol{\theta}) \times f_{(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}(\boldsymbol{\theta}) \; d\theta \; d\theta_{01} \; d\theta_{11}; \\
&\quad = \int_0^1 \theta_{01} \times \mathcal{B}eta(\theta_{0_1} \mid \alpha_{01} + n_{01}, \, \beta_{01} + n_0 - n_{01}) \; d\theta_{01}; \\
&\quad = \mathbb{E}(\theta_{01} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1); \\
&\quad = \frac{\alpha_{01} + n_{01}}{\alpha_{01} + \beta_{01} + n_0}.
\end{aligned}$$

Calculons maintenant la probabilité prédictive que le message soit un pourriel :

$$\begin{aligned}
\mathbb{P}(\tilde{Y}=0) &= \int_0^1 \int_0^1 \int_0^1 \mathbb{P}(\tilde{Y}=0 \mid \boldsymbol{\theta}) \times f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{X}_1=\boldsymbol{x}_1)}(\boldsymbol{\theta})\ d\theta\ d\theta_{01}\ d\theta_{11}; \\
&= \int_0^1 (1-\theta) \times \mathcal{B}eta(\theta \mid \alpha+n_1,\, \beta+n_0)\ d\theta; \\
&= 1 - \mathbb{E}(\theta \mid \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_1=\boldsymbol{x}_1); \\
&= 1 - \frac{\alpha+n_1}{\alpha+\beta+n} = \frac{\beta+n_0}{\alpha+\beta+n}.
\end{aligned}$$

Remarque. *Cette dernière probabilité avait déjà été calculée à la section 9.1.3. Elle correspond à la probabilité marginale de recevoir un pourriel peu importe la valeur de la variable explicative.*

Avant de calculer le dénominateur, les termes suivants peuvent être calculés de façon analogue :

$$\begin{aligned}
\mathbb{P}(\tilde{X}_1=1 \mid \tilde{Y}=1) &= \frac{\alpha_{11}+n_{11}}{\alpha_{11}+\beta_{11}+n_1}; \\
\mathbb{P}(\tilde{Y}=1) &= \frac{\alpha+n_1}{\alpha+\beta+n}.
\end{aligned}$$

Le dénominateur, qui correspond à la probabilité de retrouver le mot *http* dans un message, peut finalement être calculé à l'aide de la loi des probabilités totales :

$$\begin{aligned}
\mathbb{P}(\tilde{X}_1=1) =& \mathbb{P}(\tilde{X}_1=1 \mid \tilde{Y}=0) \times \mathbb{P}(\tilde{Y}=0) + \mathbb{P}(\tilde{X}_1=1 \mid \tilde{Y}=1) \times \mathbb{P}(\tilde{Y}=1) \\
=& \frac{\alpha_{01}+n_{01}}{\alpha_{01}+\beta_{01}+n_0} \times \frac{\beta+n_0}{\alpha+\beta+n} + \frac{\alpha_{11}+n_{11}}{\alpha_{11}+\beta_{11}+n_1} \times \frac{\alpha+n_1}{\alpha+\beta+n}.
\end{aligned}$$

Finalement, la probabilité que le nouveau message soit un pourriel est donnée par l'expression suivante :

$$\mathbb{P}(\tilde{Y}=0 \mid \tilde{X}_1=1) = \frac{\frac{\alpha_{01}+n_{01}}{\alpha_{01}+\beta_{01}+n_0} \times \frac{\beta+n_0}{\alpha+\beta+n}}{\frac{\alpha_{01}+n_{01}}{\alpha_{01}+\beta_{01}+n_0} \times \frac{\beta+n_0}{\alpha+\beta+n} + \frac{\alpha_{11}+n_{11}}{\alpha_{11}+\beta_{11}+n_1} \times \frac{\alpha+n_1}{\alpha+\beta+n}}.$$

Le message entrant $\tilde{Y}$ qui contient le mot *http* sera classé comme pourriel si la probabilité prédictive $\mathbb{P}(\tilde{Y}=0 \mid \tilde{X}_1=1)$ est suffisamment grande.

Remarque. *En pratique, on calculera rarement les probabilités exactes suivantes :*

$$\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1);$$

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = 1).$$

Il est en effet plus simple d'obtenir les expressions non-normalisées suivantes :

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = 1) \propto \frac{\alpha_{01} + n_{01}}{\alpha_{01} + \beta_{01} + n_0} \times \frac{\beta + n_0}{\alpha + \beta + n} = p_0; \tag{9.2}$$

$$\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1) \propto \frac{\alpha_{11} + n_{11}}{\alpha_{11} + \beta_{11} + n_1} \times \frac{\alpha + n_1}{\alpha + \beta + n} = p_1. \tag{9.3}$$

Le message $\tilde{Y}$ sera classifié comme courriel si p_1 est plus grand que p_0.

Exercice 4

Montrez que la probabilité prédictive qu'un message soit un courriel sachant qu'il ne contient pas le mot *http* est égal à l'expression suivante :

$$\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 0) = \frac{\frac{\beta_{11}+n_1-n_{11}}{\alpha_{11}+\beta_{11}+n_1} \times \frac{\alpha+n_1}{\alpha+\beta+n}}{\frac{\beta_{01}+n_0+n_{01}}{\alpha_{01}+\beta_{01}+n_0} \times \frac{\beta+n_0}{\alpha+\beta+n} + \frac{\beta_{11}+n_1-n_{11}}{\alpha_{11}+\beta_{11}+n_1} \times \frac{\alpha+n_1}{\alpha+\beta+n}}.$$

Par conséquent, dans un filtre anti-pourriel, un message qui ne contient pas le mot *http* sera classé comme courriel si cette probabilité est suffisamment grande.

Exemple 6

Selon l'exemple 5, la probabilité que le message soit un courriel sachant qu'il contient le mot *http* est égale à

$$\mathbb{P}(\tilde{Y} = 1 | \tilde{X}_1 = 1) \approx 0.22.$$

La probabilité que le message soit un courriel sachant qu'il ne contient contient pas le mot *http* est égale à

$$\mathbb{P}(\tilde{Y} = 1 | \tilde{X}_1 = 0) \approx 0.78.$$

La règle de décision consiste donc à classer comme pourriel tout message contenant le mot *http*. Avec cette règle, on obtient les résultats suivants :

		Réalité	
		Pourriel	Courriel
Prédiction	Pourriel	154	46
	Courriel	346	1178

Le score F_1 correspondant et de 0.86.

9.2.3 Effet de la variable explicative

Lorsqu'il n'y avait aucune variable explicative considérée, la probabilité prédictive que le message entrant soit un pourriel était de $\frac{\alpha+n_0}{\alpha+\beta+n}$ (voir la section 9.1.3). Dans l'expression de la probabilité prédictive exprimée à l'équation (9.2), on remarque que le terme $\frac{\alpha_{11}+n_{01}}{\alpha_{01}+\beta_{01}+n_0}$ corrige la probabilité marginale que le message soit un pourriel lorsque le mot *http* est présent dans le message. On pourrait appeler ce terme le *spamliness* du mot *http*.

9.3 Ajout d'une deuxième variable explicative

Considérons maintenant le cas où une deuxième variable explicative X_2 est considérée en plus de X_1.

Exemple 7

Soit la variable explicative suivante pour classifier les messages électroniques :

$$X_2 = \begin{cases} 1 & \text{si le mot } \textit{enron} \text{ se trouve dans le message ;} \\ 0 & \text{sinon.} \end{cases}$$

La variable X_2 peut donc être modélisée par la loi de Bernoulli. L'ajout d'une autre variable explicative a pour but d'améliorer la qualité du filtrage des messages.

À l'instar de la section précédente, les lois conditionnelles de la variable X_2 sont requises pour modéliser la variable en fonction de la classe de la variable d'intérêt. Dans le cas de l'exemple 3, les lois conditionnelles sont des lois de Bernoulli :

$$\begin{aligned} f_{(X_2|Y=0,\theta_{02})}(x_2) &= \mathcal{B}ernoulli(x_2 \mid \theta_{02}); \\ f_{(X_2|Y=1,\theta_{12})}(x_2) &= \mathcal{B}ernoulli(x_2 \mid \theta_{12}), \end{aligned}$$

où θ_{02} correspond à la probabilité que le mot *enron* se retrouve dans les pourriels et θ_{12} correspond à la probabilité que le mot *enron* se retrouve dans les courriels.

9.3.1 Modèle statistique

Avec la variable X_2, le nombre de paramètres du modèle est égal à 5, *i.e*

$$\boldsymbol{\theta} = (\theta, \theta_{01}, \theta_{11}, \theta_{02}, \theta_{12}).$$

La vraisemblance des paramètres $\boldsymbol{\theta}$ pour le message électronique i est donnée par la règle de multiplication :

$$f_{\{(X_{i1},X_{i2},Y_i)|\boldsymbol{\theta}\}}(x_{i1}, x_{i2}, y_i) = f_{\{(X_{i1},X_{i2})|Y_i=y_i,\boldsymbol{\theta}\}}(x_{i1}, x_{i2}) \times f_{(Y_i|\boldsymbol{\theta})}(y_i).$$

Or, nous ne possédons pas la loi conjointe du couple (X_1, X_2) sachant $Y = y$. Une simplification naïve consiste à supposer que les variables explicatives sont **conditionnellement indépendantes** :

$$f_{\{(X_{i1},X_{i2})|Y_i=y_i,\boldsymbol{\theta}\}}(x_{i1}, x_{i2}) = f_{(X_{i1}|Y_i=y_i,\boldsymbol{\theta})}(x_{i1}) \times f_{(X_{i2}|Y_i=y_i,\boldsymbol{\theta})}(x_{i2}).$$

Remarque. *L'hypothèse d'indépendance des variables X_1 et X_2 serait trop forte. Il serait en effet déraisonnable de supposer que l'occurrence des mots* http *et* enron *soit indépendante. Si l'un des mots est présent, on s'attend à ce que la probabilité de trouver l'autre soit plus faible. Cependant, sachant que le message est un courriel, on peut supposer que les deux variables sont indépendantes. La probabilité qu'un courriel contienne le mot* http *et le mot* enron *correspond donc à la probabilité qu'un courriel contienne le mot* http *multiplié par la probabilité qu'un courriel contienne le mot* enron.

Le mot *naïve* de l'expression *classification naïve bayésienne* provient du fait que l'on simplifie le modèle à l'aide de l'hypothèse d'indépendance conditionnelle. Cette simplification est rarement vraie en pratique. Néanmoins, le modèle probabiliste qui en découle performe généralement très bien même si cette hypothèse n'est pas satisfaite.

Avec l'hypothèse d'indépendance conditionnelle, la vraisemblance des paramètres $\boldsymbol{\theta}$ s'écrit de la façon suivante :

$$\begin{aligned}
&f_{\{(\boldsymbol{Y},\boldsymbol{X}_1)|\theta,\theta_{01},\theta_{11}\}}(\boldsymbol{y}, \boldsymbol{x}_1) \\
&\qquad = \prod_{i=1}^{n} f_{(X_{i1}|Y_i=y_i,\theta_{01},\theta_{11})}(x_{i1}) \times f_{(X_{i2}|Y_i=y_i,\theta_{02},\theta_{12})}(x_{i2}) \times f_{(Y_i|\theta)}(y_i) \\
&\qquad = \theta_{01}^{n_{01}}(1-\theta_{01})^{n_0-n_{01}}\ \theta_{11}^{n_{11}}(1-\theta_{11})^{n_1-n_{11}} \\
&\qquad\quad \times \theta_{02}^{n_{02}}(1-\theta_{02})^{n_0-n_{02}}\ \theta_{12}^{n_{12}}(1-\theta_{12})^{n_1-n_{12}} \\
&\qquad\quad \times \theta^{n_1}(1-\theta)^{n_0},
\end{aligned}$$

où

n_{02} **:** le nombre de pourriels où le mot *enron* apparaît ;

n_{12} **:** le nombre de courriels où le mot *enron* apparaît.

Grâce à l'hypothèse d'indépendance conditionnelle, chaque variable explicative peut être traitée indépendamment. En effet, la vraisemblance exprimée à l'équation précédente se factorise en fonction des différents paramètres.

Exercice 5

Si la loi *a priori* suivante est utilisée :

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = &\mathcal{B}eta(\theta_{01} \mid \alpha_{01}, \beta_{01}) \times \mathcal{B}eta(\theta_{11} \mid \alpha_{11}, \beta_{11}) \\ &\times \mathcal{B}eta(\theta_{02} \mid \alpha_{02}, \beta_{02}) \times \mathcal{B}eta(\theta_{12} \mid \alpha_{12}, \beta_{12}) \\ &\times \mathcal{B}eta(\theta \mid \alpha, \beta), \end{aligned}$$

montrez que la forme fonctionnelle de la loi *a posteriori* des paramètres s'exprime sous la forme suivante :

$$\begin{aligned} &f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) \\ &\propto \mathcal{B}eta(\theta_{01} \mid \alpha_{01} + n_{01}, \beta_{01} + n_0 - n_{01}) \times \mathcal{B}eta(\theta_{11} \mid \alpha_{11} + n_{11}, \beta_{11} + n_1 - n_{11}) \\ &\propto \mathcal{B}eta(\theta_{01} \mid \alpha_{02} + n_{02}, \beta_{02} + n_0 - n_{02}) \times \mathcal{B}eta(\theta_{12} \mid \alpha_{12} + n_{12}, \beta_{12} + n_1 - n_{12}) \\ &\times \mathcal{B}eta(\theta \mid \alpha + n_1, \beta + n_0). \end{aligned}$$

Exemple 8

Posons $\alpha_{02} = \beta_{02} = \alpha_{12} = \beta_{12} = 1$. Dans l'ensemble d'entraînement, le mot *enron* est présent dans 0 pourriel et 986 courriels. Selon l'exercice 5, on a donc que $n_{02} = 0$ et $n_{12} = 986$. Les lois *a posteriori* marginales correspondantes sont les suivantes :

$$\begin{aligned} f_{(\theta_{02}|\boldsymbol{Y}=\boldsymbol{y})}(\theta_{02}) &= \mathcal{B}eta(\theta_{02} \mid 1 + 0, 1 + 1000 - 0) \\ f_{(\theta_{12}|\boldsymbol{Y}=\boldsymbol{y})}(\theta_{02}) &= \mathcal{B}eta(\theta_{02} \mid 1 + 986, 1 + 2448 - 986). \end{aligned}$$

9.3.2 Loi prédictive

On reçoit un nouveau message qui contient les mots *http* et *enron*, *i.e.* $\tilde{X}_1 = 1$ et $\tilde{X}_2 = 1$, calculons la probabilité prédictive que ce message soit un courriel :

$$\begin{aligned} &\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1, \tilde{X}_2 = 1) \\ &\quad \propto \mathbb{P}(\tilde{X}_1 = 1 \cap \tilde{X}_2 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{Y} = 1) \text{ (théorème de Bayes)} \\ &\quad \propto \mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{X}_2 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{Y} = 1) \text{ (ind. cond.)} \\ &\quad \propto \frac{\alpha_{11} + n_{11}}{\alpha_{11} + \beta_{11} + n_1} \times \frac{\alpha_{12} + n_{12}}{\alpha_{12} + \beta_{12} + n_1} \times \frac{\alpha + n_1}{\alpha + \beta + n} = p_1 \text{ (calcul par cond.).} \end{aligned}$$

De façon analogue, la probabilité que le message contenant les mots *http* et *enron* soit un pourriel est proportionnelle à

$$\begin{aligned}&\mathbb{P}(\tilde{Y}=0\,|\tilde{X}_1=1,\tilde{X}_2=1)\\&\qquad\propto \frac{\alpha_{01}+n_{01}}{\alpha_{01}+\beta_{01}+n_1}\times\frac{\alpha_{02}+n_{02}}{\alpha_{02}+\beta_{02}+n_1}\times\frac{\alpha+n_1}{\alpha+\beta+n}=p_0.\end{aligned}\tag{9.4}$$

Si les mots *http* et *enron* sont présents dans le message, la probabilité que le message soit un pourriel p_0 est proportionnelle à la probabilité marginale que le message soit un pourriel, corrigée par le *spamliness* de X_1 et le *spamliness* de X_2. Cette factorisation de la probabilité prédictive est une conséquence de la simplification supposant l'indépendance conditionnelle des variables X_1 et X_2 sachant Y.

Exercice 6

Soit le nouveau message $\tilde{Y}$ ne contenant pas les mots *http* et *enron*, *i.e.* $\tilde{X}_1=0$ et $\tilde{X}_2=0$, montrez que les probabilités que ce message soit respectivement un courriel et un pourriel sont proportionnelles aux expressions suivantes :

$$\begin{aligned}&\mathbb{P}(\tilde{Y}=1\,|\tilde{X}_1=0,\tilde{X}_2=0)\\&\qquad\propto \mathbb{P}(\tilde{X}_1=0\mid\tilde{Y}=1)\times\mathbb{P}(\tilde{X}_2=0\mid\tilde{Y}=1)\times\mathbb{P}(\tilde{Y}=1)\\&\qquad\propto \frac{\beta_{11}+n_1-n_{11}}{\alpha_{11}+\beta_{11}+n_1}\times\frac{\beta_{12}+n_1-n_{12}}{\alpha_{12}+\beta_{12}+n_1}\times\frac{\alpha+n_1}{\alpha+\beta+n}.\end{aligned}$$

$$\begin{aligned}&\mathbb{P}(\tilde{Y}=0\,|\tilde{X}_1=0,\tilde{X}_2=0)\\&\qquad\propto \mathbb{P}(\tilde{X}_1=0\mid\tilde{Y}=0)\times\mathbb{P}(\tilde{X}_2=0\mid\tilde{Y}=0)\times\mathbb{P}(\tilde{Y}=0)\\&\qquad\propto \frac{\beta_{01}+n_0-n_{01}}{\alpha_{01}+\beta_{01}+n_0}\times\frac{\beta_{02}+n_0-n_{02}}{\alpha_{02}+\beta_{02}+n_0}\times\frac{\beta+n_0}{\alpha+\beta+n}.\end{aligned}$$

Exemple 9

Selon l'exemple 8, la probabilité que le message soit un courriel sachant qu'il contient les mots *http* et *enron* est égale à

$$\mathbb{P}(\tilde{Y}=1|\tilde{X}_1=1,\tilde{X}_2=1)\approx 0.99.$$

La probabilité que le message soit un courriel sachant qu'il contient le mot *http* et ne contient pas le mot *enron* est égale à

$$\mathbb{P}(\tilde{Y}=1|\tilde{X}_1=1,\tilde{X}_2=0)\approx 0.14.$$

La probabilité que le message soit un courriel sachant qu'il ne contient pas le mot *http* et qu'il contient le mot *enron* est égale à

$$\mathbb{P}(\tilde{Y} = 1 | \tilde{X}_1 = 0, \tilde{X}_2 = 1) \approx 1.$$

La probabilité que le message soit un courriel sachant qu'il ne contient contient pas les mots *http* et *enron* est égale à

$$\mathbb{P}(\tilde{Y} = 1 | \tilde{X}_1 = 0, \tilde{X}_2 = 0) \approx 0.67.$$

La règle de décision consiste donc à classer comme pourriel tout message contenant le mot *http* et ne contenant pas le mot *enron*. Avec cette règle, on obtient les résultats suivants :

		Réalité	
		Pourriel	Courriel
Prédiction	Pourriel	154	26
	Courriel	346	1198

Le score F_1 correspondant et de 0.87.

9.4 Inclusion de plusieurs variables explicatives

Soit le modèle bayésien naïf avec p variables explicatives. Dénotons le vecteur des variables explicatives par $\boldsymbol{X} = (X_1, \ldots, X_p)$. En utilisant l'hypothèse d'indépendance conditionnelle, la vraisemblance du modèle pour une observation $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ s'écrit de la façon suivante :

$$f_{\{(\boldsymbol{X}_i, Y_i)|\boldsymbol{\theta}\}}(\boldsymbol{x}_i, y_i) = \left\{ \prod_{j=1}^{p} f_{(X_j|Y_i=y_i, \boldsymbol{\theta})}(x_{ij}) \right\} \times f_{(Y_i|\boldsymbol{\theta})}(y_i).$$

Pour les n observations de l'échantillon aléatoire, la vraisemblance s'écrit de la façon suivante :

$$f_{\{(X, \boldsymbol{Y})|\boldsymbol{\theta}\}}(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{n} f_{\{(\boldsymbol{X}_i, Y_i)|\boldsymbol{\theta}\}}(\boldsymbol{x}_i, y_i).$$

Si l'on suppose que toutes les variables explicatives s'expriment sous la forme d'une loi

de Bernoulli, alors on obtient la forme suivante pour la fonction de vraisemblance :

$$\begin{aligned} f_{\{(X,\boldsymbol{Y})|\boldsymbol{\theta}\}}(\boldsymbol{x},\boldsymbol{y}) = {} & (1-\theta_{01})^{n_0-n_{01}}\ \theta_{01}^{n_{01}} \times (1-\theta_{11})^{n_1-n_{11}}\ \theta_{11}^{n_{11}} \\ & \times (1-\theta_{02})^{n_0-n_{02}}\ \theta_{02}^{n_{02}} \times (1-\theta_{12})^{n_1-n_{12}}\ \theta_{12}^{n_{12}} \\ & \vdots \\ & \times (1-\theta_{0p})^{n_0-n_{0p}}\ \theta_{0p}^{n_{0p}} \times (1-\theta_{1p})^{n_1-n_{1p}}\ \theta_{1p}^{n_{1p}} \\ & \times (1-\theta)^{n_0}\ \theta^{n_1}; \end{aligned}$$

où

n_{0j} **:** le nombre de pourriels où la variable X_j est un succès ;

n_{1j} **:** le nombre de courriels où la variable X_j est un succès.

Le calcul de la loi *a posteriori* des paramètres et de la loi prédictive se font de façon similaire au cas de la section précédente avec deux variables explicatives. On obtient que la probabilité prédictive qu'un nouveau message ayant $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}$ soit un courriel est proportionnelle à l'expression suivante :

$$\mathbb{P}(\tilde{Y}=1|\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{x}}) \propto \left\{\prod_{j=1}^{p} \mathbb{P}(\tilde{X}_j=\tilde{x}_j|\tilde{Y}=1)\right\} \times \mathbb{P}(\tilde{Y}=1)$$

De façon analogue, on trouve que la probabilité prédictive que le message soit un pourriel est proportionnelle à l'expression suivante :

$$\mathbb{P}(\tilde{Y}=0|\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{x}}) \propto \left\{\prod_{j=1}^{p} \mathbb{P}(\tilde{X}_j=\tilde{x}_j|\tilde{Y}=0)\right\} \times \mathbb{P}(\tilde{Y}=0)$$

Le terme $\mathbb{P}(\tilde{X}_j=\tilde{x}_j|\tilde{Y}=0)$ correspond au *spamliness* de la variable X_j.

Remarque. *Dans le cas d'un filtre anti-pourriel, une fonction calculant le facteur* $\mathbb{P}(\tilde{X}_j = \tilde{x}_j|Y=0,\boldsymbol{Y}=\boldsymbol{y},X=x)$ *peut être implémentée et utilisée pour chaque* $j \in \{1,\ldots,p\}$*. Une telle fonction permettrait de calculer le* spamliness *des variables explicatives.*

9.5 Le cas où la variable d'intérêt possède plus de deux catégories (OPTIONNEL)

Soit la variable d'intérêt Y pouvant prendre des valeurs dans l'ensemble $\{1,2,\ldots,m\}$. La variable aléatoire Y sert à identifier la classe d'un objet en question. La classification de

texte par sujet consitue un exemple d'une telle variable à plusieurs catégories :

$$Y = \begin{cases} 1 & \text{si les mathématiques sont le sujet du texte ;} \\ 2 & \text{si le génie informatique est le sujet du texte ;} \\ 3 & \text{sinon.} \end{cases}$$

La variable Y prenant des valeurs dans l'ensemble $\{1, 2, \ldots, m\}$ peut être modélisée par la loi catégorielle. Supposons le vecteur de probabilités $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)$ où $\alpha_k = \mathbb{P}(Y = k)$ et où $\sum_{k=1}^{m} \alpha_k = 1$. Alors la fonction de masse de la loi catégorielle Y est la suivante :

$$p_{(Y|\boldsymbol{\alpha})}(k) = \begin{cases} \alpha_k & \text{pour } k \in \{1, 2, \ldots, m\}, \\ 0 & \text{sinon.} \end{cases}$$

On dit alors que la variable aléatoire Y est distribuée selon la loi catégorielle avec le vecteur de probabilité $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. On peut dénoter cette expression par $Y \sim \mathcal{C}at(\boldsymbol{\alpha})$.

Remarque. *La distribution catégorielle constitue un généralisation de la loi de Bernoulli pour plusieurs catégories. Pour simplifier l'écriture, on dénote par $\{0, 1\}$ l'ensemble des valeurs possibles de Y lorsqu'il n'y a que deux catégories possibles. Lorsqu'il y a plus que deux catégories, on dénote l'ensemble des valeurs possibles de Y par $\{1, 2, \ldots, m\}$.*

Dans le cas où Y possède m catégories, il faut définir les m lois conditionnelles suivantes pour chacune des variables explicatives X_j :

$$\begin{cases} f_{(X_j|Y=1,\boldsymbol{\theta})}(x_j) \\ f_{(X_j|Y=2,\boldsymbol{\theta})}(x_j) \\ \quad \vdots \\ f_{(X_j|Y=m,\boldsymbol{\theta})}(x_j) \end{cases}$$

9.6 Exercices

1. Pour la classification des messages électroniques, supposons que seulement la variable explicative $X_1 =$ *le nombre de mots en majuscules dans le message électronique* soit considérée. Supposons également que $f_{(X_1|\theta_1)}(x_1) = \mathcal{P}oisson(x_1 \mid \theta_1)$.

   (a) Proposez des lois conditionnelles appropriées pour la classification bayésienne naïve.

   (b) Écrivez la vraisemblance des paramètres pour le message i de l'échantillon aléatoire.

(c) Écrivez la vraisemblance des paramètres pour les n messages de l'échantillon aléatoire.

(d) Calculez la loi *a posteriori* des paramètres correspondante à la loi *a priori* suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathcal{B}eta(\theta \mid 1,1) \times \mathcal{G}amma(\theta_{01} \mid 1,1) \times \mathcal{G}amma(\theta_{11} \mid 1,1).$$

(e) Calculez la probabilité prédictive qu'un nouveau message contenant 0 mots en majuscules soit un courriel.