CHAPITRE 8

Classification bayésienne naïve

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

Jonathan Jalbert – Automne 2019

Au chapitre 3, nous avons vu la régression logistique permettant la classification de la variable réponse en deux catégories. Dans ce chapitre, nous verrons une méthode permettant la classification de la variable réponse en m catégories basée sur le théorème de Bayes : la classification bayésienne naïve. Cette méthode est assez simple et c'est pourquoi elle est très répandue en pratique. C'est d'ailleurs la méthode privilégiée pour filtrer les messages électroniques en courriels et pourriels.

À la fin du chapitre, les étudiants devraient 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.

La théorie de ce chapitre est illustrée pour la construction d'un filtre anti-pourriel où la variable réponse ne possède que deux catégories. Le chapitre culminera d'ailleurs avec la construction d'un filtre anti-pourriel réel lors du TD8. Il faut cependant noter que la classification bayésienne naïve s'applique également à des problèmes de classification à plus de deux catégories. La dernière section généralisera le modèle pour une variable d'intérêt possédant plus que deux catégories.

8.1 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.

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 :

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

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.

Le paramètre θ peut être estimée 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* 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 :

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

Exemple 1

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 :

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

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

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 :

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

L'utilisation d'une loi *a priori* impropre est problématique si $n_0 = 0$ ou si $n_1 = 0$. Par ailleurs, lorsque la loi *a priori* propre de l'exemple 2 est utilisée, le fait que n_0 ou n_1 soit nul n'occasionne aucun problème. L'utilisation d'une loi *a priori* informative protège en quelque sorte du *sur-apprentissage*.

8.1.1 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 classer un nouveau message électronique entrant. La probabilité prédictive que le message soit un courriel constitue donc la quantité de prédilection pour effectuer ce classement :

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

Par conséquent, la loi prédictive de $\tilde{Y}$ est la suivante :

$$f_{(\tilde{Y}|\boldsymbol{Y}=\boldsymbol{y})}(y) = \mathcal{B}ernoulli\left(y \,\middle|\, \frac{n_1 + \alpha}{n + \alpha + \beta}\right).$$

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

Exercice 2

Quelle est la loi prédictive de $\tilde{Y}$ si la loi *a priori* impropre $f_\theta(\theta) \propto (1-\theta)^{-1}\,\theta^{-1}$ est utilisée ?

8.2 Inclusion d'une variable explicative

Une variable explicative est maintenant introduite dans le modèle pour améliorer la classification des messages électroniques.

Exemple 2

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

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

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. La loi marginale de X_1 peut être définie de la façon suivante :

$$f_{(X_1|\theta_1)}(x_1) = \mathcal{B}ernoulli(x_1 \mid \theta_1);$$

où θ_1 correspond à la probabilité qu'un message électronique contienne le mot *enlarge*. Dans la classification bayésienne naïve, ce sont plutôt les lois conditionnelles suivantes qui sont utiles :

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

où θ_{01} correspond à la probabilité que le mot *enlarge* se retrouve dans les pourriels et θ_{11} correspond à la probabilité que le mot *enlarge* se retrouve dans les courriels.

L'intégration de la variable explicative X_1 s'effectue par le théorème de Bayes, d'où la nomenclature *classification bayésienne*. On souhaite mettre à jour la connaissance sur le vecteur de paramètres $\boldsymbol{\theta} = (\theta, \theta_{01}, \theta_{11})$ après avoir observé l'échantillon aléatoire. La vraisemblance de $\boldsymbol{\theta}$ pour le message électronique i est donnée par la règle de multiplication :

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

La loi conditionnelle de $(X_{i1}|Y_i = y_i, \boldsymbol{\theta})$ peut ensuite être décomposée en fonction des valeurs de Y_i :

$$\begin{aligned} f_{\{(X_{i1},Y_i)|\boldsymbol{\theta}\}}(x_{i1}, y_i) &= \left\{f_{(X_{i1}|Y_i=0,\boldsymbol{\theta})}(x_{i1})\right\}^{1-y_i} \times \left\{f_{(X_{i1}|Y_i=1,\boldsymbol{\theta})}(x_{i1})\right\}^{y_i} \times f_{(Y_i|\boldsymbol{\theta})}(y_i) \\ &= \left\{\theta_{01}^{x_{i1}}(1-\theta_{01})^{1-x_{i1}}\right\}^{1-y_i} \times \left\{\theta_{11}^{x_{i1}}(1-\theta_{11})^{1-x_{i1}}\right\}^{y_i} \times \theta^{y_i}(1-\theta)^{1-y_i}. \end{aligned}$$

Pour l'ensemble des n messages électroniques, la vraisemblance des paramètres $\boldsymbol{\theta}$ s'exprime sous la forme suivante :

$$\begin{aligned} f_{\{(\boldsymbol{X}_1,\boldsymbol{Y})|\boldsymbol{\theta}\}}(\boldsymbol{x}_1, \boldsymbol{y}) &= \prod_{i=1}^{n} f_{\{(X_{i1},Y_i)|\boldsymbol{\theta}\}}(x_{i1}, y_i) \\ &= \left\{\prod_{\{i:y_i=0\}} \theta_{01}^{x_{i1}}(1-\theta_{01})^{1-x_{i1}}\right\} \times \left\{\prod_{\{i:y_i=1\}} \theta_{11}^{x_{i1}}(1-\theta_{11})^{1-x_{i1}}\right\} \times \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta_{01}^{n_{01}}\ (1-\theta_{01})^{n_0-n_{01}} \times \theta_{11}^{n_{11}}\ (1-\theta_{11})^{n_1-n_{11}} \times \theta^{n_1}\ (1-\theta)^{n_0}; \end{aligned}$$

où

n_0 **:** Le nombre de pourriels.

n_1 **:** Le nombre de courriels.

n_{01} **:** Le nombre de pourriels où le mot *enlarge* apparaît.

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

Ce modèle contient 3 paramètres, $\boldsymbol{\theta} = (\theta, \theta_{01}, \theta_{11})$, qui devront être estimés.

Exercice 3

Si la loi *a priori* suivante est utilisée :

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

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

$$\begin{aligned} f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) = \mathcal{B}eta(\theta_{01} \mid n_{01}+1, n_0-n_{01}+1) \times \mathcal{B}eta(\theta_{11} \mid n_{11}+1, n_1-n_{11}+1) \\ \times \mathcal{B}eta(\theta \mid n_1+1, n_0+1). \end{aligned}$$

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.*

8.2.1 Loi prédictive

Supposons maintenant qu'un nouveau message contenant le mot *enlarge*, *i.e.* $\tilde{X}_1 = 1$, entre dans la boîte courriel. Pour classer ce message en courriel ou pourriel, la probabilité prédictive que le message soit un courriel est calculée :

$$\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = \tilde{x}_1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1).$$

Cette probabilité prédictive peut être calculée en utilisant le théorème de Bayes :

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \\
&\quad = \frac{\mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \times \mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}{\mathbb{P}(\tilde{X}_1 = 1 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \\
&\quad = \int_0^1 \int_0^1 \int_0^1 \mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 1, \boldsymbol{\theta}) \times f_{(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}(\boldsymbol{\theta}) d\theta \; d\theta_{01} \; d\theta_{11}; \\
&\quad = \int_0^1 \int_0^1 \int_0^1 \theta_{11} \times f_{(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}(\boldsymbol{\theta}) d\theta \; d\theta_{01} \; d\theta_{11}; \\
&\quad = \mathbb{E}(\theta_{11} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1); \\
&\quad = \frac{n_{11} + 1}{n_1 + 2}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \\
&\quad = \int_0^1 \int_0^1 \int_0^1 \mathbb{P}(\tilde{Y} = 1 \mid \boldsymbol{\theta}) \times f_{(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}(\boldsymbol{\theta}) d\theta \; d\theta_{01} \; d\theta_{11}; \\
&\quad = \int_0^1 \int_0^1 \int_0^1 \theta \times f_{(\boldsymbol{\theta} \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)}(\boldsymbol{\theta}) d\theta \; d\theta_{01} \; d\theta_{11}; \\
&\quad = \mathbb{E}(\theta \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1); \\
&\quad = \frac{n_1 + 1}{n + 2}.
\end{aligned}$$

Avant de calculer le dénominateur, on peut calculer de façon analogue les termes suivants :

$$\mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 0, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) = \frac{n_{01} + 1}{n_1 + 2};$$

$$\mathbb{P}(\tilde{Y} = 0 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) = \frac{n_0 + 1}{n + 2}.$$

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

$$\mathbb{P}(\tilde{X}_1 = 1 \mid \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) = \frac{\frac{n_{11}+1}{n_1+2} \times \frac{n_1+1}{n+2}}{\frac{n_{01}+1}{n_0+2} \times \frac{n_0+1}{n+2} + \frac{n_{11}+1}{n_1+2} \times \frac{n_1+1}{n+2}}.$$

Dans un filtre anti-pourriel, un message qui contient le mot *enlarge* sera classé comme courriel si la probabilité prédictive $\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1)$ est suffisamment grande.

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

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

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

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

$$\mathbb{P}(\tilde{Y} = 1 \mid \tilde{X}_1 = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \propto \frac{n_{11} + 1}{n_1 + 2} \times \frac{n_1 + 1}{n + 2} = p_1;$$

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = 1, \boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{X}_1 = \boldsymbol{x}_1) \propto \frac{n_{01} + 1}{n_0 + 2} \times \frac{n_0 + 1}{n + 2} = p_0.$$

On classera le message $\tilde{Y}$ comme courriel si p_1 est suffisamment plus grand que p_0.

Dans l'expression de la probabilité prédictive $\mathbb{P}(\tilde{Y} = 0 | \tilde{X}_1 = 1, \boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{Y} = \boldsymbol{y})$, le terme $\frac{n_{01}+1}{n_0+2}$ corrige en fonction de la valeur de la variable explicative la probabilité marginale que le message soit un pourriel $\frac{n_0+1}{n+2}$. On pourrait appeler ce terme le *spamliness* de l'événement $\tilde{X}_1 = 1$.

Exercice 4

Montrez que la probabilité prédictive qu'un message soit un courriel sachant qu'il ne

contient pas le mot *enlarge* est égal à l'expression suivante :

$$\mathbb{P}(\tilde{Y}=1 \mid \tilde{X}_0=1, \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_1=\boldsymbol{x}_1) = \frac{\frac{n_1-n_{11}+1}{n_1+2} \times \frac{n_1+1}{n+2}}{\frac{n_0-n_{01}+1}{n_0+2} \times \frac{n_0+1}{n+2} + \frac{n_1-n_{11}+1}{n_1+2} \times \frac{n_1+1}{n+2}}.$$

Par conséquent, dans un filtre anti-pourriel, un message qui ne contient pas le mot *enlarge* sera classé comme courriel si $\mathbb{P}(\tilde{Y}=1 \mid \tilde{X}_1=0, \boldsymbol{Y}=\boldsymbol{y}, \boldsymbol{X}_1=\boldsymbol{x}_1)$ est suffisamment grande.

8.3 Ajout d'une deuxième variable explicative

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

Exemple 3

Par exemple, la variable explicative suivante pourrait être ajoutée :

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

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.

La loi marginale de X_2 peut être définie de la façon suivante :

$$f_{(X_2|\theta_2)}(x_2) = \mathcal{B}ernoulli(x_2 \mid \theta_2);$$

où θ_2 correspond à la probabilité qu'un message électronique contienne le mot *dear*. Tel que mentionné à la section précédente, ce sont plutôt les lois conditionnelles suivantes qui sont importantes pour la classification bayésienne naïve :

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

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

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* enlarge *et* dear *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 pourriel, on peut supposer que les deux variables sont indépendantes. La probabilité qu'un pourriel contienne le mot* enlarge *et le mot* dear *correspond donc à la probabilité qu'un pourriel contienne le mot* enlarge *multiplié par la probabilité qu'un pourriel contienne le mot* dear*.*

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. Il faut noter que cette simplification est rarement vraie en pratique. Néanmoins, le modèle probabiliste en découlant performe généralement très bien même si cette hypothèse n'est pas satisfaite.

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

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

Avec les distributions conditionnelles données précédemment pour les varaibles X_1 et X_2, on a que

$$f_{(X_{i1}|Y_i=y_i,\boldsymbol{\theta})}(x_{i1}) = \left\{\theta_{01}^{x_{i1}}(1-\theta_{01})^{1-x_{i1}}\right\}^{1-y_i} \times \left\{\theta_{11}^{x_{i1}}(1-\theta_{11})^{1-x_{i1}}\right\}^{y_i};$$

$$f_{(X_{i2}|Y_i=y_i,\boldsymbol{\theta})}(x_{i2}) = \left\{\theta_{02}^{x_{i2}}(1-\theta_{02})^{1-x_{i2}}\right\}^{1-y_i} \times \left\{\theta_{12}^{x_{i2}}(1-\theta_{12})^{1-x_{i2}}\right\}^{y_i}.$$

Exercice 5

Montrez que la vraisemblance des paramètres pour l'ensemble des n message peut s'exprimer sous la forme suivante :

$$\begin{aligned} f_{\{(\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{Y})|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{y}) \\ &= \theta_{01}^{n_{01}}\ (1-\theta_{01})^{n_0-n_{01}} \times \theta_{11}^{n_{11}}\ (1-\theta_{11})^{n_1-n_{11}} \\ &\times \theta_{02}^{n_{02}}\ (1-\theta_{02})^{n_0-n_{02}} \times \theta_{12}^{n_{12}}\ (1-\theta_{12})^{n_1-n_{12}} \\ &\times \theta^{n_1}\ (1-\theta)^{n_0}; \end{aligned}$$

avec

n_0 **:** Le nombre de pourriels ;

n_1 : Le nombre de courriels ;

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

n_{11} : Le nombre de courriels où le mot *enlarge* apparaît ;

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

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

Grâce à l'hypothèse d'indépendance conditionnelle, chaque variable explicative peut être traitée indépendamment. En effet, on remarquer de l'exercice 5 que la vraisemblance se factorise en fonction des variables explicatives.

Exercice 6

Si la loi *a priori* suivante est utilisée :

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathcal{B}eta(\theta_{01} \mid 1,1) \times \mathcal{B}eta(\theta_{11} \mid 1,1) \\ \times \mathcal{B}eta(\theta_{02} \mid 1,1) \times \mathcal{B}eta(\theta_{12} \mid 1,1) \\ \times \mathcal{B}eta(\theta \mid 1,1), \end{aligned}$$

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

$$\begin{aligned} f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) \propto \mathcal{B}eta(\theta_{01} \mid n_{01}+1, n_0 - n_{01}+1) \times \mathcal{B}eta(\theta_{11} \mid n_{11}+1, n_1 - n_{11}+1) \\ \times \mathcal{B}eta(\theta_{02} \mid n_{02}+1, n_0 - n_{02}+1) \times \mathcal{B}eta(\theta_{12} \mid n_{12}+1, n_1 - n_{12}+1) \\ \times \mathcal{B}eta(\theta \mid n_1+1, n_0+1). \end{aligned}$$

8.3.1 Loi prédictive

Remarque. *Afin de ne pas alourdir la notation de cette section, le conditionnement sur l'échantillon aléatoire* $(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})$ *est omis.*

On reçoit un nouveau message qui contient les mots *enlarge* et *dear*, *i.e.* $\tilde{X}_1 = 1$ et

$\tilde{X}_2 = 1$, calculons la probabilité prédictive que ce message soit un courriel :

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} = 1 \,|\tilde{X}_1 = 1, \tilde{X}_2 = 1) \\
&\qquad \propto \mathbb{P}(\tilde{X}_1 = 1 \cap \tilde{X}_2 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{Y} = 1) \text{ (théorème de Bayes)} \\
&\qquad \propto \mathbb{P}(\tilde{X}_1 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{X}_2 = 1 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{Y} = 1) \text{ (ind. cond.)} \\
&\qquad \propto \frac{n_{11} + 1}{n_1 + 2} \times \frac{n_{12} + 1}{n_1 + 2} \times \frac{n_1 + 1}{n + 2} = q_1 \text{ (calcul par conditionnement).}
\end{aligned}$$

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

$$\mathbb{P}(\tilde{Y} = 0 \mid \tilde{X}_1 = 1, \tilde{X}_2 = 1) \propto \frac{n_{01} + 1}{n_0 + 2} \times \frac{n_{02} + 1}{n_0 + 2} \times \frac{n_0 + 1}{n + 2} = q_0$$

Si $q_1 > q_0$, alors le message sera classé comme un courriel. Sinon, il sera classé comme un pourriel.

Exercice 7

Soit le nouveau message $\tilde{Y}$ ne contenant pas les mots *enlarge* et *dear*, *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 :

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} = 1 \,|\tilde{X}_1 = 0, \tilde{X}_2 = 0) \\
&\qquad \propto \mathbb{P}(\tilde{X}_1 = 0 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{X}_2 = 0 \mid \tilde{Y} = 1) \times \mathbb{P}(\tilde{Y} = 1) \\
&\qquad \propto \frac{n_1 - n_{11} + 1}{n_1 + 2} \times \frac{n_1 - n_{12} + 1}{n_1 + 2} \times \frac{n_1 + 1}{n + 2}.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{P}(\tilde{Y} = 0 \,|\tilde{X}_1 = 0, \tilde{X}_2 = 0) \\
&\qquad \propto \mathbb{P}(\tilde{X}_1 = 0 \mid \tilde{Y} = 0) \times \mathbb{P}(\tilde{X}_2 = 0 \mid \tilde{Y} = 0) \times \mathbb{P}(\tilde{Y} = 0) \\
&\qquad \propto \frac{n_0 - n_{01} + 1}{n_0 + 2} \times \frac{n_0 - n_{02} + 1}{n_0 + 2} \times \frac{n_0 + 1}{n + 2}.
\end{aligned}$$

8.4 Inclusion de p 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 :

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

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 :

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

où

n_0 **:** Le nombre de pourriels.

n_1 **:** Le nombre de courriels.

n_{0j} **:** Le nombre de pourriels où la variable X_j est vraie.

n_{1j} **:** Le nombre de courriels où la variable X_j est vraie.

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 :

$$\begin{aligned}&\mathbb{P}(\tilde{Y}=1|\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{x}},-)\\&\qquad\propto\left\{\prod_{j=1}^{p}\mathbb{P}(X_j=\tilde{x}_j|Y=1,\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{X}=\boldsymbol{x})\right\}\times\mathbb{P}(\tilde{Y}=1|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{X}=\boldsymbol{x}).\end{aligned}$$

De façon analogue, on trouve que la probabilité prédictive que le message soit un pourriel

est proportionnelle à l'expression suivante :

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

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

8.5 Le cas où la variable d'intérêt possède plus de deux catégories

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 :

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

8.6 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.