Probabilités conditionnelles
Problème de recherche : le filtre anti-spam

Une entreprise développe un filtre anti-spam pour sa messagerie. Elle a analysé un grand nombre de courriels reçus :

— 60 % des courriels reçus sont des spams ;
— le mot « gratuit » apparaît dans 30 % des spams ;
— le mot « gratuit » apparaît dans 2 % des courriels qui ne sont pas des spams.

La mission

1. Un courriel contient le mot « gratuit ». Quelle est la probabilité que ce soit un spam ?
2. Un courriel ne contient pas le mot « gratuit ». Quelle est la probabilité que ce soit un spam ?
3. Le filtre doit-il envoyer automatiquement dans les spams tous les courriels qui contiennent le mot « gratuit » ? Discuter.

Pour aller plus loin

Une autre entreprise reçoit seulement 10 % de spams (les autres pourcentages sont les mêmes). La présence du mot « gratuit » est-elle encore un bon indice ?

Coups de pouce

Bloqué ? Des indices progressifs sont disponibles en ligne. N'en ouvrir qu'un à la fois, et seulement après avoir cherché !

Éléments de correction

Un tableau d'effectifs

Imaginons 10 000 courriels :

	Contient « gratuit »	Ne contient pas « gratuit »	Total
Spam	$30\,\% \times 6\,000 = 1\,800$	4 200	6 000
Pas spam	$2\,\% \times 4\,000 = 80$	3 920	4 000
Total	1 880	8 120	10 000

1. Courriel avec « gratuit »

$P_G(S) = \frac{1\,800}{1\,880} \approx \mathbf{0{,}957}$. (Avec un arbre : $\frac{0{,}6\times 0{,}3}{0{,}6\times 0{,}3+0{,}4\times 0{,}02} = \frac{0{,}18}{0{,}188}$.)

2. Courriel sans « gratuit »

$P_{\overline{G}}(S) = \frac{4\,200}{8\,120} \approx \mathbf{0{,}517}$. L'absence du mot ne dit presque rien : la plupart des spams ne contiennent pas « gratuit ».

3. Discussion

Environ 4,3 % des courriels contenant « gratuit » sont des messages légitimes : sur 10 000 courriels, 80 messages importants seraient perdus dans les spams. Un vrai filtre combine de nombreux indices (expéditeur, liens, autres mots…) avant de décider, et laisse l'utilisateur vérifier le dossier des spams.

Pour aller plus loin

Avec 10 % de spams : $P_G(S) = \frac{0{,}1\times 0{,}3}{0{,}1\times 0{,}3+0{,}9\times 0{,}02} = \frac{0{,}03}{0{,}048} = 0{,}625$. Le mot est un indice beaucoup moins fiable : la probabilité dépend fortement de la proportion de spams.

Ce qu'on a utilisé du chapitre

— Tableau croisé d'effectifs, arbre pondéré.
— Probabilité conditionnelle ; distinguer $P_G(S)$ et $P_S(G)$.

Pour aller plus loin (enrichissement)

Les filtres réels, dits « bayésiens », calculent ce type de probabilité pour des milliers de mots et les combinent.