

2 Classification naïve bayésienne pour détecter les pourriels

La classification naïve bayésienne est un algorithme d'apprentissage supervisé qui fonctionne avec les probabilités. Un problème classique qui peut être traité avec cet algorithme est la classification de courriels. On peut tenter d'estimer la probabilité qu'un courriel soit en fait un pourriel en prenant en compte les mots particuliers qu'il contient, l'idée étant que certains mots auront tendance à être plus souvent utilisés dans des pourriels.

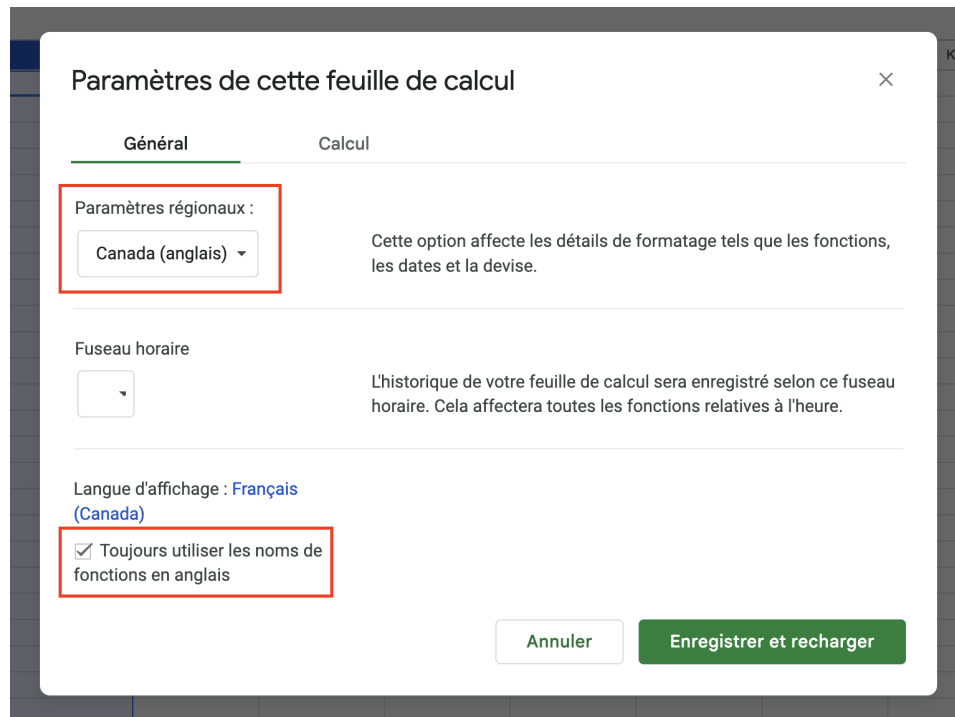
La probabilité qui nous intéresse est en fait une probabilité conditionnelle, celle du fait qu'un courriel particulier soit ou non un pourriel, étant donné les mots particuliers qui le composent. Un courriel sera classifié en tant que pourriel si :

$$\text{Prob}(\text{oui c'est un pourriel} \mid \text{mots}) > \text{Prob}(\text{non ce n'est pas un} \mid \text{mots})$$

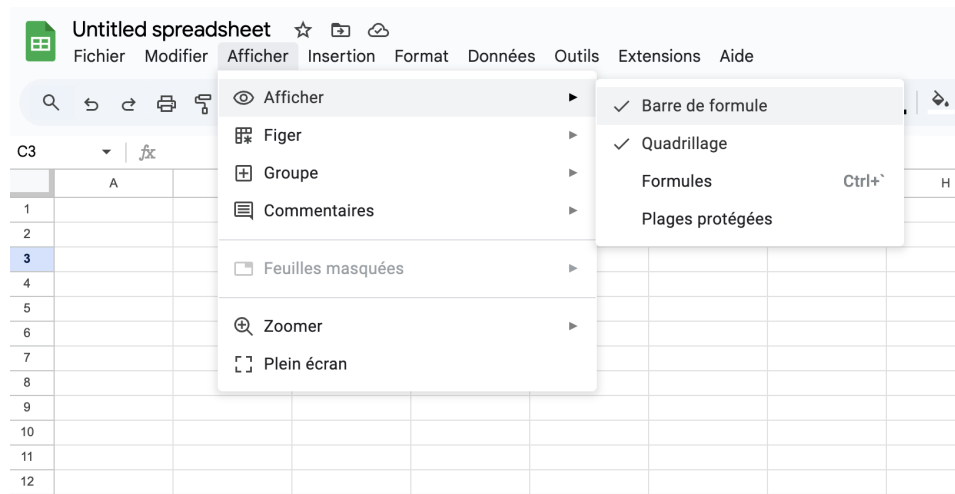
2.1 Entraînement du modèle

Voyons comment il est possible de calculer ces probabilités en entraînant un modèle de classification sur une série de courriels particuliers.

Nous allons utiliser le tableur Google Sheets au lieu de Excel, car Google Sheets est plus accessible, et le langage de ses formules est plus facile à gérer (celui d'Excel dépend de la langue et des paramètres régionaux de votre système d'exploitation). Pour éviter la confusion dans le contexte de ce travail, nous devons tout d'abord nous assurer que la langue des fonctions et des paramètres régionaux est l'anglais :



Assurez-vous ensuite que la "barre de formules" soit visible :



Copiez tout d'abord ces 10 courriels dans la colonne A d'une nouvelle "feuille" Google Sheets, un courriel par rangée :

voici le colis est arrivé
bonjour voici le lien
offre spéciale colis gratuit
merci pour votre colis
colis livré demain matin
voici votre carte gratuite
réunion demain à midi
voici le code pour carte
livraison spéciale pour vous
merci encore pour votre carte

Pour avoir un aperçu de la tâche d'étiquetage des données (qui dans un scénario réel peut s'avérer très coûteuse et laborieuse), vous êtes invités à tenter tout d'abord de catégoriser les courriels dans la colonne B, en utilisant la valeur "oui" si vous considérez qu'il s'agit d'un pourriel, ou "non" (ce n'est pas un pourriel) sinon.

Si vous n'avez pas envie de vous soumettre à cet exercice à ce stade, vous pouvez toujours copier ces valeurs (dans la colonne B) :

non
non
oui
non
non
oui
non
non
oui
non

À ce stade, votre feuille devrait ressembler à ceci :

	A	B	C	D	E
1	voici le colis est arrivé	non			
2	bonjour voici le lien	non			
3	offre spéciale colis gratuit	oui			
4	merci pour votre colis	non			
5	colis livré demain matin	non			
6	voici votre carte gratuite	oui			
7	réunion demain à midi	non			
8	voici le code pour carte	non			
9	livraison spéciale pour vous	oui			
10	merci encore pour votre carte	non			
11					
12					

Calculons tout d'abord dans la colonne C la probabilité à priori qu'un courriel quelconque soit un pourriel ou non (sans prendre en considérations les mots donc, pour le moment) :

`=MAP(UNIQUE(B1:B10), LAMBDA(x, COUNTIF(B1:B10, x) / COUNTA(B1:B10)))`

Ces probabilités à priori nous serviront plus loin. Définissez ensuite la colonne D avec cette formule :

=UNIQUE(TRANPOSE(SPLIT(TEXTJOIN(" ", TRUE, A:A), " ")))

La colonne D devrait maintenant contenir le vocabulaire des courriels :

Feuille de calcul sans titre

Fichier Modifier Afficher Insertion Format Données Outils Extensions Aide

100% CA\$ % .0 .00 123 Par dé... - 10 + B I ÷ A

D1 =UNIQUE(TRANPOSE(SPLIT(TEXTJOIN(" ", TRUE, A:A), " ")))

	A	B	C	D	E	F
1	voici le colis est arrivé	non	0.7	voici		
2	bonjour voici le lien	non	0.3	le		
3	offre spéciale colis gratuit	oui		colis		
4	merci pour votre colis	non		est		
5	colis livré demain matin	non		arrivé		
6	voici votre carte gratuite	oui		bonjour		
7	réunion demain à midi	non		lien		
8	voici le code pour carte	non		offre		
9	livraison spéciale pour vous	oui		spéciale		
10	merci encore pour votre carte	non		gratuit		
11				merci		
12				pour		
13				votre		
14				livré		
15				demain		
16				matin		
17				carte		
18				gratuite		
19				réunion		
20				à		
21				midi		
22				code		
23				livraison		
24				vous		
25				encore		
26						

La colonne E devrait ensuite correspondre au nombre de fois où les mots de la colonne D apparaissent dans les courriels valides (qui donc "non", ne sont pas des pourriels) :

=SUMPRODUCT((B\$1:B\$10="non") * ISNUMBER(SEARCH(D1, A\$1:A\$10)))

et de manière similaire pour la colonne F et la fréquence des mots qui apparaissent dans les courriels qui "oui", sont des pourriels :

=SUMPRODUCT((B\$1:B\$10="oui") * ISNUMBER(SEARCH(D1, A\$1:A\$10)))

Notez que les colonnes E et F doivent avoir le même nombre d'éléments que la colonne D (il faut donc utiliser la fonction de remplissage automatique, pour laquelle le plus simple est de soit glisser (drag) la première cellule vers

le bas, une fois qu'elle a été calculée, ou encore de double-cliquer sur le petit "+" noir qui apparaît en bas à droite de la première cellule).

Feuille de calcul sans titre

Fichier Modifier Afficher Insertion Format Données Outils Extensions Aide

100% CA\$ % .0 .00 123 Par dé... 10 + B I A

E1 =SUMPRODUCT((B\$1:B\$10="non") * ISNUMBER(SEARCH(D1, A\$1:A\$10)))

	A	B	C	D	E	F	G
1	voici le colis est arrivé	non	0.7	voici	3		
2	bonjour voici le lien	non	0.3	le			
3	offre spéciale colis gratuit	oui		colis			
4	merci pour votre colis	non		est			
5	colis livré demain matin	non		arrivé			
6	voici votre carte gratuite	oui		bonjour			
7	réunion demain à midi	non		lien			
8	voici le code pour carte	non		offre			
9	livraison spéciale pour vous	oui		spéciale			
10	merci encore pour votre carte	non		gratuit			
11				merci			
12				pour			
13				votre			
14				livré			
15				demain			
16				matin			
17				carte			
18				gratuite			
19				réunion			
20				à			
21				midi			
22				code			
23				livraison			
24				vous			
25				encore			
26							

Feuille de calcul sans titre

Fichier Modifier Afficher Insertion Format Données Outils Extensions Aide

100% CA\$ % .0 .00 123 Par dé... - 10 + B I A

F1 =SUMPRODUCT((B\$1:B\$10="oui") * ISNUMBER(SEARCH(D1, A\$1:A\$10)))

	A	B	C	D	E	F	G
1	voici le colis est arrivé	non	0.7	voici	3	1	
2	bonjour voici le lien	non	0.3	le	3	2	
3	offre spéciale colis gratuit	oui		colis	3	1	
4	merci pour votre colis	non		est	1	0	
5	colis livré demain matin	non		arrivé	1	0	
6	voici votre carte gratuite	oui		bonjour	1	0	
7	réunion demain à midi	non		lien	1	0	
8	voici le code pour carte	non		offre	0	1	
9	livraison spéciale pour vous	oui		spéciale	0	2	
10	merci encore pour votre carte	non		gratuit	0	2	
11				merci	2	0	
12				pour	3	1	
13				votre	2	1	
14				livré	1	0	
15				demain	2	0	
16				matin	1	0	
17				carte	2	1	
18				gratuite	0	1	
19				réunion	1	0	
20				à	1	0	
21				midi	1	0	
22				code	1	0	
23				livraison	0	1	
24				vous	0	1	
25				encore	1	0	
26							

À partir de ces fréquences de mots pour chaque classe ("oui" ou "non"), on peut maintenant calculer la probabilité conditionnelle de chaque mot du vocabulaire, étant donné le fait qu'un courriel soit "oui" ou "non" un pourriel. Donc la colonne G correspond à la probabilité des mots étant donné que "non" il ne s'agit pas d'un pourriel :

$$=(E1 + 1) / (\text{SUM}(E:E) + \text{COUNTA}(D:D))$$

et de manière similaire la colonne H est la probabilité des mots quand on sait que "oui" il s'agit d'un pourriel :

$$=(F1 + 1) / (\text{SUM}(F:F) + \text{COUNTA}(D:D))$$

Encore une fois les colonnes G et H doivent avoir la même taille que celle du vocabulaire (colonne D), il faut donc s'assurer d'utiliser le mécanisme du remplissage automatique décrit précédemment.

Feuille de calcul sans titre								
Fichier Modifier Afficher Insertion Format Données Outils Extensions Aide								
100% CA\$ % 123 Par dé... 10 B I A								
H1	= (F1 + 1) / (SUM(F:F) + COUNTA(D:D))							
	A	B	C	D	E	F	G	H
1	voici le colis est arrivé	non	0.7	voici	3	1	0.07142857143	0.05
2	bonjour voici le lien	non	0.3	le	3	2	0.07142857143	0.075
3	offre spéciale colis gratuit	oui		colis	3	1	0.07142857143	0.05
4	merci pour votre colis	non		est	1	0	0.03571428571	0.025
5	colis livré demain matin	non		arrivé	1	0	0.03571428571	0.025
6	voici votre carte gratuite	oui		bonjour	1	0	0.03571428571	0.025
7	réunion demain à midi	non		lien	1	0	0.03571428571	0.025
8	voici le code pour carte	non		offre	0	1	0.01785714286	0.05
9	livraison spéciale pour vous	oui		spéciale	0	2	0.01785714286	0.075
10	merci encore pour votre carte	non		gratuit	0	2	0.01785714286	0.075
11				merci	2	0	0.05357142857	0.025
12				pour	3	1	0.07142857143	0.05
13				votre	2	1	0.05357142857	0.05
14				livré	1	0	0.03571428571	0.025
15				demain	2	0	0.05357142857	0.025
16				matin	1	0	0.03571428571	0.025
17				carte	2	1	0.05357142857	0.05
18				gratuite	0	1	0.01785714286	0.05
19				réunion	1	0	0.03571428571	0.025
20				à	1	0	0.03571428571	0.025
21				midi	1	0	0.03571428571	0.025
22				code	1	0	0.03571428571	0.025
23				livraison	0	1	0.01785714286	0.05
24				vous	0	1	0.01785714286	0.05
25				encore	1	0	0.03571428571	0.025
26								

Notre modèle est maintenant entièrement entraîné, et il est donc prêt pour son utilisation!

2.2 Utilisation du modèle (inférence)

Nous allons maintenant utiliser le modèle pour déterminer si un nouveau courriel (qui n'a pas servi à l'entraînement) est un pourriel ou non. Dans la colonne I entrez un courriel à tester :

voici votre carte spéciale

Faites l'extraction des mots du courriel dans la colonne J :

=TRANSPOSE(SPLIT(I1, " "))

Nous avons maintenant besoin, dans la colonne K, de la probabilité des mots de ce courriel de test dans l'hypothèse où "non", ça ne serait pas un pourriel :

=IFERROR(XLOOKUP(J1, D:D, G:G), 1E-5)

et de manière similaire pour la colonne L, avec la probabilité des mots du courriel dans l'hypothèse où "oui" il s'agit d'un pourriel :

2.3 Questions

1. Que se passe-t-il si vous changez le mot "spéciale" par le mot "livrée" dans le courriel de test de la cellule I1?
2. Expliquez en vos mots ce qui se passe avec les probabilités conditionnelles du mot "livrée", aux cellules K4 et L4. Pourquoi a-t-on besoin de faire en sorte que ça fonctionne ainsi?
3. Est-ce que ce modèle est paramétrique ou non? Expliquez pourquoi.
4. S'il s'agit d'un modèle paramétrique, quels sont les paramètres du modèle exactement?
5. Quelle est la signification des nombres dans les cellules G11 et H11, comment peut-on les interpréter?
6. Quelles sont les différentes probabilités conditionnelles de ce modèles?
7. Quelles sont les probabilités non-conditionnelles (à priori)?
8. Est-ce qu'il serait possible d'utiliser seulement ces probabilités non-conditionnelles pour faire un modèle? Quelles conséquences ça entraînerait?
9. De quelle manière peut-ton dire que ce modèle généralise?
10. Est-ce que l'ordre des mots joue un rôle dans les décisions de ce modèle? Expliquez pourquoi c'est ainsi
11. Si l'ordre des mots ne joue pas de rôle, comment pourrait-on modifier le modèle de manière à ce qu'il en joue un?
12. Est-ce que certains mots aident particulièrement le modèle? Si oui pourquoi?
13. Est-ce que certains mots sont moins utiles? Si oui pourquoi?