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 :

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

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.

Copiez tout d'abord ces 10 courriels dans la colonne A d'un feuillet Excel, un courriel par rangée :

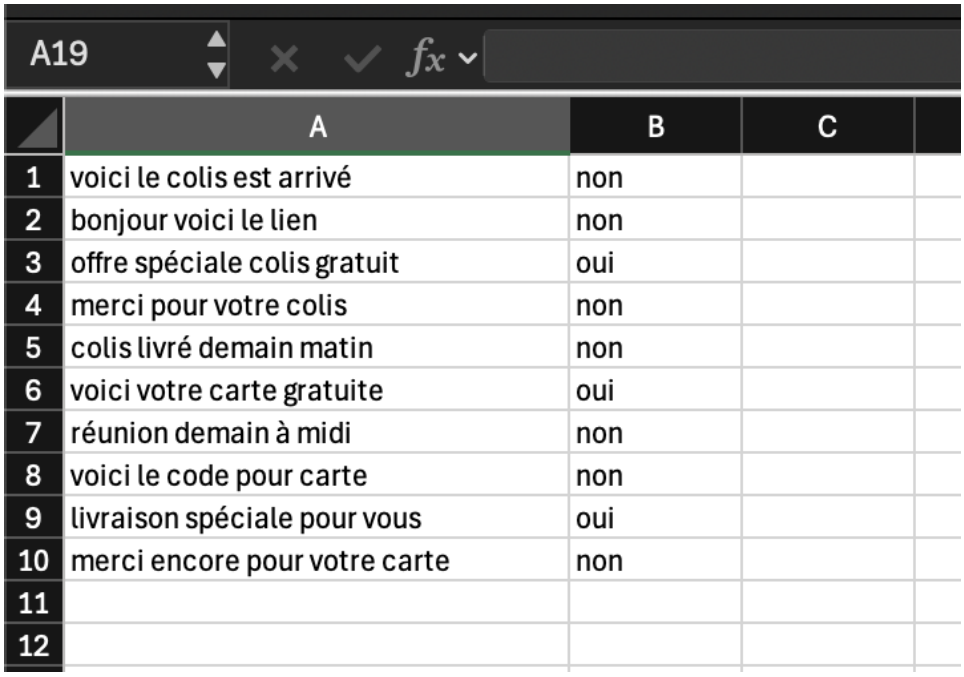
```
voici le colis est arrivé
bonjour voici le lien
offre spéciale colis gratuit
merci pour votre colis
colis livré demain matin
voici votre carte gratuite
réunion demain à midi
voici le code pour carte
livraison spéciale pour vous
merci encore pour votre carte
```

Pour avoir un aperçu de la tâche d'étiquettage 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 feuillet Excel devrait ressembler à ceci :



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) :

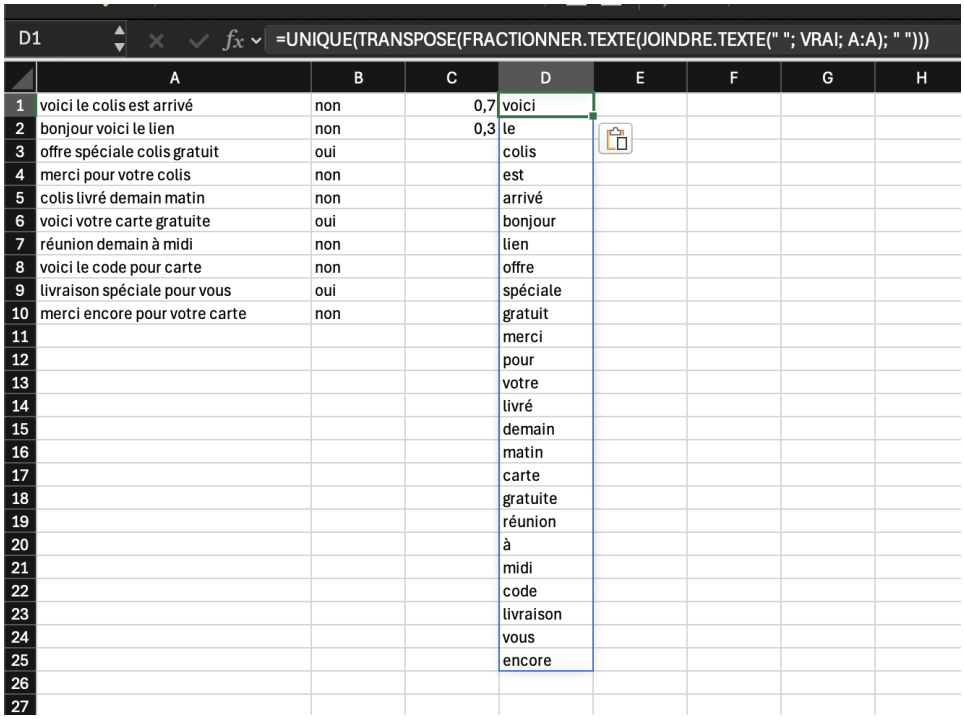
```
=NB.SI(B1:B10; UNIQUE(B1:B10)) / NBVAL(B1:B10)
```

Ces probabilités à priori nous serviront plus loin. Définissez ensuite la

colonne D avec cette formule :

```
=UNIQUE(TRANSPOSE(FRACTIONNER.TEXTE(JOINDRE.TEXTE(" "; VRAI; A:A); " ")))
```

La colonne D devrait maintenant contenir le vocabulaire des courriels :



	A	B	C	D	E	F	G	H
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								
27								

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) :

```
=SOMMEPROD((B$1:B$10="non") * ESTNUM(CHERCHE(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 :

```
=SOMMEPROD((B$1:B$10="oui") * ESTNUM(CHERCHE(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 apparait en bas à droite de la première cellule).

E1 =SOMMEPROD((B$1:B$10="non") * ESTNUM(CHERCHE(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							

F1 =SOMMEPROD((B$1:B$10="oui") * ESTNUM(CHERCHE(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

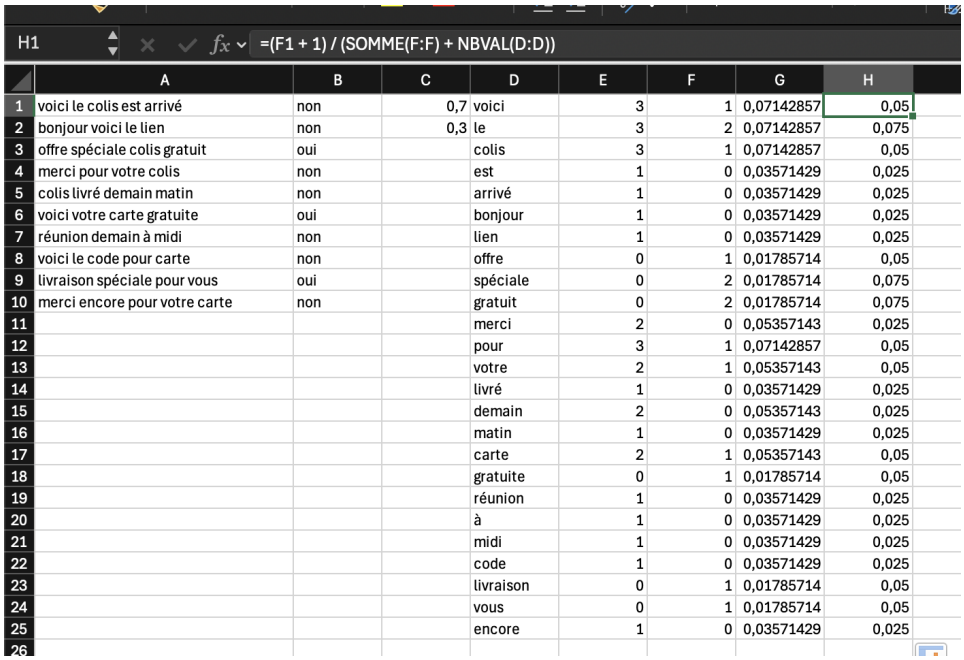
du vocabulaire, étant donné le fait qu'un courriel soit "oui" ou "non" un pourriel. Donc la colonne G correspond à la probabilité des mots étant donné que "non" il ne s'agit pas d'un pourriel :

```
=(E1 + 1) / (SOMME(E:E) + NBVAL(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) / (SOMME(F:F) + NBVAL(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.



	A	B	C	D	E	F	G	H
1	voici le colis est arrivé	non	0,7	voici	3	1	0,07142857	0,05
2	bonjour voici le lien	non	0,3	le	3	2	0,07142857	0,075
3	offre spéciale colis gratuit	oui		colis	3	1	0,07142857	0,05
4	merci pour votre colis	non		est	1	0	0,03571429	0,025
5	colis livré demain matin	non		arrivé	1	0	0,03571429	0,025
6	voici votre carte gratuite	oui		bonjour	1	0	0,03571429	0,025
7	réunion demain à midi	non		lien	1	0	0,03571429	0,025
8	voici le code pour carte	non		offre	0	1	0,01785714	0,05
9	livraison spéciale pour vous	oui		spéciale	0	2	0,01785714	0,075
10	merci encore pour votre carte	non		gratuit	0	2	0,01785714	0,075
11				merci	2	0	0,05357143	0,025
12				pour	3	1	0,07142857	0,05
13				votre	2	1	0,05357143	0,05
14				livré	1	0	0,03571429	0,025
15				demain	2	0	0,05357143	0,025
16				matin	1	0	0,03571429	0,025
17				carte	2	1	0,05357143	0,05
18				gratuite	0	1	0,01785714	0,05
19				réunion	1	0	0,03571429	0,025
20				à	1	0	0,03571429	0,025
21				midi	1	0	0,03571429	0,025
22				code	1	0	0,03571429	0,025
23				livraison	0	1	0,01785714	0,05
24				vous	0	1	0,01785714	0,05
25				encore	1	0	0,03571429	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(FRACTIONNER.TEXTE(MINUSCULE(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 :

`=SIERREUR(RECHERCHEX(FILTRE(J:J; J:J<>""); D:D; G:G); 1E-10)`

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 :

`=SIERREUR(RECHERCHEX(FILTRE(J:J; J:J<>""); D:D; H:H); 1E-10)`

Calculons dans la colonne M la probabilité que "non" le courriel n'est pas un pourriel :

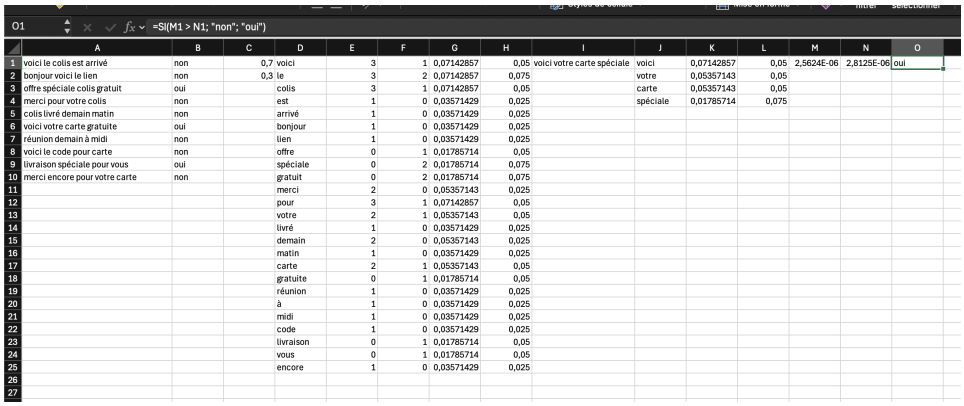
`=PRODUIT(K:K) * C1`

Et dans la colonne N la probabilité que "oui" le courriel est un pourriel :

`=PRODUIT(L:L) * C2`

Notre classification finale sera dans la colonne O :

`=SI(M1 > N1; "non"; "oui")`



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	voici le colis est arrivé	non	0,7	voici	3	1	0,07142857	0,05	voici votre carte spéciale	voici	0,07142857	0,05	2,5624E-06	2,8125E-06	oui
2	bonjour voici le lien	non	0,3	le	3	2	0,07142857	0,075		votre	0,05357143	0,05			
3	offre spéciale colis gratuit	oui		colis	3	1	0,07142857	0,05		carte	0,05357143	0,05			
4	merci pour votre colis	non		est	1	0	0,03571429	0,025		spéciale	0,01785714	0,075			
5	colis livré demain matin	non		arrivé	1	0	0,03571429	0,025							
6	voici votre carte gratuite	oui		bonjour	1	0	0,03571429	0,025							
7	réunion demain à midi	non		lien	1	0	0,03571429	0,025							
8	voici le code pour carte	non		offre	0	1	0,01785714	0,05							
9	livraison spéciale pour vous	oui		spéciale	0	2	0,01785714	0,075							
10	merci encore pour votre carte	non		gratuit	0	2	0,01785714	0,075							
11				merci	2	0	0,05357143	0,025							
12				pour	3	1	0,07142857	0,05							
13				votre	2	1	0,05357143	0,05							
14				livré	1	0	0,03571429	0,025							
15				demain	2	0	0,05357143	0,025							
16				matin	1	0	0,03571429	0,025							
17				carte	2	1	0,05357143	0,05							
18				gratuite	0	1	0,01785714	0,05							
19				réunion	1	0	0,03571429	0,025							
20				à	1	0	0,03571429	0,025							
21				midi	1	0	0,03571429	0,025							
22				code	1	0	0,03571429	0,025							
23				livraison	0	1	0,01785714	0,05							
24				vous	0	1	0,01785714	0,05							
25				encore	1	0	0,03571429	0,025							
26															
27															

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 entrainerait?
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 le soit?
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?