Fouille de données TP4 : Classification documentaire

Pierre Petitbon Florian Privé Xinrui Xu

Q1) Nous avons un fichier contenant, sur chaque ligne, un vecteur correspondant à chaque document. Chaque composante i d'un vecteur correspond au nombre d'occurrence du $i^{ème}$ mot du vocabulaire. Pour optimiser l'espace de stockage, ce vecteur est codé dans le fichier sous la forme indice-valeur qui consiste à obtenir une représentation compacte des vecteurs de documents en ne codant que les mots qui sont présents dans le document ainsi que le nombre d'occurrence associé à ce mot.

Ainsi, pour trouver la taille du vocabulaire, il faut écrire un parser qui lit les indices présents dans chaque vecteur et trouver le maximum de ces indices. Le plus grand des indices correspond au nombre de mots différents qui sont présents dans l'ensemble des documents de la base. On obtient ainsi la taille du vocabulaire : $|V| = 141144$

Le premier nombre de chaque ligne correspond à la classe à laquelle appartient ce document. Pour trouver le nombre de documents dans une classe k, il suffit de compter le nombre de lignes commençant par k. On obtient ainsi :

classe	nombre de documents
1	5894
2	1003
3	2472
4	2207
5	6010
6	2992
7	1586
8	1226
9	2007
10	3982
11	7757
12	3644
13	3405
14	2307
15	1040
16	1460
17	1191
18	1733
19	4745
20	1411
21	1016
22	3018

classe	nombre de documents
23	1050
24	1184
25	1624
26	1296
27	1018
28	1049
29	1376

Q2) On veut scinder aléatoirement les 70703 documents en deux ensembles :
— base d'entraînement (52500 documents)
— base de test (18203 documents)

Soit $ratio = \frac{52500}{70703}$
On va tirer, en suivant une lui uniforme[0,1], une valeur r, pour chaque document di :
— Si $r < ratio$ on place le document di dans la base apprentissage.
— Sinon on le place dans la base test.
Ainsi la taille de la base d'apprentissage suit une loi binômiale de paramètres $n = 70703$ et $p = ratio$. L'espérance de cette loi est de $n * p = 52500$ et son écart-type est de $sqrtn * p * (1 - p) = 116$. Vu la taille du problème, la différences entre la taille de la base d'apprentissage obtenue et celle voulue est négligeable.

Q3) — Le modèle de Bernouilli :

Le modèle a pour paramètres (apprentissage) :
$\hat{\theta}_{t_i|k} = \frac{df_{t_i}(k)+1}{N_k(S)+2}$(lissage de Laplace pour ne que $\hat{\theta}_{t_i|k}$ soit non nul)
$\hat{\pi}_k = \frac{N_k(S)}{m}$
La prédiction :
$k(d^{'}) = argmax_{k\in[1..K]}(ln(\hat{\pi}_k) + \sum\limits_{t_i \in d'} ln(\hat{\theta}_{t_i|k}) + \sum\limits_{t_i \notin d'} ln(1 - \hat{\theta}_{t_i|k}))$

— Le modèle multinomial :

Le modèle a pour paramètres (apprentissage) :
$\hat{\theta}_{t_i|k} = \frac{\sum\limits_{d \in S_k} tf_{t_i,d}+1}{\sum\limits_{1}^{V} \sum\limits_{d \in S_k} tf_{t_i,d}+V}$
$\hat{\pi}_k = \frac{N_k(S)}{m}$
La prédiction :
$k(d^{'}) = argmax_{k\in[1..K]}(ln(\hat{\pi}_k) + \sum\limits_{t_i \in d'} w_{id'} ln(\hat{\theta}_{t_i|k}))$

Nous avons codé les formules du modèle Bernouilli et du modèle multinomial pour la phase d'apprentissage et la phase de prédiction. Nous avons modifié quelques formules du modèle Bernouilli afin d'optimiser le code.

— Lors de la phase d'apprentissagede Bernouilli, nous avons choisi de calculer seulement les $df_{t_i}(k)$ et les $N_k(S)$ au lieu des $\hat{\theta}_{t_i|k}$ et des $\hat{\pi}_k$ pour ne stocker que des uint16_t au lieu des doubles (ce qui prendrait beaucoup plus de mémoire).

— Lors de la phase de prédiction de Bernouilli, nous n'avons pas codé la formule exactement. Nous avons calculé $ln(\hat{\pi}_k) + \sum_{t_i \in Vocab} ln(1 - \hat{\theta}_{t_i|k}) + \sum_{t_i \notin d'} (ln(\hat{\theta}_{t_i|k}) - ln(1 - \hat{\theta}_{t_i|k}))$ de façon à minimiser le nombre de calculs de log qui ralentirait beaucoup le programme.

Q4) Avec le modèle Bernouilli : Taux de bonne classification = 55%.
Avec le modèle Multinomiale : Taux de bonne classification = 75%.

Q5) A l'issue de 20 expériences, nous calculons la moyenne, la variance et l'écart-type des taux de bonne classification :

— Avec le modèle de Bernouilli : La moyenne vaut : 54.814625%. La variance vaut : 0.082031. L'ecart-type vaut : 0.286411.

— Avec le modèle Multinomial : La moyenne vaut : 74.771996%. La variance vaut : 0.085938. L'ecart-type vaut : 0.293151.

Conclusion :
Avec les deux modèles, la variance et l'écart-type des taux de bonne classification restent plutôt faible, donc les résultats des deux modèles sont très stables. La moyenne sur les taux de bonne classification avec les deux modèles sont assez différents. Le modèle Multinomial présente un taux de bonne classification largement supérieur à celui du modèle de Bernouilli (différence de 20%). Le modèle Multinomial est donc meilleur. Cela peut s'expliquer par le fait que le modèle de Bernouilli ne prend que en compte l'absence ou la présence d'un mot dans un document, alors que le modèle Multinomial prend également en compte la fréquence d'apparition d'un mot dans un document.