***Feuille de route : Test du chi-deux :***

1. Objectif :

On cherche à faire un test d’indépendance entre les différents termes et les deux catégories que nous avons (ham et spam) le but étant de choisir les termes qui sont le plus dépendant de la catégorie.

1. Formalisation des hypothèses :

H0 : le fait d’avoir un terme précis ne nous permet pas de prédire la catégorie (en d’autres termes le terme et la catégorie sont indépendants)

H1 : : le fait d’avoir un terme précis nous permet de prédire la catégorie (en d’autres termes le terme et la catégorie sont dépendants)

1. Critères de sélection :

* On cherche à rejeter H0 parce qu’on cherche les termes qui permettent de connaître la catégorie
* On cherche ainsi à sélectionner les plus petites p-value (la p-value dans le cadre d’un khi-deux est la probabilité de ne pas rejeter H0)

1. Format de la matrice (qui va servir d’Input à la fonction chi-square) :

Dans ce cas nous aurons besoin d’une matrice avec en colonne nos prédicteurs qui sont les termes en plus de la variable cible qui est la catégorie (on peut faire mettre une variable binaire avec 0 ham et 1 spam). Tandis que les lignes seront les différents documents. Ça va donner quelque chose qui ressemble à ça :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Term1 | Term2 | Terme3 | … | TermeT | catégorie |
| Document1 | frequence |  |  |  |  | 1 |
| Document2 |  |  |  |  |  | 0 |
| Document3 |  |  |  |  |  | 0 |
| ….. |  |  |  |  |  | 0 |
| DocumentD |  |  |  |  |  | 1 |

1. Modalité d’application :

Je pense à une sorte de boucle pour faire le test du chi-deux tour à tour pour chacun des termes avec la variable catégorie pour voir s’ils sont dépendent ou indépendant ( car pour chacun des terme nous avons besoin d’une p-value du test du chi-deux pour savoir si on va la garder ou pas).

Ainsi dans la première itération on va faire

chisq.test(x,y)

avec x= factor contenant les fréquences du Terme1 dans chacun des documents

|  |  |
| --- | --- |
|  | Term1 |
| Document1 | frequence |
| Document2 |  |
| Document3 |  |
| ….. |  |
| DocumentD |  |

Et y = factor contenant si le document appartient à la catégorie spam ou ham

|  |  |
| --- | --- |
|  | catégorie |
| Document1 | 1 |
| Document2 | 0 |
| Document3 | 0 |
| ….. | 0 |
| DocumentD | 1 |

Le y va rester le même dans chacune des itérations. C’est le x qui va changer à chaque fois (terme1, term2, term3 etc.).

Enfin on pourra faire une sélection pour les 25 plus petites valeurs du p-values pour voir avoir les termes les plus dépendants aux catégories.

Références :

Ref1 : <https://stackoverflow.com/questions/38538821/feature-selection-in-document-feature-matrix-by-using-chi-squared-test>

Ref2 : <https://data-flair.training/blogs/chi-square-test-in-r/>

Ref3 : <https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223>