TD : *Latent Dirichlet Allocation*

Romain Tavenard

Dans cette séance, nous nous focaliserons sur la manipulation d'un modèle de type *topic model* vu en cours : le modèle *Latent Dirichlet Allocation*.

Pour s'entrainer à manipuler ce modèle, vous allez utiliser un (petit) sous-ensemble d'un jeu de données "historique" du domaine de la fouille de texte qui se nomme `20 Newsgroup` et qui consiste, comme son nom l'indique, en de courts documents textuels issus de 20 forums (*newsgroups*) thématiques différents.

Pour charger ces données enregistrées au format `numpy` compressé, vous utiliserez la fonction `numpy.load` comme suit (après avoir téléchargé le fichier `20newsgroup.npz` sur CURSUS) :

```
content = numpy.load("20newsgroup.npz")
X, y, vocab = content["X"], content["y"], content["vocabulary"]
```

`X` est une matrice terme-document (dans laquelle j'ai supprimé un certain nombre de *stop-words* pour obtenir des *topics* un peu plus lisibles), y un vecteur indiquant la catégorie dans laquelle chacun des documents a été posté et `vocab` est un vecteur permettant de faire le lien entre des indices de termes et les termes en question. Ainsi, l'expression suivante affiche les termes indicés 20, 32 et 334 du corpus :

```
print(vocab[[20, 32, 334]])
```

Pour information, les étiquettes de classe correspondent aux *newsgroups* suivants :

```
alt.atheism 1
comp.graphics 2
comp.os.ms-windows.misc 3
comp.sys.ibm.pc.hardware 4
comp.sys.mac.hardware 5
comp.windows.x 6
misc.forsale 7
rec.autos 8
rec.motorcycles 9
rec.sport.baseball 10
rec.sport.hockey 11
sci.crypt 12
```

```
sci.electronics 13
sci.med 14
sci.space 15
soc.religion.christian 16
talk.politics.guns 17
talk.politics.mideast 18
talk.politics.misc 19
talk.religion.misc 20
```

1. Combien de documents contient la collection ? Et combien de termes sont définis ? Parmi ces termes, combien sont présents dans au moins un document du corpus ?

1 Impact du paramètre η

2. Observez l'impact du paramètre η sur la probabilité moyenne d'apparition des *topics* dans les documents. Cette probabilité moyenne d'apparition des *topics* dans les documents est une quantité importante qui permet de se rendre compte de l'importance relative de chacun des topics dans le mélange.
3. Quels sont les indices des 5 *topics* les plus importants extraits par LDA ? Et quels sont les 20 mots les plus probables dans chacun de ces *topics* ?
4. Faites de même document par document, pour 5 documents de votre choix.

2 Exercice de synthèse

On souhaite maintenant se focaliser sur un problème de classification supervisée à deux classes entre les articles issus des *newsgroups* `soc.religion.christian` et `comp.sys.mac.hardware`. Pour cela, un nouveau fichier de données a été préparé sur CURSUS, appelé `20newsgroup_binary.npz`

Créez un modèle de classification supervisée qui fonctionne en deux étapes :

a. Extraire une représentation LDA des documents du sous-corpus en question ;
b. Effectuer une classification SVM à noyau linéaire dans l'espace des *topics* LDA.

Vous devrez effectuer une validation croisée pour choisir le nombre de composantes de la décomposition LDA et l'hyper-paramètre C du SVM.

Quel est le contenu des *topics* des plus discriminants pour cette classification ?