INF8225 TP2 - Rapport de laboratoire

Apprentissage automatique par descente de gradiant

Maxime Schmitt
1719088

19 février 2016

1 Partie I - Descente par batch et mini-batch

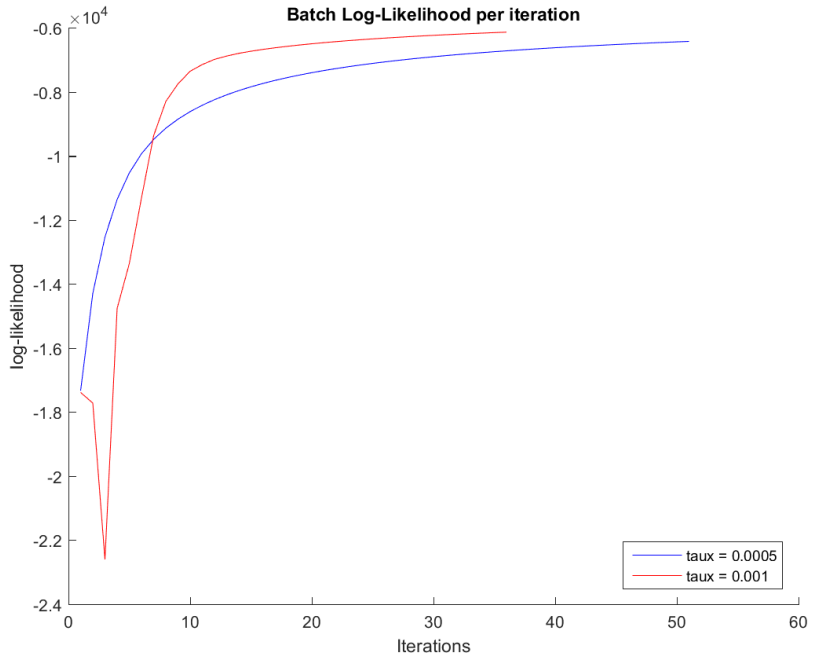
Le code pour la descente par batch peut être trouvé dans le fichier *gradientDescent.m* tandis que le code pour la descente par mini-batch peut être trouvé dans le fichier *miniBatchMultipleShuffle.m*. Dans la suite de cette partie, nous allons présenter puis analyser des résultats de l'apprentissage par ces deux méthodes en fonction de la valeur du taux d'apprentissage ainsi que celle du nombre de mini-batch utilisés pour la seconde méthode, avant de comparer les deux méthodes d'apprentissage. Pour toutes les expériences, la valeur du seuil de convergence qui détermine la condition de terminaison de l'apprentissage, lorsque la différence de log-vraisemblance entre deux itérations est inférieur à cette valeur, a été fixée à 15, car on a observé la combinaison de résultats et d'une durée de convergence raisonnable pour celle-ci.

1.1 Étude de l'influence du taux d'apprentissage sur l'apprentissage

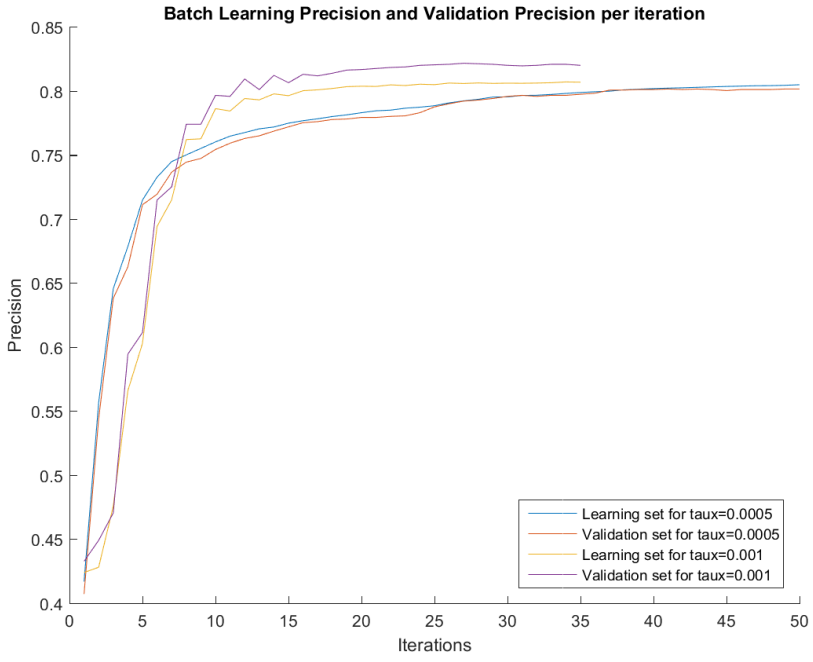
On trouve dans les figures 1 et 2 les résultats des apprentissages respectivement pour la méthode par batch et celle par mini-batch. Dans chaque figure, la figure (a) illustre les courbes de log-vraisemblance à chaque itération tandis que la figure (b) représente la précision sur l'ensemble d'apprentissage et sur l'ensemble de validation à chaque itération. Pour la méthode par mini-batch on a choisi un nombre de mini-batches de 20, comme proposé dans l'énoncé.

Une valeur de taux d'apprentissage plus grande entraîne une convergence plus rapide, en nombre d'itérations, même si cela n'est pas toujours observable en raison de la nature aléatoire de la séparation des données en les différents ensembles (apprentissage, validation, test) qui peuvent apporter des résultats très variés. De même, on devrait en général observer une meilleure précision sur l'ensemble d'apprentissage avec un taux plus faible avec éventuellement même un risque d'over-fitting plus important. Là encore, l'observation est difficilement réalisable pour la même raison.

Enfin, il est important de noter qu'il existe des valeurs pour ce taux d'apprentissage qui représentent un problème pour la construction du modèle. La figure 3 présente un tel cas avec une valeur de taux d'apprentissage de 0.005. Dans ce cas, le taux est trop grand et les oscillations très grandes. On a alors de fortes chances de ne jamais remplir la condition de convergence, et donc de ne pas obtenir de système intéressant, alors même que le nombre d'itérations devient important.

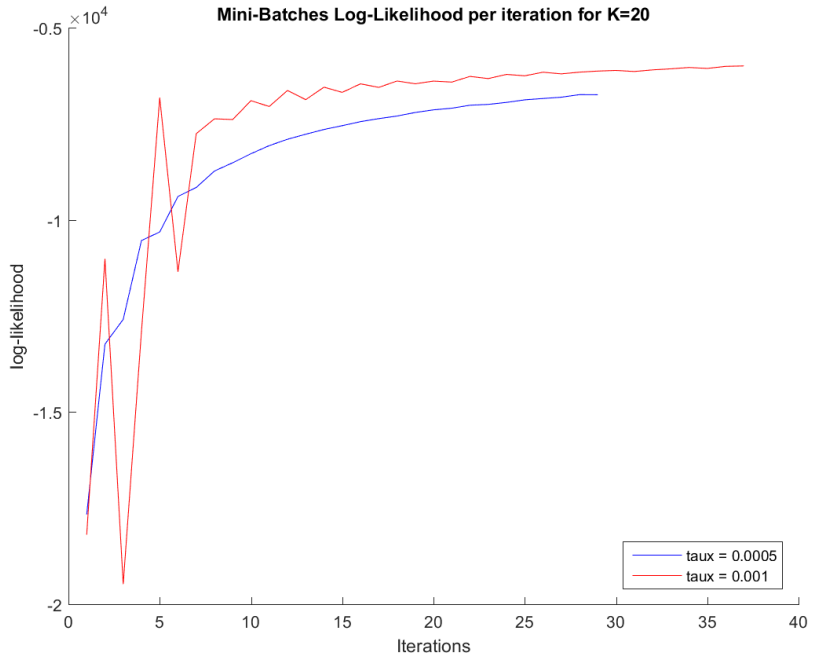


(a) Log-vraisemblance par itération

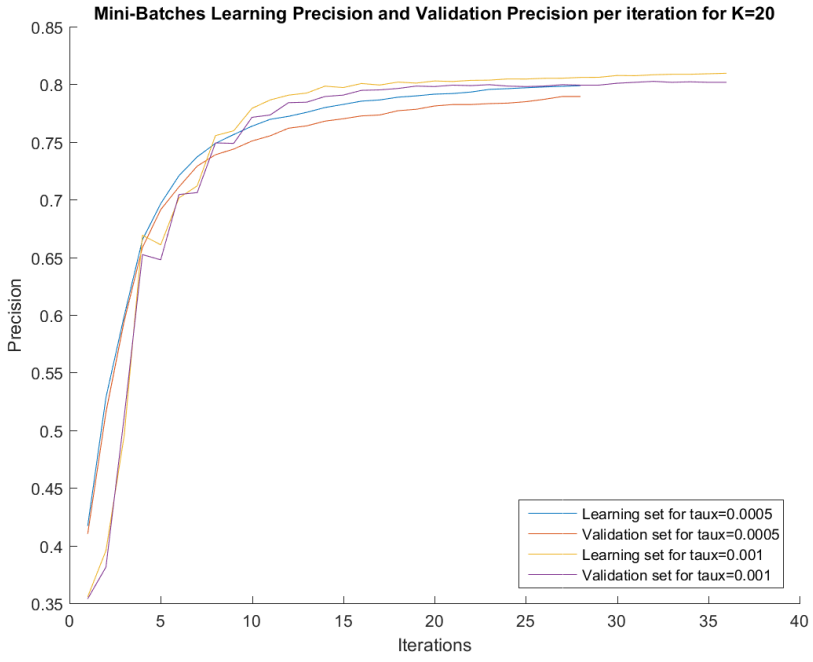


(b) Précisions par itération

FIGURE 1 – Courbes de l'apprentissage par batch

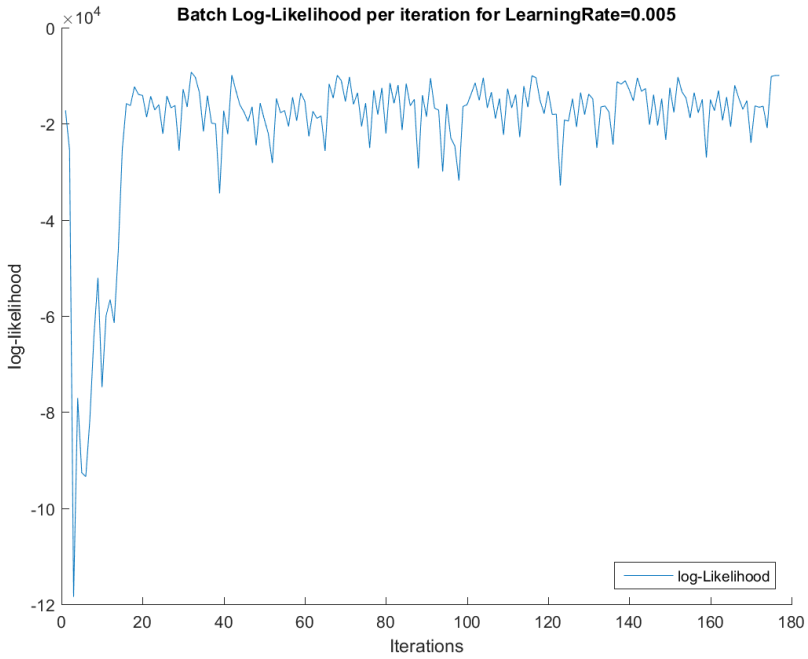


(a) Log-vraisemblance par itération

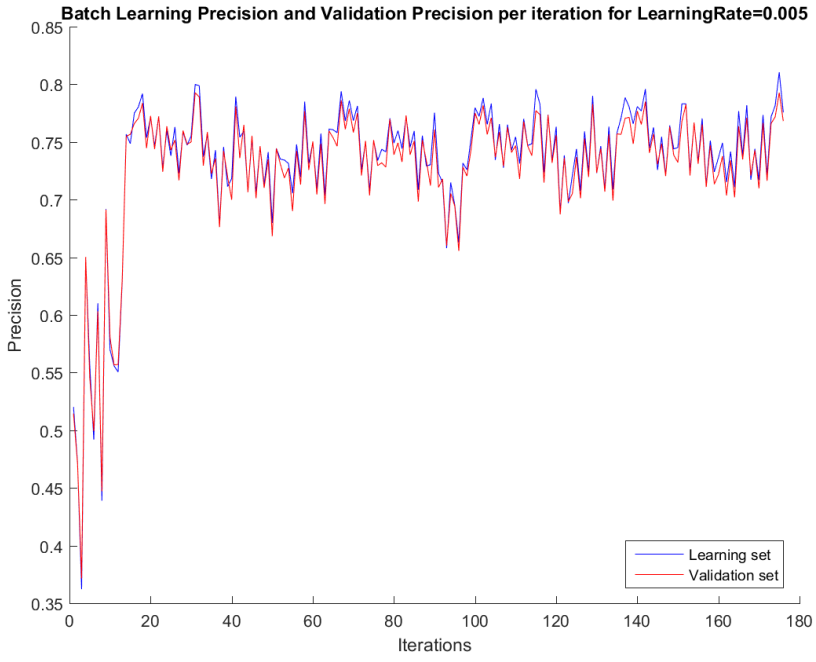


(b) Précisions par itération

FIGURE 2 – Courbes de l'apprentissage par mini-batch pour un nombre de batches de 20



(a) Log-vraisemblance par itération



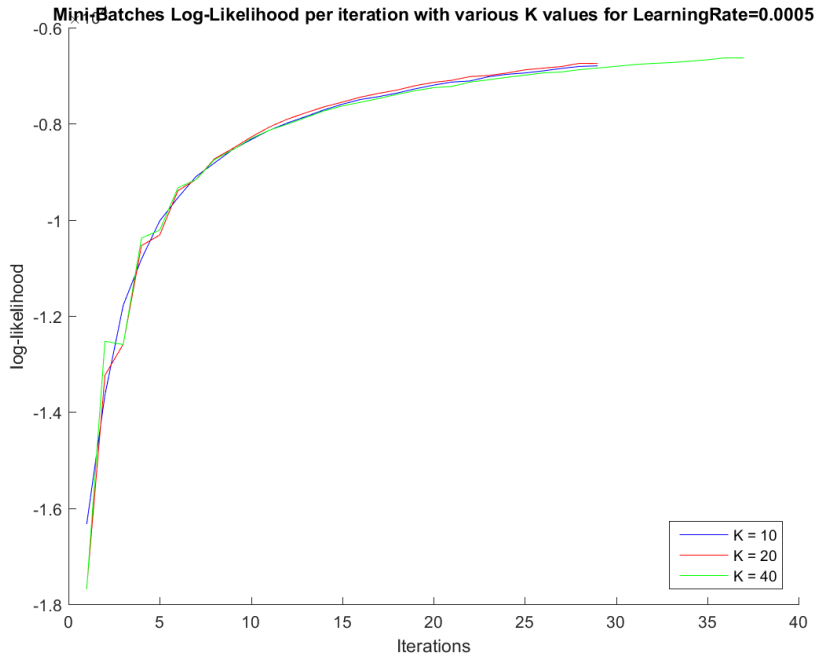
(b) Précisions par itération

FIGURE 3 – Courbes de l'apprentissage par batch

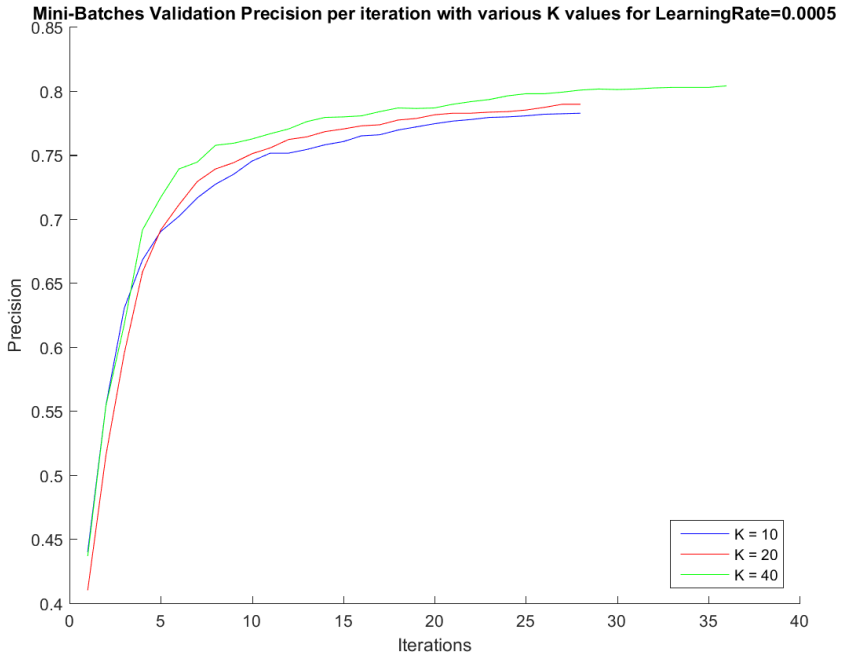
1.2 Étude de l'influence de la taille des mini-batch pour l'apprentissage par mini-batch

La figure 4 présente, pour un taux d'apprentissage de 0.0005, l'effet du nombre de mini-batches sur l'évolution du log-vraisemblance (figure 4(a)) et la précision sur l'ensemble de validation (figure 4(b)).

On constate qu'un grand nombre de mini-batches entraîne une meilleure précisions, ce qui est logique puisque les combinaisons possibles d'arrangement des données forment une représentation plus exhaustive de l'ensemble des données. Mais en contre-partie, le nombre d'itérations nécessaires pour arriver à convergence est plus grand, en raison de ce plus grand nombre de combinaisons possibles.



(a) Log-vraisemblance par itération

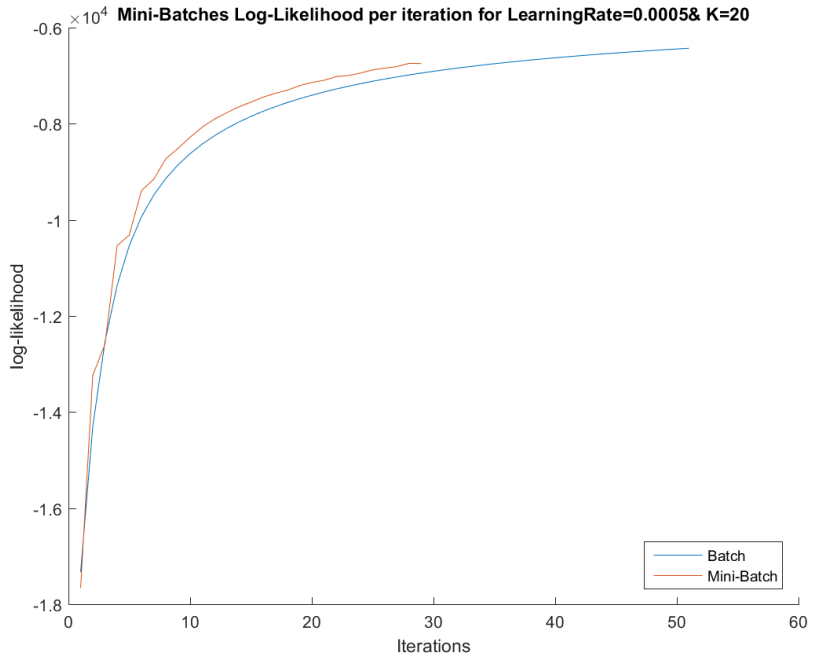


(b) Précision par itération

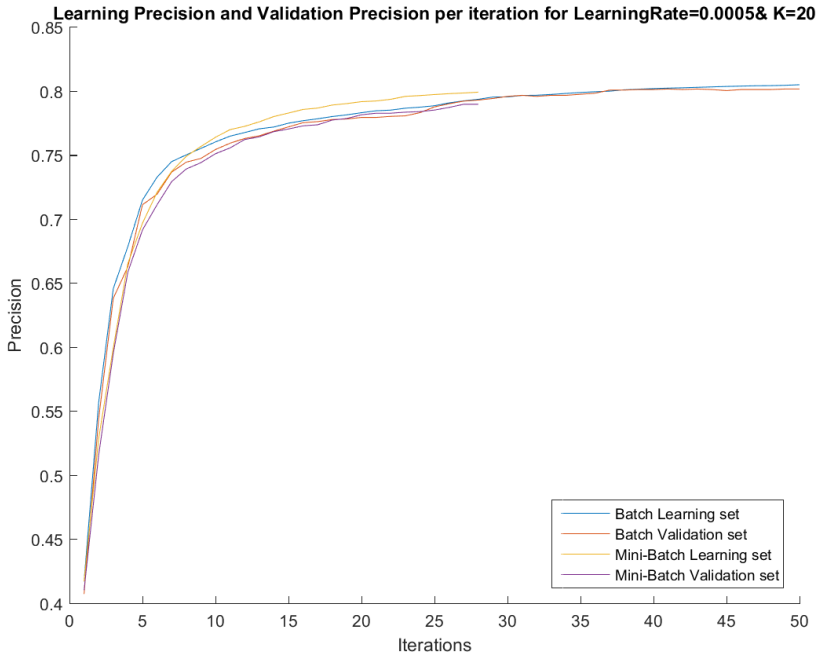
FIGURE 4 – Courbes de l'apprentissage par mini-batch pour un nombre de batches de 20

1.3 Comparaison des deux méthodes d'apprentissage

La figure 5 met en parallèle les résultats obtenus, pour un taux d'apprentissage de 0.0005, avec la méthode par batch et avec la méthode par mini-batch pour un nombre de mini-batches de 20. On constate donc que cette dernière offre une convergence plus rapide, en nombre d'itérations, et des résultats au moins équivalents à ce que l'on obtient avec la première méthode. De plus, comme on l'a vu dans la partie précédente, en modifiant le nombre de mini-batches à utiliser, on peut encore améliorer ce résultat tout en gardant une convergence plus rapide.



(a) Log-vraisemblance par itération



(b) Précision par itération

FIGURE 5 – Courbes de l'apprentissage par mini-batch pour un nombre de batches de 20

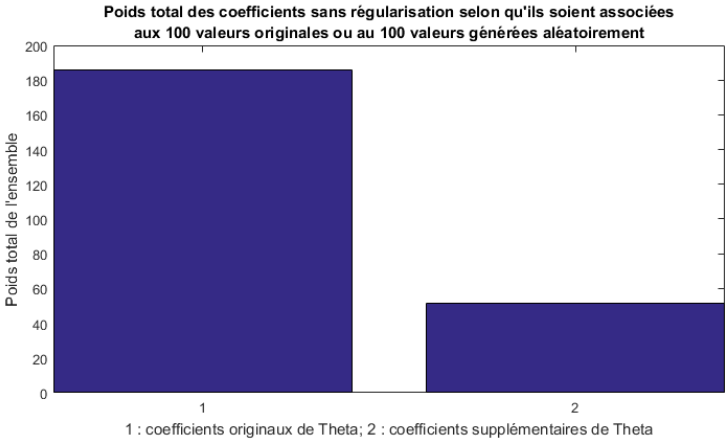
2 Partie II - Régularisation de type Elastic Net

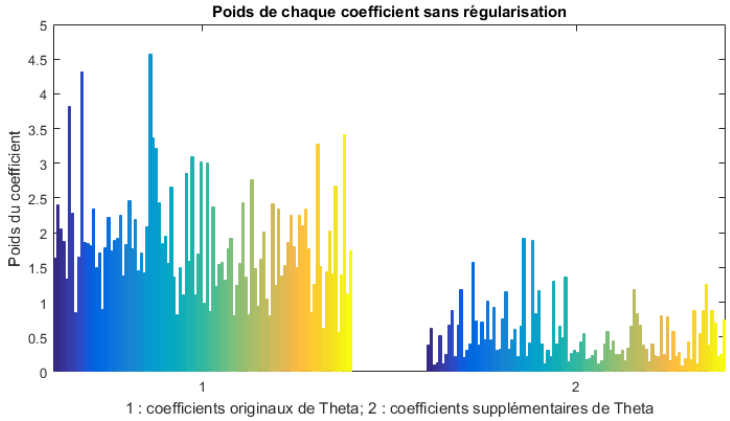
Le code pour cette partie peut être trouvé dans le fichier *elasticNet.m*. Les fichiers *elasticNetAllTests.m* et *experimentAnalysis.m* contiennent également du code relatif à cette partie. Le premier contient le code qui a permis de faire les expériences pour toutes les combinaisons de $\lambda 1$ et $\lambda 2$ avec le même découpage de l'ensemble de données en les ensembles d'apprentissage, de validation et de test. Le second contient le code ayant permis de déterminer la meilleur couple ($\lambda 1$,$\lambda 2$) différent de (0,0) en terme de précision sur l'ensemble de validation, dans notre expérience ce couple est (0.01,0.1).

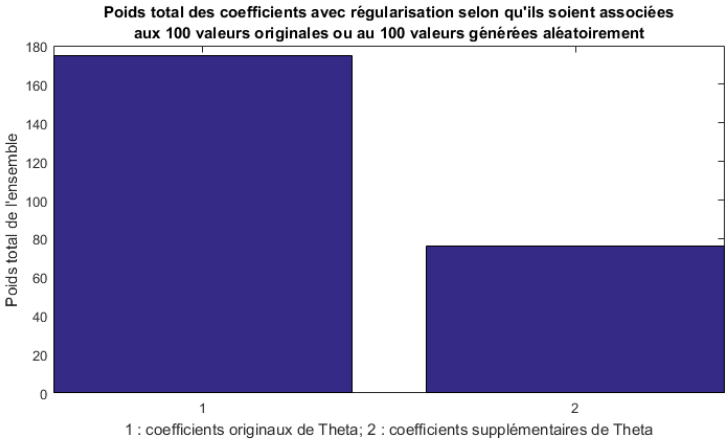
Pour cette partie, nous avons ajouté 100 caractéristiques aléatoires à notre modèle et nous proposons d'effectuer l'apprentissage avec une régularisation dite "Elastic Net". Pour analyser les résultats qui suivent, on définit le poids d'un x_i de l'ensemble de données (un x_i représentant donc la présence ou nom d'un mot dans un post) comme la somme des valeurs absolues des coefficients de θ associés à cet x_i. Cette quantité mesure donc l'importance de la dite caractéristique dans la catégorisation de la page dans l'une des quatre catégories. Par extension, on définit le poids d'un ensemble de x_i comme la somme de leurs poids respectifs.

La figure 6 présente les histogrammes des poids des x_i en séparant ceux-ci en deux catégories selon qu'ils soient originaux ou aient été introduits de façon aléatoire comme décrit précédemment. Les histogrammes de gauche représentent les poids totaux de ces ensembles dans le cas sans régularisation (en haut) et avec régularisation (en bas). Les histogrammes de droite présentent, pour les mêmes deux cas, les poids de chacun des x_i individuellement.

La première constatation est que les nouvelles caractéristiques ont des poids bien plus faibles que les caractéristiques originales, ce qui est le résultats attendu puisqu'elles ne devraient pas avoir d'influence sur l'appartenance à l'une ou l'autre des catégories ayant été déterminées aléatoirement. En revanche on constate qu'elles ont un poids plus important lorsqu'on effectue la régularisation de type Elastic Net que dans l'autre cas. On voit alors que cette régularisation permet de limiter l'over-fitting ce qui se traduit ici par une prise en compte plus importante des paramètres aléatoires.


Poids total des coefficients sans régularisation selon qu'ils soient associées aux 100 valeurs originales ou au 100 valeurs générées aléatoirement
Poids total de l'ensemble
0
20
40
60
80
100
120
140
160
180
200
1
2
1 : coefficients originaux de Theta; 2 : coefficients supplémentaires de Theta


Poids de chaque coefficient sans régularisation
Poids du coefficient
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
1
2
1 : coefficients originaux de Theta; 2 : coefficients supplémentaires de Theta


Poids total des coefficients avec régularisation selon qu'ils soient associées aux 100 valeurs originales ou au 100 valeurs générées aléatoirement
Poids total de l'ensemble
0
20
40
60
80
100
120
140
160
180
1
2
1 : coefficients originaux de Theta; 2 : coefficients supplémentaires de Theta

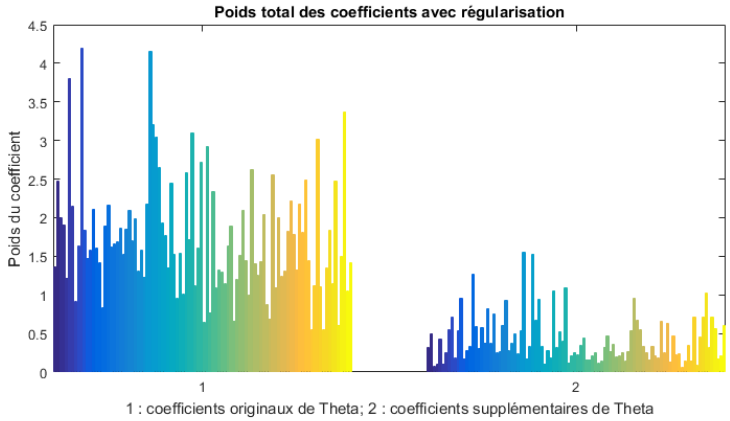

Poids total des coefficients avec régularisation
Poids du coefficient
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
1
2
1 : coefficients originaux de Theta; 2 : coefficients supplémentaires de Theta

FIGURE 6 –