Régression (II) - Régression linéaire, compléments

Dans les exercices qui suivent, sauf mention explicite du contraire, les données d'entrée seront systématiquement normalisées en appliquant une standardisation.

Exercice 1 (*). On considère les données du fichier `data_quad.csv` pour lesquelles on souhaite effectuer une régression polynomiale de y en x. L'erreur quadratique totale sera utilisée lors des deux premières questions, tandis que l'erreur absolue totale sera utilisée lors de la dernière question.

1. Donner la solution au problème de régression obtenue en utilisant un modèle polynomial de degré 2, représenter graphiquement les données ainsi que la courbe de régression.
2. Reprendre la question précédente en modifiant manuellement les données de la façon suivante (afin de faire apparaître des valeurs extrêmes) : remplacer les valeurs des ordonnées de rang multiple de 5, numérotés de 0 à 49, par la valeur 0.
3. *Facultatif.* En considérant toujours les données "extrêmes", donner la solution au problème de régression et représenter graphiquement les données ainsi que la courbe de régression, en utilisant cette fois-ci l'erreur absolue totale. Utiliser pour cela l'algorithme de descente de gradient.

Exercice 2 (**). On reprend les données du fichier `data.csv` (exercice 2 TD3) pour lesquelles on souhaite toujours effectuer une régression polynomiale de y en x au sens des moindres carrés. Néanmoins, le but de cet exercice est de construire un modèle qui réalise un surapprentissage volontaire des données. Choisir pour cela un modèle qui vous semble pertinent, puis résoudre le problème de régression et expliquer en quoi il réalise bien le surapprentissage voulu.

Exercice 3 (**). On reprend les données et le modèle de l'exercice précédent, qui faisait du surapprentissage. Lui appliquer une régularisation ridge, puis une régularisation LASSO. Donner les solutions obtenues dans les deux cas, et les comparer. Représenter les chemins de régularisation.

Exercice 4 (**). On reprend le contexte de l'exercice 3 du TD3 (données du prix de logements en Californie). Justifier l'intérêt d'intégrer une régularisation au problème de régression linéaire résolu lors du TD précédent, mettre en oeuvre cette régularisation et interpréter le résultat pour dégager la variable explicative ayant le plus d'impact sur la variable expliquée.

Exercice 5 (***). Montrer que la fonction d'erreur utilisée dans le cadre de la régression linéaire robuste, $\mathscr{E}(\theta) = \|y - X\theta\|_1$, est une fonction convexe de θ.