Master SID
Année universitaire 2021-2022

TP1 Apprentissage Automatique 2

Régression linéaire

simon.bernard@univ-rouen.fr

Le but de ce TP est d'implémenter et de tester une méthode de régression linéaire sur des problèmes réels. Pour cela, vous utiliserez les librairies python *Scikit-learn* [1], *Numpy* [2] et *Matplotlib.pyplot* [3] (et tout autre librairie que vous jugerez utiles).

1 Description des données

Vous allez travailler sur deux jeux de données :

- *Boston Housing*
- *Diabetes*

tous deux disponibles directement dans la librairie *Scikit-learn*, sous `sklearn.datasets`. Pour chaque exercice de ce TP, il vous est demandé de tester ces deux jeux de données.

▷ Dans un script python, chargez les deux jeux de données. Si vous ne savez pas comment faire, vous pouvez vous référer à la documentation de *Scikit-learn* [4]

▷ Analysez le contenu des deux jeux que vous avez chargez et donnez une description des données et de la problématique sous-jacente

▷ Visualisez l'ensemble des données en fonction de chaque variable explicative et de la réponse. Un exemple est montré en figure 1, où chaque sous-figure est la représentation des données en fonction d'une des variables explicatives en abscisse, et de la réponse en ordonnée.

2 Régression des moindres carrés ordinaires

▷ Créez la matrice d'apprentissage $\mathbf{X}$ telle que nous l'avons vu en cours, en ajoutant une colonne de 1 aux exemples. Cette matrice est appelé "matrice augmentée" par la suite.

▷ Estimez les paramètres des moindres carrés sur les données. Stockez ces paramètres dans un vecteur $\mathbf{w}$ et un biais b

▷ Affichez le nuage de points (*scatterplot*) des valeurs réponses prédites et réelles

▷ Calculez et affichez l'erreur quadratique moyenne et le coefficient de corrélation entre les valeurs prédites et réelles

3 Régression Ridge

Dans cette partie, vous allez tester et comparer plusieurs modèles entre eux. Pour cela, il faut séparer les données en deux sous-ensembles : un pour l'apprentissage et l'autre pour le test des modèles.

▷ Utilisez la fonction `train_test_split` de *Scikit-learn* [5] pour créer ces deux sous-ensembles. Vous pourrez par exemple, utiliser un tier des données pour le test et le reste pour l'apprentissage.

▷ Créez une fonction permettant de calculer la solution d'un problème de régression ridge en fonction de $\mathbf{X}$, $\mathbf{y}$ et λ

▷ Pour un ensemble de valeurs possibles de λ, calculez la solution du problème et calculez l'erreur quadratique et le coefficient de corrélation sur l'ensemble d'apprentissage et de test. On pourra utiliser par exemple les valeurs `lambda = numpy.logspace(-4,2,20)`

1. `https://scikit-learn.org/`
2. `https://numpy.org/`
3. `https://matplotlib.org/3.1.1/tutorials/introductory/pyplot.html`
4. `https://scikit-learn.org/stable/datasets/toy_dataset.html` et `https://scikit-learn.org/stable/auto_examples/index.html#dataset-examples`
5. `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

- ▷ Affichez sur une figure l'erreur en test et en apprentissage en fonction de λ. Dans la mesure du possible, on choisira deux échelles différentes pour les deux erreurs.

4 Normalisation

Il est souvent important lorsque l'on fait de la régression (et pas que) de normaliser les données pour faciliter l'apprentissage ou améliorer la qualité du modèle. Cela consiste souvent à uniformiser les valeurs prises par chaque variable explicative, par exemple pour faire en sorte que leurs moyennes soient 0 (on centre la variable) et leurs écarts-type soit 1 (on réduit la variable). Nous allons voir dans cet exercice quelques effets de ce type de pré-traitements.

- ▷ Séparez les données en deux sous-ensembles apprentissage/test
- ▷ Centrez les données d'apprentissage et leurs réponses, en retranchant aux valeurs la moyenne de l'échantillon
- ▷ Calculez le modèle de regression ridge sur les données centrées
- ▷ Affichez les valeurs
  - ○ du biais du modèle appris sur les données non-centrées (exercice précédent)
  - ○ de la moyenne des réponses de l'ensemble d'apprentissage
  - ○ du biais du modèle appris sur les données centrées mais avec une réponse non-centrée
  - ○ du biais du modèle appris sur les données centrées avec une réponse centrée
- ▷ Comparez maintenant les performances obtenus avec les deux modèles (sur les données centrées et non-centrées). Attention les données de test doivent être centrées avec la moyenne de l'ensemble d'apprentissage (car nous ne sommes pas sensé connaitre celle de l'ensemble de test).
- ▷ Que constatez-vous ?
- ▷ Refaites cette comparaison avec des données centrées et réduites

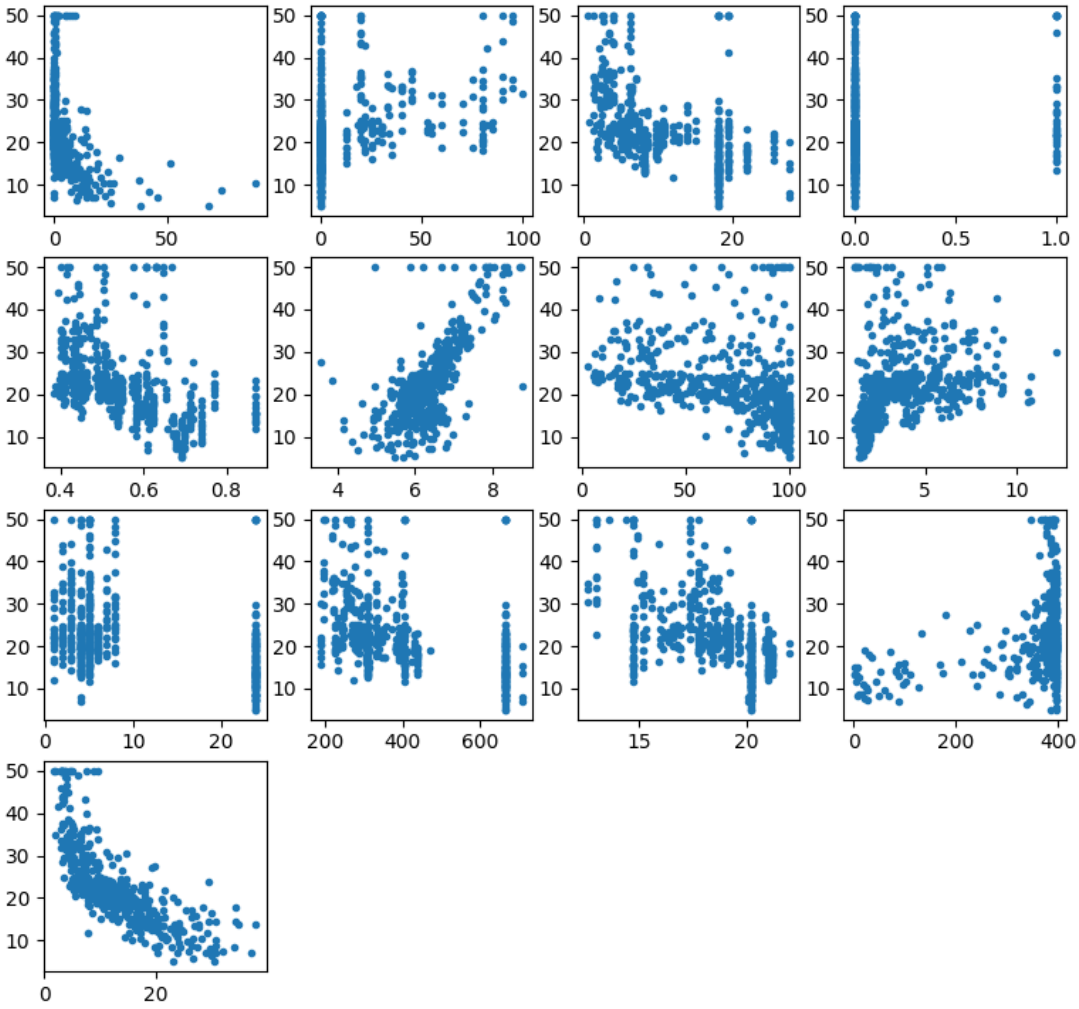
50
40
30
20
10
0
50
0
50
100
0
20
0.0
0.5
1.0
0.4
0.6
0.8
4
6
8
0
50
100
5
10
0
10
20
200
400
600
15
20
0
200
400
0
20

FIGURE 1