Estimation matricielle
en haute dimension
M2 DS

GEMA
GRANDES ECOLES DES METIERS D'AVENIR

Modèles de Régularisation

Objectif

Comparer les modéles de regression linéaire et régression linéaire régularisée.

Exercice 1.

1. Importer le dataset `auto-mpg`.
2. Séparer la variable endogène et les variables exogènes.
3. Diviser les données en `train` et `test`.
4. Transformer les données pour qu'elles soient sur la même échelle.
5. Construire des modèles de régression ridge, lasso et régression linéaire classique.
6. Donner la valeur RMSE des données `train` et `test` de chaque modèle.
7. Lequel de ces 3 modéles est plus efficace ? Pourquoi ?

Exercice 2. La base de données `BostonHousing` contient les variables suivantes :

Variable	Description
`CRIM`	Taux de criminalité par habitant dans la ville
`ZN`	Proportion de terrains résidentiels zonés pour des lots de plus de 25 000 pieds carrés
`INDUS`	Proportion de terrains non commerciaux (industrie) par ville
`CHAS`	Variable binaire pour la rivière Charles (1 si le quartier est en bord de rivière, 0 sinon)
`NOX`	Concentration d'oxydes d'azote (en parties par 10 millions)
`RM`	Nombre moyen de pièces par logement
`AGE`	Proportion de logements occupés construits avant 1940
`DIS`	Distances pondérées aux cinq principaux centres d'emploi de Boston
`RAD`	Indice d'accessibilité aux autoroutes radiales
`TAX`	Taux d'imposition foncière pour $10 000 de valeur
`PTRATIO`	Ratio élèves / enseignants par ville
`B`	$1000 \cdot (Bk - 0.63)^2$, où Bk est la proportion de résidents noirs (variable problématique)
`LSTAT`	Pourcentage de population à statut socio-économique faible
`MEDV`	**Valeur médiane des maisons occupées par leur propriétaire (en $1000s)**

1. Importer le dataset `BostonHousing`.
2. Explorer le dataset et afficher ses informations.
3. Vérifier s'il y a des données manquantes.
4. Transformer les données pour qu'elles soient sur la même échelle.

5. Étudier la corrélation entre les variables.
6. Existe-t-il des variables à enlever ? Lesquelles ?
7. Diviser les données en `train` et `test`.
8. Faire un modèle de régression linéaire pour prédire le prix des maisons `medv`.
9. Construire un modèle de régression `LASSO`.
   (a) Représenter les coefficients du modèle selon plusieurs valeurs de α.
   (b) Justifier le choix du paramètre α.
   (c) Combien de variables sont-elles éliminées ?
10. Construire un modèle de régression `RIDGE`, en choisissant la bonne valeur de α.
11. Donner le coefficient de détermination R^2 de chaque modèle.
12. Lequel de ces 3 modèles est plus efficace ? Pourquoi ?

Exercice 3. On travaille avec un jeu de données nommé `pisa.csv`, contenant une variable cible `non_cognitive` et plusieurs variables explicatives. L'objectif est de prédire cette variable en utilisant des techniques de régression pénalisée.

1. Chargement et division des données
   (a) Charger le jeu de données `pisa.csv`.
   (b) Diviser les données en un ensemble d'entraînement (70%) et un ensemble de test (30%).
2. Régression `Ridge` (sans mise à l'échelle)
   (a) Entraîner un modèle `Ridge` avec un paramètre de régularisation $\alpha = 0.001$.
   (b) Prédire sur les données d'entraînement.
   (c) Calculer les métriques suivantes : R^2 et `RMSE` (Root Mean Squared Error).
   (d) Commenter la performance du modèle.
3. Mise à l'échelle des variables
   (a) Appliquer une mise à l'échelle (standardisation) sur toutes les variables explicatives.
   (b) Réentraîner le modèle `Ridge` avec $\alpha = 0.001$.
   (c) Comparer les performances avec celles obtenues à l'étape précédente.
4. Grid Search sur `Ridge`
   (a) Mettre en place une validation croisée à 5 plis.
   (b) Tester 10 valeurs de α entre 10^{-4} et 10^4.
   (c) Afficher le score R^2 moyen pour chaque valeur.

(d) Tracer le graphe : α vs R^2.

(e) Quelle valeur de α donne la meilleure performance ?

5. Régression Lasso

   (a) Répéter la même démarche que pour `Ridge`, en utilisant un modèle `Lasso`.

   (b) Tester les mêmes valeurs de α (10 valeurs).

   (c) Afficher les résultats sous forme graphique.

   (d) Comparer avec `Ridge` :

   - Quel modèle est le plus performant ?
   - Le `Lasso` met-il certains coefficients à zéro ?

6. Régression `Elastic Net`

   (a) Entraîner un modèle `Elastic Net` avec double tuning :

   - $\alpha \in \{10^{-4}, \ldots, 10^{1}\}$
   - l_1-ratio $\in \{0, 0.25, 0.5, 0.75, 1\}$

   (b) Réaliser une validation croisée sur la grille complète.

   (c) Visualiser les résultats avec une carte de chaleur (`heatmap`).

   (d) Quel couple (α, l_1-ratio) donne le meilleur score ?

7. Comparaison finale

   (a) Pour les trois modèles (`Ridge`, `Lasso`, `Elastic Net`), prédire sur les jeux d'entraînement et de test.

   (b) Calculer le R^2 et le `RMSE` pour chaque modèle.

   (c) Présenter les résultats dans un tableau comparatif.

   (d) Quel modèle généralise le mieux ?

   (e) Préférez-vous un modèle très performant ou un modèle interprétable (comme `Lasso`) ? Justifiez.