Séance 3.1: Exemple d'algorithme d'apprentissage automatique supervisé

Démonstration

Aoudou Njingouo

SICSS - Montréal

13 June 2024

Plan de présentation

Session 2 : Régression linéaire

Concept de la régression linéaire Méthode des moindres carrés Évaluation des performances : RMSE, MAE, R^2 Travail pratique : Implémentation de la régression linéaire avec lm() et visualisation avec ggplot2.

Introduction : pourquoi la régularisation ?

- Les moodèles de regresssion classique peuvent souvent conduire à un surajustement(overfitting)
- Le surapprentissage peut être causé par la présence de variables inutiles ou redondadantes, ou par une quantitité insuffisante de donnée d'entrainement.
- Les modèles pénalisés, sont des techniques de régularisation qui permettent de limiter le surajustement en ajoutant des termes de pénalité à la fonction de coût de la régression.

Introduction : pourquoi la régularisation ?

- Les termes de pénalité encouragent le modèle à avoir des coefficients plus petits, ce qui reduit la complexité du modèle et limite l'influence des variables inutiles ou redondantes.
- Les modèles pénalisés permettent également de sélectionner automatiquement les variables les plus importantes, ce qui peut améliorer la compréhension du modèle et la qualité des prédictions.
- Les modèles pénalisés sont particulièrement utiles lorsque le nombre de variables indépendantes est élevé ou lorsque les données d"entrainement sont limitées.
- En resumé, les modèles pénalisés sont une technique importante pour limiter l'overffiting et améliorer la généralisation des modèles de régression

Types de régularisation

- Régression ridge
- Régression Lasso
- Régression Elastic Net

Régression Ridge ou L2

1. Principe de la régularisation L2

- La régression Ridge ajoute une pénalité L2 à la fonction de coût
- La pénalité L2 ajoute une contrainte sur les coefficients de régression, en les limitants à des valeurs proche de Zéro
- Ridge permet de réduire la variance du modèle et améliorer la capacité à généraliser

Régression Ridge ou L2

2 Fonction de coût

$$\sum_{i=1}^{n}(Y_i - X\beta|_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, 0 \leq \lambda \leq 1.$$

- Faire une régression Ridge c'est donc résoudre le problème d'optimisation suivant :

$$\text{Minimiser} \sum_{i=1}^{n}(Y_i - X\beta|_i)^2 \text{sous la contrainte} \sum_{j=1}^{p} \beta_j^2 \leq s$$

Régression Lasso OU L1

1. Principe de la régularisation L1

- La régression Lasso ajoute une pénalité L1 à la fonction de coût de la régression.
- La pénalité L1 impose une contrainte plus forte sur les coefficients de régression, *en les poussant à zéro pour certains coefficients.*
- Cette méthode permet de selectionner automatiquement les variables les plus importantes et de réduire la dimensionnalité du modèle.

2. Fonction de coût

$$\sum_{i=1}^{n} (Y_i - X\beta|_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, 0 \leq \lambda \leq 1.$$

Régression Elastic Net

- Elastic Net est une Combinaison de la régularisation L1 et L2
- Sa fonction de coût est :

$$\sum_{i=1}^{n}(Y_i - X\beta|_i)^2 + (1-\alpha)(\sum_{j=1}^{p}\beta_j^2) + (\alpha)(\lambda\sum_{j=1}^{p}|\beta_j|)$$

- Cette méthode permet de combiner les avantages de la régression Ridge et Lasso, en *réduisant la variance du modèle et en sélectionnant les variables les plus importantes.* - Avantages par rapport aux autres méthodes

Méthodes d'estimation

- Méthode des moindres carrés
- Méthode de descente de gradient

Comparaison des méthodes

- Forces et faiblesses de chaque méthode
- Choix en fonction du contexte

Exemple pratique

- Application des régressions régularisées sur un jeu de données
- Comparaison des résultats avec la régression classique

Conclusion

- Avantages des régressions régularisées
- Utilisation dans différents domaines

Ressources supplémentaires pour approfondir le sujet

- https://glmnet.stanford.edu/articles/glmnet.html
- https://github.com/Labo-Lacourse/Code_chap_23_logistic_regression_regularization

Modèles d'apprentissage automatique en arbre

Qu'est-ce qu'un modèle d'apprentissage automatique en arbre ?

- Définition : Un modèle d'apprentissage automatique en arbre est une méthode d'apprentissage supervisé qui utilise une structure arborescente pour prendre des décisions ou effectuer des prédictions.
- Caractéristiques : Les modèles en arbre sont faciles à comprendre, interprétables et adaptés à la fois pour des problèmes de classification et de régression.

Arbre de décision

- Définition : Un arbre de décision est un modèle en arbre qui représente les décisions et leurs conséquences sous forme d'arbre.
- Fonctionnement : L'arbre est construit de manière récursive en divisant les données en fonction de certaines caractéristiques jusqu'à obtenir des feuilles contenant des résultats finaux.
- Avantages : Interprétabilité, prise de décision transparente, robustesse aux valeurs manquantes.

Arbre de décision (suite)

- Limitations : Tendance à l'overfitting (surapprentissage) sur des ensembles de données complexes, sensibilité aux variations mineures des données d'entraînement.

Random Forest

- Définition : La Random Forest est un modèle en ensemble (ensemble learning) qui combine plusieurs arbres de décision.
- Fonctionnement : Chaque arbre est entraîné sur un sous-ensemble aléatoire des données et des caractéristiques, et les prédictions sont agrégées pour produire un résultat final.
- Avantages : Réduction de l'overfitting grâce à la combinaison de multiples arbres, performances élevées, robustesse aux valeurs aberrantes.

Random Forest (suite)

- Utilisations : Classification, régression, détection d'anomalies, sélection de variables.
- Importance des caractéristiques : Les Random Forests permettent de calculer l'importance relative des différentes caractéristiques dans la prédiction.

Comparaison entre l'arbre de décision et la Random Forest

- Complexité : L'arbre de décision est plus simple et interprétable, tandis que la Random Forest est plus complexe mais souvent plus précise.
- Overfitting : L'arbre de décision a tendance à l'overfitting, tandis que la Random Forest le réduit.
- Adaptabilité : L'arbre de décision peut être adapté facilement, tandis que la Random Forest nécessite plus de ressources et de paramétrages.

Merci !