Régression Bayésienne

Idriss FELLOUSSI

12 octobre 2023

Plan de la Présentation

Méthode : Régression Bayésienne

- La régression bayésienne est un type de régression linéaire
- Elle est basée sur les statistiques bayesiennes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

- Son but est de trouver la meilleure estimation des paramètres d'un modèle linéaire qui décrit la relation entre les variables indépendantes et dépendantes

Exemple

Pour illustrer la régression bayésienne, prenons comme exemple le dataset "diabetes" de Scikit-Learn, qui contient des données sur le diabète de patients. L'objectif est de prédire une mesure de la progression de la maladie en fonction de plusieurs caractéristiques médicales.

- Charger le dataset "diabetes".
- Diviser les données en ensembles d'entraînement et de test.
- Appliquer la régression bayésienne pour modéliser la relation entre les caractéristiques et la progression de la maladie.
- Évaluer les performances du modèle.

Format des données

Les données du problème peuvent être représentées comme suit :

X : Matrice de conception de forme (n, p)

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

y : Vecteur cible de forme $(n, 1)$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

θ : Vecteur des coefficients de régression de forme $(p, 1)$

Prétraitement des données

1. Gestion des données manquantes
2. Normalisation/Standardisation
3. Gestion des valeurs aberrantes
4. Encodage des variables catégoriques
5. Séparation des ensembles de données
6. Examen de la corrélation
7. Gestion des variables inutiles
8. Traitement des déséquilibres

Hypothèse

Hypothèse de modèle bayésien : θ est une variable aléatoire.

$$P(\theta) = \mathcal{N}(0, \sigma^2 I)$$

où $(0, \sigma^2 I)$ est une distribution gaussienne multivariée centrée autour de zéro avec une variance σ^2 pour chaque coefficient θ_i, et I est la matrice identité.

Fonction de coût

La régression bayésienne, y compris le modèle Bayesian Ridge, peut être formulée de la manière suivante :

1. Distribution a priori sur les coefficients (θ) :

$$P(\theta)$$

2. Likelihood (Vraisemblance) :

$$P(Y|X,\theta)$$

3. Distribution a posteriori sur les coefficients (θ) :

$$P(\theta|X,Y) = \frac{P(Y|X,\theta)P(\theta)}{P(Y|X)}$$

4. Prédiction : Pour faire une prédiction pour de nouvelles données, utilisez la distribution a posteriori sur les coefficients pour estimer la distribution des valeurs cibles prédites.

Ici, X représente les caractéristiques, Y représente les valeurs cibles, et θ représente les coefficients du modèle.

Résultats Expérimentaux

- Présentation des résultats obtenus

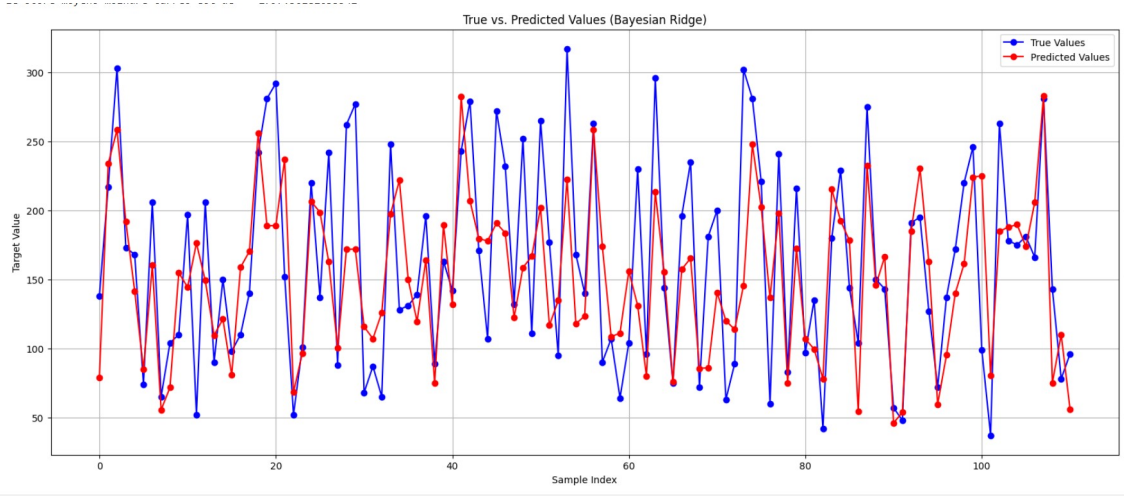

Figure – résultats d'apprentissage

Questions ?

Merci pour votre attention. Des questions ?