Cài đặt giải thuật hồi quy Ridge (Ridge Regression)

Họ và tên: Vũ Công Luật

MSSV: 20142745

1, Chuẩn hóa dữ liệu

Với mỗi thuộc tính của dữ liệu (mỗi cột của ma trận X) chuẩn hóa theo phương pháp ***StandardScaler*** như sau:

x_norm=(x-mean)/(standard_deviation)

mean: kì vọng

standart_deviation: độ lệch chuẩn

Cài đặt: sử dụng lớp **StandardScaler()**của thư viện sklearn để chuẩn hóa:

class sklearn.preprocessing.**StandardScaler**(*copy=True*, *with_mean=True*, *with_std=True*)

2. Tìm λ cho mô hình

Sử dụng phương pháp ***Cross-Validation:***

Thử $\lambda \in S=(0, 100, 0.1)$ (Tập S gồm những số thực từ 0 đến 100, mỗi số cách nhau 0.1)

Với mỗi λ_i :

K-fold cross validation:

(i) Chia tập training T thành K tập tách biệt có kích thước bằng nhau:

Giả sử $T = (T_1, T_2, \ldots, T_K)$

Thường chọn K = 5 hoặc K = 10

(ii) Với mỗi k = 1, 2, . . . ,K, train mô hình với tham số λ_i trên tập tập dữ liệu $\{ T/T_k \}$ (loại T_k ra khỏi tập tranning)

(ii) Phán đoán đầu ra cho tập T_k bằng mô hình vừa được training: $f_{(\lambda i)k}(z)$

(iv) Tính lỗi trên tập T_k:

$$(\text{CV Error})_{(\lambda i)k} = |T_k|^{-1} \sum_{(z,y)\in T\_k} (y - f_{(\lambda i)k}(z))^2$$

Tìm $\lambda^* = \lambda_i \in S$ sao cho $(\text{CV Error})_{(\lambda i)k}$ nhỏ nhất

Nếu cần thiết tiếp tục chia nhỏ tập S rồi lặp lại K-fold cross validation:

để tìm được λ tốt hơn nữa

VD: lần đầu tiên S=(0, 100, 0.1) tìm được λ bằng 0.90. Chia lại tập S=(0.85, 0.95, 0.001)

Và tìm được λ*=0.86499

Dùng λ^* training lại mô hình trên toàn bộ tập training

Cài đặt: Sử dụng lớp RidgeCV() của thư viện sk-learn để tìm λ^* và training mô hình:

class sklearn.linear_model.**RidgeCV**(*alphas=(0.1*, *1.0*, *10.0)*, *fit_intercept=Tru*

e, *normalize=False*, *scoring=None*, *cv=None*, *gcv_mode=None*, *store_cv_values=False*)

3. Phán đoán đầu ra cho tập test

Đọc tập X_test từ test dataset. Đọc X_train từ train dataset ghép lại thành tập X
Chuẩn hóa dữ liệu cho tập X.
Mục đích gộp cả phần input của cả tập train và test lại rồi cùng chuẩn hóa để tránh trường hợp tập test có ít dữ liệu quá dẫn đến việc chuẩn hóa sai lệch quá nhiều.
X_norm_test = Tách phần input của test đã chuẩn hóa từ X
Sử dụng phương thức predict(X_norm_test) của lớp RidgeCV() để phán đoán đầu ra.