Regresia Liniară. Regresia Ridge

1. Regresia Liniară

Dorim să găsim o funcție *g* astfel încât:

$$y_{hat} = g(X) = \Sigma_{i=1}^{i=n} x_i w_i + b$$

și care interpolează cel mai bine o mulțime de exemple $(X_1, y_1), (X_2, y_2), .., (X_n, y_n)$. Pentru a găsi această funcție, vom minimiza valoarea funcției **M**ean **S**quared **E**rror pe mulțimea de antrenare.

$$MSE(y, y_{hat}) = \Sigma_{i=1}^{i=n} (y_{hat_i} - y_i)^2$$

2. Regresia Ridge

Regresia Ridge adaugă o nouă ”penalizare” funcției de cost, pe lângă faptul că diferența între etichetele *ground-truth* și etichetele *prezise* trebuie să fie minimă, dorim ca ponderile pe care le învățăm să fie mici. Pentru a forța ponderile să fie mici, vom adaugă la funcția de cost norma L_2 a ponderilor.

$$cost_{Ridge}(y, y_{hat}) = \Sigma_{i=1}^{i=n} (y_{hat_i} - y_i)^2 + \alpha \left|\left|W\right|\right|_2$$

Parametrul α controlează cât de mici să fie ponderile.

3. Regresia Lasso

Regresia Lasso adaugă norma L_1 a ponderilor la funcția de cost, creând o reprezentare *sparse* a ponderilor.

$$cost_{Lasso}(y, y_{hat}) = \Sigma_{i=1}^{i=n} (y_{hat_i} - y_i)^2 + \alpha \left|\left|W\right|\right|_1$$

În acest laborator vom folosi modelele implementate în biblioteca Scikit-Learn.

```
from sklearn.linear_model import LinearRegression, Ridge, Lasso
# definirea modelelor
linear_regression_model = LinearRegression()
ridge_regression_model = Ridge(alpha=1)
lasso_regression_model = Lasso(alpha=1)

# calcularea valorii MSE și MAE
from sklearn.metrics import mean_squared_error, mean_absolute_error
mse_value = mean_squared_error(y_true, y_pred)
mae_value = mean_absolute_error(y_true, y_pred)
```

Car Price Prediction

În continuare, vom lucra pe baza de date ***Car Price Prediction*** pentru a prezice prețul unei mașinii în funcție de caracteristicile ei.

Această bază de date este formată din 4879 exemple de antrenare. Neavând o mulțime separată de testare vom folosi tehnica de validare încrucișată (*cross-validation*) pentru a valida parametrii modelelor pe care le vom antrena.

În figura de mai jos, vedem 4 exemple din mulțime de antrenare.

Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Price
2010	72000	CNG	Manual	First	26.6 km/kg	998 CC	58.16 bhp	5	1.75
2012	87000	Diesel	Manual	First	20.77 kmpl	1248 CC	88.76 bhp	7	6
2013	40670	Diesel	Automatic	Second	15.2 kmpl	1968 CC	140.8 bhp	5	17.74
2012	75000	LPG	Manual	First	21.1 km/kg	814 CC	55.2 bhp	5	2.35

După procesarea datelor (extragerea datelor din CVS și salvarea lor în format .npy) atributele au fost rearanjate în felul următor:

1. anul fabricației
2. numărul de kilometrii
3. mileage
4. motor
5. putere
6. numărul de locuri
7. numărul de proprietarii (valori între 1 și 4)

8-12. tipul de combustibil - fiind 5 tipuri de combustibil, acesta a fost recodat într-un one-hot vector de 5 componente.

13-14. tipul de transmisie - fiind 2 tipuri de transmisie, acesta a fost recodat într-un one-hot vector de 2 componente. 10 - „Manual”; 01 - ”Automatic”.

Descărcați arhiva care conține datele de antrenare de aici.

Codul următor ne ajută să citim datele de antrenare:

```
import numpy as np
from sklearn.utils import shuffle

# load training data
training_data = np.load('data/training_data.npy')
prices = np.load('data/prices.npy')
# print the first 4 samples
print('The first 4 samples are:\n ', training_data[:4])
print('The first 4 prices are:\n ', prices[:4])
# shuffle
training_data, prices = shuffle(training_data, prices, random_state=0)
```

Exerciții

1. Definiți o metodă care primește doi parametrii, datele de antrenare și cele de testare și returnează datele normalizate. Folosiți o metodă de normalizare ***corespunzătoare*** pentru setul de date ***Car Price Prediction***.

2. Folosind mulțimea de antrenare din setul de date ***Car Price Prediction*** antrenați un *model de regresie liniară* folosind validarea încrucișată cu 3 fold-uri. Calculați valoarea medie a funcțiilor *MSE* și *MAE*.
   Nu uitați să normalizați datele folosind metoda definită anterior.

3. Folosind mulțimea de antrenare din setul de date ***Car Price Prediction*** antrenați un *model de regresie ridge* folosind validarea încrucișată cu 3 fold-uri. Calculați valoarea medie a funcțiilor *MSE* și *MAE*. Verificați care valoare a lui α, $\alpha \in \{1, 10, 100, 1000\}$ obține o performanță mai bună.
   Nu uitați să normalizați datele folosind metoda definită anterior.

4. Folosind cel mai performant *alpha* de la punctul anterior, antrenați un *model de regresie ridge* pe întreaga mulțime de antrenare, afișați coeficienți și bias-ul regresiei. Care este *cel mai semnificativ* atribut? Care este al doilea *cel mai semnificativ atribut*? Care este ***cel mai puțin semnificativ*** atribut?
   Nu uitați să normalizați datele folosind metoda definită anterior.