**Universitatea “DUNĂREA DE JOS” Galați**

Facultatea de Automatică, Calculatoare, Inginerie Electrica și Electronică,

Specializarea Calculatoare

**Tema 2**

**Lasso Regression**

S.l.dr.inf. Student:

Crăciun Viorel Marian Abeaboeru Alexandru Cristian

Cuprins

[Introducere 3](#_Toc153143652)

[Rezultate și analiză 5](#_Toc153143653)

[Concluzie 6](#_Toc153143654)

[Bibliografie: 7](#_Toc153143655)

# Introducere

Regresia

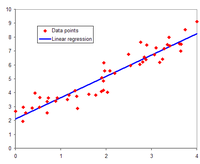
În domeniul statisticilor, regresia reprezintă un ansamblu de tehnici folosite pentru a evalua și înțelege relațiile dintre o variabilă dependentă și una sau mai multe variabile independente. Una dintre formele cele mai utilizate în acest domeniu este regresia liniară, care încearcă să găsească o linie sau o relație liniară între variabilele care se potrivește cel mai bine datelor în conformitate cu anumite criterii matematice, cum ar fi metoda celor mai mici pătrate.

Această metodă ajută cercetătorii să estimeze valoarea așteptată a variabilei dependente în funcție de valorile luate de variabilele independente. Există și alte forme de regresie mai puțin obișnuite, care explorează variantele neliniare sau estimează diferite caracteristici ale datelor, precum regresia cuantile sau Analiza Condiției Necesare. [1]

Scopurile principale ale analizei de regresie sunt:

1.Predicție și Prognoză: Utilizarea regresiei pentru a prezice sau estima valori viitoare, un domeniu în care există o suprapunere semnificativă cu tehnologiile din învățarea automată.

2. Inferență Cauzală: În unele contexte, regresia poate fi folosită pentru a deduce legături de cauzalitate între variabilele independente și cele dependente. Este important să se sublinieze că regresiile oferă relații între variabile într-un set specific de date; pentru a utiliza aceste relații pentru predicții sau pentru a stabili cauzalitatea, cercetătorii trebuie să argumenteze și să justifice temeinic valabilitatea acestor relații în contextul dorit. Acest aspect devine crucial atunci când se încearcă estimarea relațiilor de cauzalitate folosind date observaționale.



Linia de regresie pentru 50 de puncte aleatoare într-o distribuție gaussiană în jurul liniei y=1.5x+2

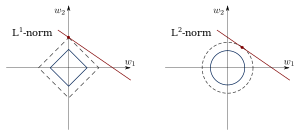
Există mai multe tipuri de regresie care oferă abordări diferite în analiza datelor, cum ar fi regresia liniară, regresia logistică pentru clasificare binară și regresia polinomială. Alte tehnici precum Ridge și Lasso Regression folosesc regularizarea pentru a gestiona overfitting-ul, în timp ce Bayesian Linear Regression adoptă o perspectivă probabilistică în estimarea parametrilor modelului. Aceste variante diversificate de regresie oferă instrumente puternice pentru a explora și înțelege relațiile dintre variabile într-o gamă largă de domenii și aplicații.

Regresia Lasso

Regresia Lasso reprezintă un tip de regresie liniară ce utilizează o tehnică numită reducere (shrinkage). Această metodă constă în ajustarea valorilor datelor spre un punct central, precum media. Regresia Lasso promovează modele simple și cu mai puțini parametri. Este deosebit de utilă pentru modelele care prezintă niveluri ridicate de multicolinearitate sau atunci când se dorește automatizarea selecției modelelor, inclusiv a eliminării variabilelor/parametrilor neesențiali.

Acronimul "LASSO" provine de la denumirea în limba engleză, "Least Absolute Shrinkage and Selection Operator" (Operatorul de Selectare și Reducere a Celor Mai Mici Valori Absolute).

Regresia Lasso utilizează regularizare de tip L1, adăugând o penalizare echivalentă cu valoarea absolută a coeficienților. Acest tip de regularizare poate conduce la modele cu puțini coeficienți, unde anumiți coeficienți pot deveni zero și pot fi eliminați din model. Penalizările mai mari determină coeficienții să fie mai aproape de zero, contribuind astfel la crearea unor modele mai simple. În contrast, regularizarea L2 (cum ar fi în regresia Ridge) nu elimină coeficienții sau nu conduce la modele cu puțini coeficienți. Această caracteristică face regresia Lasso mai ușor de interpretat în comparație cu regresia Ridge. [2]



Prin ajustarea parametrului de regularizare în regresia Lasso, se poate controla complexitatea modelului. Un parametru mai mare de regularizare duce la mai mulți coeficienți zero și, implicit, la un model mai simplu.

Deși regresia Lasso are multiple avantaje, există și anumite limite. De exemplu, atunci când variabilele sunt puternic corelate, Lasso are tendința de a selecta una dintre ele și a ignora celelalte, ceea ce poate conduce la pierderea informațiilor importante.

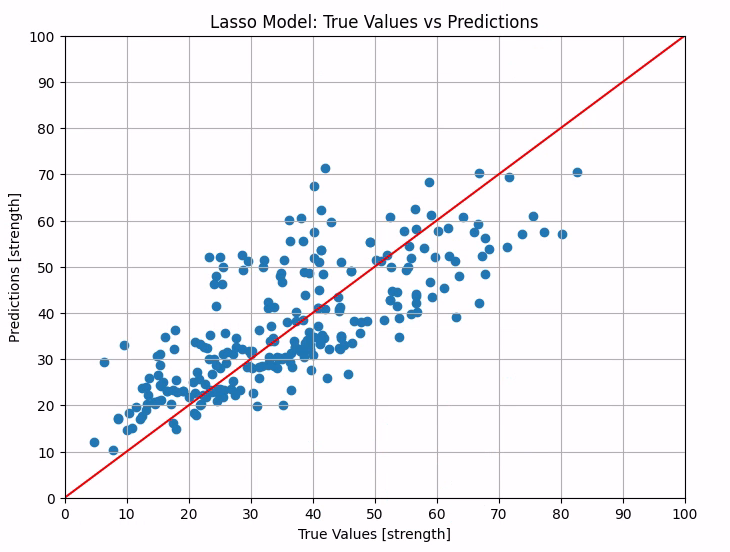
# Rezultate și analiză

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alpha | Lasso coefficients for | | | | | | | | Score | MAE |
| Cement | Slag | Ash | Water | Superplastic | Coarseagg | Fineagg | Age |
| 4.0 | 4.3138 | 0 | 0 | 0 | 2.1087 | 0 | 0 | 1.7515 | 0.2927 | 11.132 |
| 3.0 | 5.3036 | 0.5823 | 0 | 0 | 3.2223 | 0 | 0 | 2.8817 | 0.3749 | 10.532 |
| 2.0 | 6.5559 | 2.1304 | 0 | -1.1410 | 3.5539 | 0 | 0 | 4.1963 | 0.4709 | 9.7326 |
| 1.0 | 7.7721 | 3.6922 | 0.2465 | -2.7834 | 3.5684 | 0 | -0.4931 | 5.5579 | 0.5305 | 9.0611 |
| 0.1 | 11.3132 | 7.3257 | 4.2034 | -4.2440 | 1.6040 | -0.0445 | 0 | 6.9851 | 0.5559 | 8.6624 |
| 0.01 | 12.0521 | 8.0673 | 4.9210 | -4.0518 | 1.4335 | 0.2248 | 0.4981 | 7.1429 | 0.5558 | 8.6423 |
| 0.001 | 12.2989 | 8.3173 | 5.1344 | -3.8595 | 1.4574 | 0.4191 | 0.7365 | 7.1639 | 0.5567 | 8.6243 |
| 0.001&  w/o Water | 15.0458 | 10.9709 | 7.3347 | - | 2.8294 | 3.1336 | 3.8544 | 7.0758 | 0.5587 | 8.5012 |

Scorul pe setul de test este o măsură a performanței modelului tău pe datele pe care nu le-a văzut în timpul antrenării, adică pe setul de date de testare. Este o metrică care indică cât de bine se generalizează modelul tău la datele noi, necunoscute, pe baza a ceea ce a învățat din setul de date de antrenament.

Mean Absolute Error (MAE) este o măsură a erorii medii absolute între predicțiile modelului și valorile reale din setul de date de testare. Pentru fiecare punct de date, MAE reprezintă diferența absolută dintre predicția modelului și valoarea reală, iar apoi calculează media acestor diferențe absolute pentru a oferi o singură măsură a erorii.

Coeficienții Lasso sunt valorile asociate fiecărei caracteristici (sau predictor) în modelul de regresie Lasso, care indică contribuția relativă a fiecărei caracteristici la predicția variabilei țintă. Regresia Lasso este un tip de regresie liniară care aplică regularizare L1 (Lasso) asupra coeficienților pentru a forța unele dintre ele să devină zero, efectuând implicit selecție de caracteristici.



# Concluzie

În urma analizei rezultatelor pe setul de date, observăm că utilizarea unui alpha mai mic în regresia Lasso a condus la performanțe mai bune în general. Alegerea unui alpha mai mic a avut următoarele efecte:

Reducerea erorilor de predicție: Modelul antrenat cu un alpha mai mic a înregistrat erori de predicție mai scăzute, exprimata prin Mean Absolute Error (MAE). Aceste rezultate indică faptul că predicțiile modelului Lasso cu un alpha redus s-au apropiat mai mult de valorile reale din setul de date de testare.

Coeficienți non-zero pentru mai multe caracteristici: Un alpha mai mic în regresia Lasso a determinat păstrarea unui număr mai mare de coeficienți non-zero. Aceasta sugerează că regularizarea mai mică a permis modelului să considere mai multe caracteristici drept semnificative pentru predicție.

În plus, s-a observat că coeficientul Lasso asociat coloanei "water" a fost negativ, ceea ce a indicat o relație negativă între această caracteristică și variabila țintă. Din considerente de îmbunătățire a performanței modelului și a reducerii erorilor de predicție, am ales să ignorăm această caracteristică în procesul de testare. Această decizie a contribuit la obținerea unui scor mai bun și a unui MAE mai mic, întrucât eliminarea acestei caracteristici a condus la o îmbunătățire a generalizării modelului.

# Bibliografie:

1. <https://en.wikipedia.org/wiki/Regression_analysis>
2. <https://www.statisticshowto.com/lasso-regression/>