Journal of Engineering and Basic Sciences, 2024, 01, 1422252
Published Online 25.01.2024 (https://dergipark.org.tr/en/pub/joebs)

Bağlantısallık Probleminin Cezalı Regresyon Yöntemleri İle Giderilmesi

Emel Ciger* **Evrim Ersin Kangal***

Mersin Üniversitesi, Sosyal Bilimler Enstitüsü, Mersin, Türkiye

ÖZET

Gelişen teknoloji ile yapay zekâ uygulamalarına olan ilgi artmış ve bu uygulamalar kurumların, akademik çalışmaların ilgi odağı olmuştur. Makine öğrenmesinde karar ağaçları ve yapay sinir ağları (artificial neural network) algoritmaları sıkça kullanılan yöntemler olsa da araştırma yapılan çalışmanın amacı veya kullanılan veri setlerine uygunluklarından dolayı regresyon modelleri de hala en çok kullanılan yöntemlerdendir. Ancak bazı regresyon modellerinde "Çoklu Doğrusal Bağlantı Problemi" olarak adlandırılan, bağımsız değişkenlerden iki veya daha fazlası arasında doğrusal ya da doğrusala yakın ilişki olması durumu ortaya çıkabilmektedir. Çoklu doğrusal bağlantı problemi(multicollinearity) ile karşılaşılan durumlarda Lasso Regresyon'u ve Ridge Regresyon'u gibi alternatif yöntemler ele alınabilir. Bu çalışmada Kaggle veri bankasında açık kaynak olarak sunulan öğrencilerin not performanslarının olduğu 1000 kayıttan oluşan bir veri seti kullanılmıştır. Veri setine, Python 3.8.5 yazılım dili kullanılarak sırasıyla Lineer Regresyon, Lasso Regresyon ve Ridge Regresyon makine öğrenmesi modelleri uygulanmıştır. Sonuç olarak, bu çalışmada cezalı regresyon yöntemlerinin denetimli makine öğrenmesine etkisi bir örnek üzerinde denenmiş ve sonuçları tartışılmıştır. Sistem üzerinde ayrı ayrı uygulanan modellerin performans değerleri; Lineer Regresyonda "0,839", Lasso Regresyonda "0,843" ve Ridge Regresyonda "0,846" olarak gerçekleşmiştir.

Anahtar Kelimeler: Lasso, Ridge, Lineer Regresyon, Makine Öğrenmesi

Eliminating The Connectivity Problem With Penalized Regressıon Methods

ABSTRACT

With the developing technology, interest in artificial intelligence applications has increased and has become the center of attention of institutions and academic studies. Although decision trees and artificial neural network algorithms are frequently used methods in machine learning, regression models are still among the most commonly used methods due to their suitability for the purpose of the study or the data sets used. However, in some regression models, there may be a linear or near-linear relationship between two or more of the independent variables, which is called the "Multicollinearity Problem". In cases where multicollinearity is encountered, alternative methods such as Lasso Regression and Ridge Regression can be considered. This thesis uses a dataset of 1000 records of students' grade performance, which is available as open source in the Kaggle database. Linear Regression, Lasso Regression and Ridge Regression machine learning models are applied to the dataset using Python 3.8.5 software language. As a result, in this study, the effect of penalized regression methods on supervised machine learning is tested on an example and the results are discussed. The performance values of the models applied separately on the system were realized as "0.839" in Linear Regression, "0.843" in Lasso Regression and "0.846" in Ridge Regression.

Key Words: Lasso, Ridge, Linear Regression, Machine Learning

1. GİRİŞ

Regresyon analizi istatistiksel bir öğrenme metodu olup, bağımsız değişken ile bağımlı değişken arasındaki ilişkileri anlamak ve bir değişkenin diğerine bağlılığını ölçmek bu yöntemin odak noktasını oluşturmaktadır. Algoritmanın kazandığı deneyim algoritmaya hedef değişkenin gelecekteki değerini tahmin etmesine olanak sunacaktır. Modelin kazandığı deneyim ya bir doğru ya da bir eğri ile temsil edilmektedir. Örneğin; basit lineer regresyonda bağımlı değişken ile bağımsız değişken arasındaki ilişki bir doğru denklemi ile ifade edilirken, çoklu regresyon durumunda bağımsız değişkenlerin sayısı 2'den fazla olacağından doğrudan basit doğru denklemi ile ifade edilmesi imkânsızdır. Bu noktada model geçmiş verilerden kazanılan deneyim doğrultusunda bağımsız değişkenlerin katsayılarının alacağı değeri belirlemektedir. Elde edilen bu katsayılar algoritmanın performansını değerlendirmekten sorumlu olan R2 hakkında bilgi vermektedir. R2 değeri, 0 ile 1 arasında yer almaktadır. Sıfır durumu başarısızlığı, bir ise mükemmel başarıyı temsil etmektedir. Ekonometri, finans, biyoloji, mühendislik ve sosyal bilimler gibi birçok

* eciger@gmail.com * evrimersin@gmail.com

alanda regresyon yaklaşımı ile ilgili birçok çalışma görmek mümkündür.

Çoklu doğrusallık probleminin beraberinde getirdiği bağlantısallık sorunu ilk olarak Frish tarafından gün yüzüne çıkarılmıştır [1]. Sonrasında, Hoerl ve Kennard bu sorunun üstesinden gelmek adına Ridge Regresyon Yaklaşımı olarak bilinen çözüm perspektifi önermiştir[2, 3] Bu modelleme temel olarak veri kümesindeki hangi özniteliğin karar mekanizmasında etkin rol üstlendiğini belirlemek esasına dayanmaktadır ve sonuç itibari ile karar mekanizması model içerisinde yer alan ayar parametresi yardımı ile belirlenmektedir. Öte yandan, gelişen teknoloji ile birlikte yapay zekâ uygulamalarına olan dikkat önemli ölçüde artmış ve böylece kurumlar ve akademik çalışmaların ilgi odağı haline gelmiştir. Her ne kadar Makine Öğrenmesi bağlamında Karar Ağaçları ve Yapay Sinir ağları algoritmaları sıkça kullanılan yöntemler olarak dikkat çekse de üzerinde çalışılan konunun amacı veya kullanılan veri setlerinin karakteristik özellikleri bazı durumlarda regresyon modellerini de başarılı çıkarımlara ulaştırın yaklaşımlar ailesine katmaktadır. Ancak bazı regresyon modellerinde "Çoklu Doğrusal Bağlantı Problemi" olarak adlandırılan, bağımsız değişkenlerden iki veya daha fazlası arasında doğrusal ya da doğrusala yakın ilişki olması durumu ortaya çıkabilmektedir. Çoklu doğrusal bağlantı problemi (ya da terminolojide bilinen adıyla multi-collinearity) ile karşılaşılan durumlarda Lasso Regresyon Yaklaşımı ve Ridge Regresyon Analizi gibi alternatif yöntemler kullanılabilmektedir. Bu çalışmasının araştırma konusu, temel olarak Kaggle veri bankasında açık kaynak olarak sunulan öğrenci performans notlarından oluşan bir veri setinin Lasso ve Ridge Regresyon modelleri aracılığı ile analiz edilmesine dayanmaktadır.

2. KAYNAK ARAŞTIRMALARI

Doğrusal regresyon analizindeki karmaşıklık çeşitli nedenlerden dolayı ortaya çıkabilir. Verinin özellikleri her zaman aynı değildir veya analizin amacı her zaman aynı olmayabilir. Tüm bu konulara bağlı olarak doğrusal regresyon yöntemlerine ilişkin istatistiksel araştırmalar on dokuzuncu yüzyılda başlamıştır ve halen oldukça aktiftir[4]. Doğrusal Regresyonun tercih edilme sebepleri; doğrusal form nedeniyle model parametreleri kolaylıkla yorumlanabilir, doğrusal model teorileri matematiksel olarak iyi bir şekilde oluşturulmuştur ve en önemlisi birçok modern modelleme aracının yapı taşıdır. Özellikle örneklem boyutu küçük olduğunda temeldeki regresyon fonksiyonuna tatmin edici bir performans gösterir[5]. Regresyon analizi üç şeyi mümkün kılan bir istatistiksel değerlendirme türüdür: Bağımlı değişkenler ile bağımsız değişkenler arasındaki ilişkiler, regresyon analizi yoluyla istatistiksel olarak tanımlanabilir. Bağımlı değişkenlerin değerleri, bağımsız değişkenlerin gözlemlenen değerlerinden tahmin edilebilir. Böylece sonucu etkileyen risk faktörleri kolaylıkla belirlenebilir [6]. Tahmine dayalı regresyonlar yaygın olarak kullanılmaktadır çünkü öngörülebilirlik yıllardır öncelikli hedef olmuştur. Ancak tahmine dayalı regresyonlardaki en önemli sorun gürültü sorunudur. Ne kadar çok değişken ile çalışılırsa algoritmalar daha az tahmin yeteneği göstermektedir. Bu nedenle değişken seçimleri çok önemlidir. Yaklaşık olarak 30 yılda yapılan araştırmalar ele alındığında bu gürültüyü en aza indirmek için kullanılan tahmine dayalı başlıca yöntemlerden biri de mutlak küçültme ve seçme operatörü olan Lasso [7] ve RIDGE [8] öne çıkmaktadır. Lasso değişken seçimi tutarlılığında sahip bir yöntem olduğu için birçok çalışmada tercih edilmiştir [8]. Ridge ve genelleştirilmiş ters tahmincilerin avantajlarından biri, hesaplamaların kolaylığıdır. Bu yöntemlerde X'X ve X'Y matrisleri bir kez hesaplanır ve ardından ölçeklendirilerek korelasyon matrisi oluşturulur. Sırt regresyonu için, genellikle (X'X + kI)'nın tersinin alınması gibi, her bir λ (lambda) değeri için bir kez yapılacak basit inversiyon işlemleri, sırt izinin nerede stabilize olduğunu belirlemek için yeterlidir [9]. Ridge ve Lasso'nun etkinliği, kullanılan veri setinin özelliklerine ve hedefine bağlıdır. Literatürde yapılan pek çok çalışma, bu yöntemlerin özellikle yüksek boyutlu (high-dimensional) veri setlerinde ve çoklu doğrusal bağlantı sorunuyla karşılaşıldığında faydalı olduğunu göstermektedir [1, 10-14].

3. MATERYAL VE METOT

Bu çalışmanın amacı; makine öğrenmesi kullanılarak Lineer regresyon, Lasso regresyon ve Ridge regresyon modellerinin uygulamaları yapılarak analizlerinin çıkarılması, sonuçlarının incelenerek karşılaştırılma yapılmasıdır. Çalışmada; makine öğrenmesi ve analizler için Python 3.7 yazılım programı ve derleyicileri kullanılmıştır. Kullanılan veri seti kamuya açık olup, içerisinde veri bilimi çalışmaları yapmak için 50.000'den fazla veri kümesi bulunan www.kaggle.com sitesinden alınmıştır. Kullanılan veri seti; Amerika'da lise öğrencilerinin çeşitli derslerde kazandıkları notlardan oluşmaktadır. Bu veri; 1.000 satır ve 8 değişkenden ("gender (Cinsiyet)", "race/ethnicity (Irk/Etnik)", "parental level of education (Ebeveyn eğitim düzeyi)", "lunch (Öğle yemeği)", "test preparation course" (Sınava hazırlık kursu)", "math score (Matematik puanı), "reading score (Okuma puanı)", "writing score (Yazma puanı)" oluşan, lise öğrencilerinin bazı derslerden aldıkları notlar ve bu notları etkileyen çeşitli kişisel, sosyal ve ekonomik faktörleri içeren bir veri setidir.

4. BULGULAR

Araştırmada 518 kadın, 482 erkek öğrenci üzerinde çalışılmıştır. Bu öğrencilerin; 89'u Grup A, 190'ı Grup B, 319'u Grup C, 262'si Grup D ve 140'ı Grup E milletindendir. Öğrencilerin ebeveynlerinin; 118'i lisans, 226'sı kolej, 59'u lisans, 222'si ön lisans, 196'sı lise, 179'u lise dengi okul mezunudur. Araştırmadaki öğrencilerin; 645'i standart öğle yemeği yerken, 355'i ücretsiz/azaltılmış öğle yemeği yemektedir ve 358 öğrenci kurstaki test hazırlıklarını tamamlamışken 642'si test hazırlığı yapmamıştır. Veri setinde hedef değişken "Matematik Puanı", diğer değişkenler ise "Matematik Puanı"nı bulmak için kullanacak bağımsız değişkenlerdir. Veri setinin %80'i sistemin eğitilmesi için, kalan %20'si ise test için kullanılacaktır. Makine öğrenmesi ile sistem eğitilerek denetimli öğrenme algoritmalarından Lineer Regresyon, Lasso ve Ridge Regresyonları sırasıyla uygulanarak sistem performansı ölçülmüştür. Alınan sonuçlar Tablo 1.'de gösterilmiştir.

Tablo 1. Çalışmada uygulanan modellerin performans tablosu.

	Eğitim Seti Üzerinde R-kare	**Test Seti Üzerinde R-kare**	**Ortalama Karesel Hata**	**Optimum Alpha**
Lineer Regresyon	0,872	0,839	31,6	
Lasso Regresyon	0,87	0,843	30,85	0.188
Ridge Regresyon	0,868	0,846	30,25	30

Tablo 1'de Lineer Regresyon dan sonra uygulanan ceza alpha uygulamaları ile çalışan Ridge ve Lasso regresyonları sonucunda, test verileri üzerindeki doğruluk oranında az da olsa bir artış sağlandığı ve ortalama karesel hata da bir azalma olduğu görülmektedir. Ayrıca Tablo 2.'de Lasso ve Ridge Regresyon modellerinde bağımsız değişkenlerin sisteme katkıları listelenmiştir. Her iki modele göre de "Cinsiyet"in en önemli değişken olduğu görülmektedir. Ayrıca Lasso Regresyon modeline göre "Ebeveyn_Eğitimi" değişkeninin modele hiçbir etkisi olmadığı saptanmıştır. Ancak Ridge regresyonda yapısı gereği modelin değişken katsayısını sıfırlama potansiyeli olmadığından "Ebeveyn_Eğitimi"nin sisteme 0,011 oranında bir katkısı gözlemlenmiştir.

Tablo 2. Lasso ve Ridge modellerinde değişkenlerin modele katkıları

Değişken (Özellik)	Lasso Regresyon Katsayıları	Ridge Regresyon Katsayıları	Değişim
Cinsiyet	12,639	11,534	-9%
Irk	0,786	0,931	18%
Ebeveyn_Egitim	0,000	0,011	-
Ogle_Yemegi	2,981	3,361	13%
Kursta_Test_Hazirligi	2,033	2,313	14%
Okuma_Notu	0,390	0,400	2%
Yazma_Notu	0,560	0,540	-4%

Tablo 3. Lasso ve Ridge değişken katsayıları arasındaki yüzdelik değişim

Değişken (Özellik)	Lasso Regresyon Katsayıları	Ridge Regresyon Katsayıları
Cinsiyet	12,639	11,534
Irk	0,786	0,931
Ebeveyn_Egitim	0,000	0,011
Ogle_Yemegi	2,981	3,361
Kursta_Test_Hazirligi	2,033	2,313
Okuma_Notu	0,390	0,400
Yazma_Notu	0,560	0,540

5. SONUÇ

Tablo 1.'den elde edilen veriler doğrultusunda eğitim sırasında Lasso yaklaşımı Ridge yaklaşımına göre daha baskınken test noktasında Ridge yaklaşımı daha etkili olmaya başlamaktadır. Diğer taraftan hata payı açısından bakıldığında Lasso yaklaşımı Ridge yaklaşımına göre daha yüksek hata oranı içermektedir. Fakat Ridge yaklaşımının ayar parametresinin optimizasyonu hata minimizasyon yöntemi ile belirlenirken Lasso yaklaşımı için R2 parametresi ile optimize edilmiştir. Ridge yaklaşımında hata payı üzerinden gidilmesindeki temel neden, ceza teriminin kare şeklinde hata fonksiyonunda yer almasıdır. Bu durum katsayıların baskılanmasını engellemekte ve buna paralel olarak ayar parametresinin yüksek değerlere ulaşmasına neden olmaktadır. Diğer taftan Lasso yaklaşımında ise hata fonksiyonunda ayar parametresi birinci dereceden bağımlı olduğu için katsayıların birbirlerinin baskılanmasına neden olmaktadır. Sonuç olarak bu veri seti için bakıldığında Lasso yaklaşımının Ridgeye göre daha uygun olduğu açık olarak görülmektedir. Bunun temel nedeni ise Alpha ayar parametresinin Tablo 1.'den anlaşılacağı üzere daha düşük değerde anlamlı R2 değerine ulaştığı açık olarak görülmektedir. Sonuç olarak, Ridge ve Lasso regresyon modellerinin benzer performans gösterdiği ancak belirli durumlarda bir modelin diğerine göre daha iyi sonuçlar elde ettiği gözlemlenmiştir. Model seçiminin, uygulamanın gereksinimlerine ve veri setinin özelliklerine bağlı olarak değerlendirilmesi gerektiği bu çalışmada net olarak görülmektedir.

Tablo 3.'de değişken katsayılarının Ridge regresyon sonuçlarının, Lasso regresyon sonuçlarına göre yüzdelik değişimi hesaplanmıştır. Bu tabloda görüldüğü gibi "Irk", "Ogle_Yemegi", "Kursta_Test_Hazirligi" ve "Okuma_Notu" değişkenlerinin modele etkisini gösteren katsayılarda pozitif yönde yüzdelik bir değişim gözlemlernmiş yani bu değişkenlerin modele katkısı Ridge Regresyonu'nda daha fazla olduğu hesaplanmıştır. Bunun tersine, "Cinsiyet" ve "Yazma_Notu"'nda negatif yönde bir değişim meydana gelmiş ve bu değişkenlerin modele katkısı Lasso Regresyonunda daha fazla olmuştur. "Ebeveyn_Egitim" değişkenin modele katkısı ise Lasso Regresyonunda sıfırlandığından Lasso ve Ridge arasındaki yüzdelik değişim hesaplanamamıştır. Ancak bu değişkenin çok az da olsa Ridge Regresyonda bir katkısı hesaplanmıştır.

KAYNAKLAR

[1] **Frisch, R., Statistical confluence analysis by means of complete regression systems. (No Title). (1934)**

[2] **Hoerl, A. E. and R. W. Kennard, Ridge regression: applications to nonorthogonal problems. Technometrics. 12, (1970) 69-82.**

[3] **Hoerl, A. E. and R. W. Kennard, Ridge regression: Biased estimation for**

nonorthogonal problems. Technometrics. 12, (1970) 55-67.

[4] Feigelson, E. D. and G. J. Babu, Linear regression in astronomy. II. Astrophysical Journal, Part 1 (ISSN 0004-637X), vol. 397, no. 1, p. 55-67. 397, (1992) 55-67.

[5] Su, X., X. Yan and C. L. Tsai, Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics. 4, (2012) 275-294.

[6] Schneider, A., G. Hommel and M. Blettner, Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Ärzteblatt International. 107, (2010) 776.

[7] Tibshirani, R., Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology. 58, (1996) 267-288.

[8] Hoerl, A. E., R. W. Kannard and K. F. Baldwin, Ridge regression: some simulations. Communications in Statistics-Theory and Methods. 4, (1975) 105-123.

[9] Marquardt, D. W. and R. D. Snee, Ridge regression in practice. The American Statistician. 29, (1975) 3-20.

[10] McDonald, G. C. and D. I. Galarneau, A Monte Carlo evaluation of some ridge-type estimators. Journal of the American Statistical Association. 70, (1975) 407-416.

[11] Hocking, R. R., A Biometrics invited paper. The analysis and selection of variables in linear regression. Biometrics. (1976) 1-49.

[12] JF, L. and W. P, A simulation study of ridge and other regression estimators. Communications in Statistics-Theory and Methods. 5, (1976) 307-323.

[13] Pasha, G. and M. Shah, Application of ridge regression to multicollinear data. Journal of research (Science). 15, (2004) 97-106.

[14] Dorugade, A. and D. Kashid, Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences. 4, (2010) 447-456.