3. Metot - Kullanılan Makine Öğrenme Algoritmaları

Bu bölümde makine öğrenme yöntemlerinden gözetimli öğrenim metotlarının regresyon algoritmalarını inceleyeceğiz. Regresyon, bir değişkenin bir veya daha fazla değişkenle arasındaki bağlantının matematiksel bir fonksiyonla gösterilmesidir[1]. Yani bağımlı(y), bağımsız(x) değişkenleri düşünüldüğünde y'nin, x'in bir fonksiyonu olarak ifade edilen ilişki biçimi regresyon olarak tanımlanır. Bu durum değişkenler arasında bir sebep-sonuç bağlantısını özetlemektedir. Matematiksel fonksiyon durumuna göre doğrusal ve doğrusal olmayan regresyon çeşitleri bulunmaktadır[2]. Bu çalışma da fonların fiyatlarının nasıl tahmin edileceği sorusu bağlamında değerlendirildiğinde, veri setindeki bağımlı değişken olarak fon fiyatlarının olması ve birden fazla bağımsız değişken değerin bulunması kapsamında regresyon algoritmalarının kullanılması daha uygun bulunmuştur. Regresyon metotlarından doğrusal çoklu regresyon algoritmalarından Kısmi En Küçük Kareler ve Ridge regresyon ile doğrusal olmayan çoklu regresyon algoritmalarından ise Destek Vektör ve Yapay Sinir Ağları kullanılacaktır.

3.1 Doğrusal Çoklu Regresyon

Birden fazla bağımsız değişken ile bir bağımlı değişken arasındaki doğrusal bağlantıyı incelemektedir. Doğrusal çoklu regresyonun matematiksel olarak gösterimi ise;

$$Y_i = (b_0 + b_1 X_1 + b_2 X_2 + \ldots. b_n X_n) + e_i$$

olarak ifade edilir.

3.1.1 Kısmi En Küçük Kareler Regresyonu (PLSR)

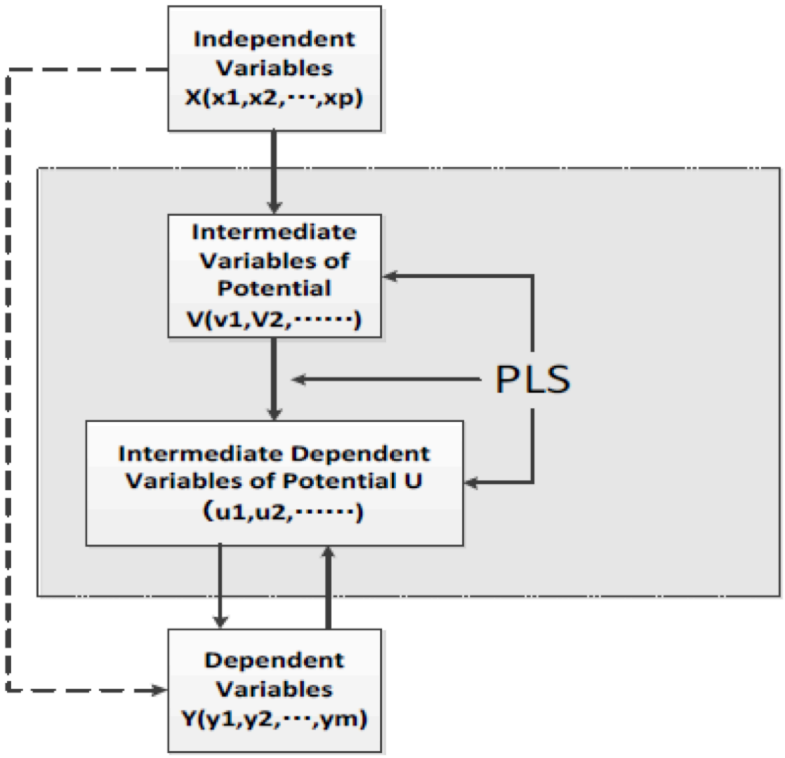
3.1.1.1 Teori

PLS, 1960'larda Herman Wold tarafından bir ekonometri tekniği olarak geliştirilmiş ve temel bileşen analizi ile çoklu regresyon özelliklerini genelleştiren ve birleştiren yeni bir teknik olmaktadır. PLS yöntemi, bağımsız(X) değişken grubunun çok fazla olması ve bir dizi bağımlı(Y) değişken tahmin etmemiz gerektiğinde kullanılmaktadır. Teknik olarak ise, X ve Y değişkenleri arasındaki gizli bağlantıdan yararlanarak, veri blokları arasındaki ilişkiyi bulmayı amaçlamaktadır [3]. Ayrıca PLS çalışmaları ilkin, ekonomik ve sosyal durumları modelleme için kullanılmıştır. Fakat yaygın olarak kullanımına oğlu Svante Wold tarafından kemometrik alanındaki çalışmalarla başlanmıştır [4]. Kemometri, istatistik ve matematik tekniklerinin kullanılarak, kimyasal

sistemlerden bilgi alma bilimidir. Kimyasal analizlerde, veriden doğru bilginin ya da gizli bilginin açığa çıkarılmasına imkan tanıyan bir araçtır[5]. İstatistik alanındaki ilk kullanımına ise 1988 yılında Höskulddson [6] ve 1989 yılındaki Naes ve Marten [7] tarafından yapılan çalışmalar örnek verilebilir.

3.1.1.2 Matematiksel Model ve Algoritma

Bu kısımda PLS yöntemin temel matematiksel diyagramından bahsedilecektir. PLS için kullanacağım algoritma ise deneysel çalışma bölümünde fonların fiyat tahmin için geliştirilecek olan model algoritması olarak kullanılacaktır.



Şekil 1. PLS Matematiksel Diyagramı[8]

PLS matematiksel diyagram şemasından anlaşılacağı üzere, PLS genel olarak potansiyel bağımsız değişkenler ile potansiyel bağımlı değişkenler arasındaki doğrusal ilişki kurmayı amaçlamaktadır. M tane bağımlı değişkenler Y (y1, y2, y3,…, ym) ile p tane bağımsız değişkenler X (x1, x2, x3,…, xp) arasındaki ilişkiyi ise dolaylı olarak yansıtır. Potansiyel bağımsız değişkenler ve potansiyel

bağımlı değişkenler, PLS regresyonundaki değişkenlerin doğrusal kombinasyonunu yansıtır ve iki varsayımı mevcuttur:

1. İki potansiyel değişken grubu, bağımsız değişken veya bağımlı değişkenler mutasyon bilgilerini taşımaktadır.

2. Potansiyel değişkenler arasındaki korelasyon değeri maksimize edilmektedir.[8]

Bağımlı Değişken Y ve Bağımsız Değişken X için PLS regresyon matematiksel gösterim şeması şu şekildedir;

$$\mathbf{X} = \sum_{j=1}^{A} \mathbf{t}_j \mathbf{p}_j' + \mathbf{E} \text{ ve } \mathbf{Y} = \sum_{j=1}^{A} \mathbf{u}_j \mathbf{q}_j' + \mathbf{F}$$

Şekil 2. Bağımlı ve Bağımsız Denklem[9]

Yukarıdaki matematiksel gösterimde $\mathbf{t_j}$, $\mathbf{u_j}$ ler gizli değişkenler olup, $\mathbf{t_j}$ değişkenleri birbirlerine ve aynı bağlamda sonraki $\mathbf{u_j}$ değişken değerine diktir. PLS model için maksimum sayıda gizli değişken sayısı, bağımsız ve bağımlı değişkenler arasındaki kovaryans değeri maksimum olacak şekilde elde edilir.[9]

Matematiksel model anlatımından sonra bu çalışmada kullanılacak PLS algoritması klasik ve standart olan NIPALS (Non-Linear Iterative Partial Least Squares; Doğrusal olmayan yinelemeli kısmi en küçük kareler) algoritması olacaktır. NIPALS bağımsız değişken değerleri birden fazla olan ve bağımlı değişken değeri tekil olan veri setleri için güçlü ve sağlam bir algoritma olduğundan dolayı tercih edilmiştir. Temel de tüm PLS algoritmalarının amacı kovaryans matrislerini en fazla sayıya ulaştıracak bileşenlerin elde edilmesidir.[10]

NIPALS algoritmasının adımları;

j= 1,2,…J, bileşen sayısını gösterir.
$\mathbf{X_1} = \mathbf{X}$, $\mathbf{Y_1} = \mathbf{Y}$, orijinal matrisler.

1. Veri setimizde bağımlı değişken sayısı tek olduğu için direkt **Y** değişken sütunu $\mathbf{u_j}$ vektörü olarak tanımlanır.

2. **X** ve **Y** bileşenlerinin $\mathbf{u_j}$ üzerinedeki regresyonu **X** ve **u** arasındaki kovaryans değerini en çoklayan **w** ağırlık vektörü $w_j = X'_j u / (u'_j u)$ ile elde edilir.

3. $\mathbf{w_j} / ||\mathbf{w_j}||$ ve $\mathbf{w_j}$ vektörü normuna bölünüp vektör boyu 1 buluncak şekilde ölçeklendirilir.

4. $t_j = X_j w_j$ eşitliği **X'**in bileşeni $\mathbf{t_j}$, $\mathbf{w_j}$ ise ağırlık vektörü ile **X** 'in değerinin bir kombinasyonu olacak formda hesaplanır.

5. $\mathbf{t_j}$ bileşen değerinin **Y** 'yi modelleme değerini açıklayan $\mathbf{c_j}$ ağırlık vektörü ise $c_j = Y'_j t_j / (t'_j t_j)$ ile **Y** 'nin $\mathbf{t_j}$ üzerindeki regresyon değeri bulunur.

6. $\mathbf{c_j}$ vektörü norm değerine bölünerek boyu 1 bulunacak şekilde ölçeklendirilir. $\mathbf{c_j} / ||\mathbf{c_j}||$ şeklinde hesaplanır.

7. **Y** değerinin ilgili bileşeni $\mathbf{u_{j(yeni)}}$, $\mathbf{c_j}$ ağırlık vektörü ve **Y** 'nin doğrusal kombinasyonu ise $\mathbf{Y_j\ c_j / (\ c'_j c_j)}$ şeklinde hesaplanır.

8. İkinci adımda kullanılan $\mathbf{u_j}$ değeri ile yedinci adımda kullanılan $\mathbf{u_{j(yeni)}}$ değerileri arasında bir benzerliğin olup olmadığına bakılır. Bu benzerlik, iki vektörün fark normu $\mathbf{10^{-6}}$ gibi çok düşük bir değer olması sağlanır. Bu benzerlik oranı sağlanır ise bir sonraki adıma geçilerek algoritma sonlandırılır, yoksa yedinci adımda bulunan $\mathbf{u_{j(yeni)}}$ değeri ikinci adımdaki yerine yazılarak algoritmaya devam edilir.

9. **X** bileşeni $\mathbf{t_j}$ üzerine regresyonu, bileşen değerinin açıklayıcı değişken üzerindeki etki faktörünü gösteren yük vektörü $\mathbf{p_j}$, $X'_j t_j / (t'_j t_j)$ ile elde edilir.

10. **Y** bileşeni $\mathbf{u_j}$ üzerine regresyonu, bileşen değerinin bağımlı değişken üzerindeki etki faktörünü gösteren yük vektörü $\mathbf{q_j}$, $Y'_j u_j / (u'_j u_j)$ ile elde edilir.

11. Hem **X** hem de **Y** için bileşen değerleri ayrı ayrı hesaplandığında değerler arasında zayıf bir ilişki bulunduğu görülmektedir. Bu zayıf değer ilişkisinin kaldırılması için her bir bileşen değeri için **Y** 'nin bileşeni $\mathbf{u_j}$ 'nın **X** 'in bileşeni $\mathbf{t_j}$ üzerine regresyon değeri bulunan iç bir $\mathbf{b_a}$ katsayısı $b_a = u'_j t_j / (t'_j t_j)$ şeklinde hesaplanır.

12. Bulunan bileşen değerleri, yüklerin bağımlı değerleri ve açıklayıcı değişkenleri modellemenin yapılmasında kullanılır. Bileşenler açıklayıcı ve bağımlı değişken $\boldsymbol{X = TP'}$ ve $\boldsymbol{Y = BTC'}$ şeklinde modellenir. Bu aşamdan sonra algoritmanın bir sonraki bileşenlerini elde etmek için kullanılacak olan $\mathbf{X_{j+1}}$ ve $\mathbf{Y_{j+1}}$ artık matrisleri $\boldsymbol{X_{j+1} \rightarrow X_j - t_j p'_j}$ ve $\boldsymbol{Y_{j+1} \rightarrow Y_j - b t_j c'_j}$ şekilde hesaplanmaktadır.[11,12]

NIPALS algoritmasının 12 adımda olaşan çalışma akışına göre açıklayıcı ve bağımlı değişkenlerdeki değişimin büyük bir kısmı açıklık elde edilene kadar devam etmektedir. Yani X bileşenindeki değer matrisinin sıfır matrisi oluncaya kadar algoritma çalışır. Ayrıca, algoritma hesaplama sürecinde, ihtiyaç duyulacak en az sayıda bileşen değerini vermektedir.[11]

Kaynakça

1. Orhunbilge, N. (2002), Uygulamalı Regresyon ve Korelasyon Analizi, İstanbul, İ.Ü. İşletme Fakültesi.

2. GÖK, M. (2017). MAKİNE ÖĞRENMESİ YÖNTEMLERİ İLE AKADEMİK BAŞARININ TAHMİN EDİLMESİ. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım ve Teknoloji. 5(3): 148-139.

3. Abdi, H. (2003). Partial Least Squares (PLS) Regression, Lewis-Beck M., Bryman, A., Futing T. (Eds.) Encyclopedia of Social Sciences Research Methods. Thousand Oaks (CA): Sage.

4. Tobias, R.D. (1995). An Introduction to Partial Least Squares Regression. UGI Proceedings, Orlando 2-5 April, pp. 1-8.

5. Polat, E., Günay, S. (2009). Kısmi En Küçük Kareler ve Bir Uygulama. Ondokuz Mayıs Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, VI. İSTATİSTİK GÜNLERİ SEMPOZYUMU BİLDİRİLER KİTABI. S,438.

6. Höskuldson, A. (1988). PLS Regression Methods. Journal of Chemometrics, 2: 211-228.

7. Marten, H. ve Naes, T. (1989). Multivariate Calibration. John Wiley & Sons.

8. Wu, J. (2015). Research on Several Problems in Partial Least Squares Regression Analysis. The Open Electrical & Electronic Engineering Journal, 8, 754-758

9. Bulut, E., Alma, G., Ö. (2011). Kısmi En Küçük Kareler Regresyonu Yardımıyla Optimum Bileşen Sayısını Seçmede Model Seçme Kriterlerinin Performans Karşılaştırılması. İstanbul Üniversitesi, İktisat Fakültesi, Ekonometri ve İstatistik Dergisi, Sayı:15 , 38-52.

10. Rosipal, R., Trejo, L. J. (2001). Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. Journal of Machine Learning Research 2 , 97-123.

11. Bulut, E., Alın, A. (2009). Kısmi En Küçük Kareler Regresyon Yöntemi Algoritmalarından Nipals ve PLS - Kernel Algoritmalarının Karşılaştırılması ve Bir Uygulama. Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Dergisi, Cilt:24, Sayı:2, ss.127-138.

12. Ümit, Ö.A., Bulut, E. (2013). TÜRKİYE'DE İŞSİZLİĞİ ETKİLEYEN FAKTÖRLERİN KISMİ EN KÜÇÜK KARELER REGRESYON YÖNTEMİ İLE ANALİZİ: 2005-2010 DÖNEMİ. Dumlupınar Üniversitesi Sosyal Bilimler Dergisi. (37): -.