TOBB ETU

8 Temmuz 2023

Son Teslim Tarihi:

18 Temmuz 2023 23:59

BİL 470/570 Ödev 2

AMAÇ:

Bu ödev üç aşamalıdır.

(1) EDA: "Gender-Height-Weight-Body Mass Index" veri kümesi üzerinde keşifsel veri analizi (exploratory data analysis - EDA) gerçekleştirilmesi
(2) LİNEER REGRESYON MODELİ: Herhangi bir kütüphaneden yararlanmadan **lineer regresyon modeli** implemente etmeniz beklenmektedir. Eğitim veri seti üzerinden modeli eğitmeniz ve test veri setindeki verilerin tahmini Body Mass Index değerinin hesaplanması.
(3) SONUÇLAR: Test veri seti için oluşturulan tahminlerden yararlanılarak modelin başarımının hesaplanması.

1 GÖREVLER

1.1 EDA (Explanatory Data Analysis)

https://www.kaggle.com/datasets/yersever/500-person-gender-height-weight-bodymassindex üzerinden veri setini indiriniz. İlk olarak sizden beklenen; veri setinde bulunan Gender sütununu kaldırmanız çünkü tahminlerinizi boy (Height) ve Kilo (Weight) değerlerine göre yapacaksınız.

Daha sonra, bu veri kümesinde bir EDA gerçekleştireceksiniz. EDA, veri seti özetini ve özniteliklerin (features) her birinin dağılımını içermelidir. Bunun yanında verilerin boy-kilo 2 boyutlu uzayındaki dağılımlarını göstermeniz beklenmektedir. İsteyen veri seti hakkında daha fazla ayrıntı verebilir.

Sonuçlarınızı bu bölümde göstermek için **pandas**, **numpy**, **seaborn**, **matplotlib** gibi kütüphanelerden yaralanabilirsiniz. İsterseniz ek kütüphaneler eklemekten çekinmeyin.

1.2 Lineer Regresyon Modeli

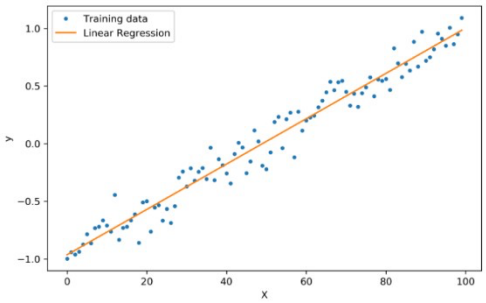
Derste öğrendiğiniz Lineer Regresyon modelini **herhangi bir kütüphane kullanmadan** implemente ediniz. Aşağıda modelin implementasyonu için gerekli açıklamalar ve kısa bir konu özeti yer almaktadır.

Konu Özeti

Genel Bakış:

Lineer Regresyon modeli ile temel amacımız veriler arasındaki lineer ilişkiyi en iyi şekilde yansıtan ve veri dağılımına en iyi şekilde fit edecek "lineer" doğruyu oluşturmaktır.
Eğer elimizde sadece bir öznitelik ve ona bağlı değişen bir target değer varsa, eğitim aşaması sonucunda aşağıda verilen Grafik1'deki gibi veri dağılımına fit etmiş doğru çıktısı oluşacaktır.

Grafik 1: Örnek Dağılım Grafiği



Grafik 1'de y ekseni target değerlerini, x ekseni ise Target'ın bağımlı olduğu özniteliği temsil etmektedir. Eğitim verilerinin dağılımından x ve y arasında lineere yakın bir ilişki olduğu görülmektedir. Lineer regresyon modeli ile bu ilişkiyi en iyi yansıtan doğru oluşturulmuştur.

Modelin eğitim aşamasından sonra, test verilerinin x değerleri, oluşturulan doğru üzerine yerleştirilerek karşılık geldiği y değeri belirlenir. Kısaca, modelin oluşturduğu doğru yardımıyla, test verilerinin x değerleri için olası y değerleri tahmin edilir. Tahmin edilen değerler ile test verilerinin orijinal y değerleri (Target değerler) karşılaştırılarak çizilen doğrunun başarımı hesaplanır.

En İyi Lineer Doğruyu Oluşturmak:

Yukarıda belirtilen bir değişken ve bir targettan oluşan ilişki,

$$Y = mx + b \quad \textbf{(d1)}$$

denklemi ile ifade edilir. Veriler arasındaki ilişkiyi en iyi şekilde yansıtacak doğruyu oluşturmak için, her x değerini Target değerine götürecek m ve n değerlerini belirlemek gerekmektedir. Lineer Regresyon modeli ile eğitim aşamasında, başlangıçta belirlenen bir m ve b değeri kullanılarak tüm eğitim verilerinin x değerlerinin karşılık geldiği tahmini y değerleri hesaplanır (1 epoch). Belirlenen *Loss fonksiyonu* ile tahmin sonuçlarıyla target değerler arasındaki fark hesaplanır. Sonrasında bu farkı en aza indirgeyecek şekilde m ve b değeri güncellenir. Güncel m ve b değerleri kullanılarak eğitim verilerinin x değerleri güncel tahminler oluşturulur. Bu güncelleme işlemi, modelin olması gereken target değerlere yakın tahminler üretmeye başlamasıyla sonlanır.

Model Eğitimi ile ideal m ve n değerlerini bulmak için;

1. m ve b için başlangıç değerleri atanır.

2. Loss fonksiyonu tanımlanır. **Loss fonksiyonu** ile her epoch sonunda seçtiğimiz m ve b değerlerinin sonucunda elde edilen Y değerleri ile olması gereken (target) Yt değerleri arasındaki kayıp bulunur. Lineer Regression modelinde Loss fonksiyonu olarak ***mean square error*** fonksiyonu kullanılabilir**.**

$$Loss = \frac{\sum(Y - Yt)^2}{n} \qquad \text{(d2)}$$

$$Loss = \frac{\sum(mx+b - Yt)^2}{n} \qquad \text{(d3)}$$

   d2 ve d3 denklemlerini açıklamak gerekirse; her bir epoch içerisinde, eğitim verisindeki x değerleri için seçilen m ve b değerleri kullanılarak $(Y - Yt)^2$ işlemi yapılır. Epoch sonunda bu işlemlerin sonucu toplanır ve ortalaması alınır.

3. Loss sonucunu minimize edecek m ve b değerlerini belirlemek ve güncellemek için **Gradient Descent** Algoritması kullanılır. Her epoch sonunda, belirlenen tahmini değerleri ile target değerleri arasındaki farkı minimize etmek için Loss fonksiyon sonucununu minimize etmek gerekir. Bunun için denklem de değiştirilebilecek parametreler olan m ve b değerleri üzerinde değişiklik yapılır.

Grafik 2: Örnek Hata-Ağırlık Fonksiyonu

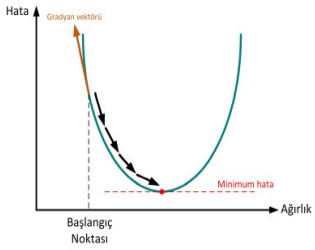


Figure 2: Gradient descent optimization algorithm methodology

d3 denkleminde bulunan ***mean square error*** fonksiyonu incelendiğinde, Loss fonksiyonun bağımlı olduğu değişkenlerin yani modelin ağırlıklarının m ve b olduğu görülmektedir.
(Denklemde yer alan,
Yt: Eğitim Veri setinde belli olan target değerleri,
x: Eğitim veri setindeki belli olan x değişkenini ve
n Eğitim veri setindeki her bir epochda eğitilen veri sayısını temsil etmektedir.)

Bu durumda d3 denklemindeki Yt, x ve n değişkenlerini sabit düşünürsek Loss fonskiyonun m ve b 'ye bağlı parabolik bir fonksiyon olduğu ve Grafik 2 deki fonksiyona benzediği görülmektedir.

Amaç tahmin değerleri ile target değerleri arasındaki farkı minimize edecek ağırlıkları bulmak. Bu durumda, Grafik 2 deki minima noktasını sağlayacak m ve b değerlerini bulunması gerekmektedir. Bu minima noktası ise Loss fonksiyonunun ağırlıklarına göre türevinin 0 olduğu noktadır. Ancak her zaman değişkenlere göre fonksiyonun türevini alıp 0'a eşitleyip uygun minima bulunamaz. Rastgele bir ağırlıktan başlanarak her epoch sonucunda Loss fonksiyonunun m ve b değişkenlerine göre türevi alınır. Belirlenen türevler bir learning rate katsayısı (μ) ile çarpılır ve çıkan sonuç eski m ve b değerlerinden çıkarılarak o epoch için güncel ağırlılar hesaplanır. Bu şekilde Grafik 2'de de gösterildiği gibi her epoch sonucunda adım adım minimaya yaklaşılmış olunur.

$$m = m - \mu * \frac{\partial Loss}{\partial m} \qquad \text{(d4)}$$

$$b = b - \mu * \frac{\partial Loss}{\partial b} \qquad \text{(d5)}$$

Implementasyon

- Problem tanımı 1 target değişken ve 2 bilinmeyen değişkenden oluşacaktır. Amacımız kilo ve boy özniteliklerinin target öznitelik olan BMI ' e olan etkisini gözlemlemektir. Lineer regresyon modelinizin ile oluşturulması beklenen doğru denklemi aşağıda verilmiştir.

$$BIM = m1 * Height + m2 * Weight + b$$

$$Z = m1 * x + m2 * y + b$$

- Loss Fonksiyonu olarak yukarıda belirtilen ***mean square error*** kullanmanız beklenmektedir.
- m1, m2 ve b değerlerinin Loss fonksiyonuna göre türevini alırken aşağıda belirtilen denklemleri kullanabilirsiniz.

$$\frac{\partial Loss}{\partial m1} = \frac{2}{n}\sum(m1 * xi + m2 * yi + b - zi) * xi \qquad \text{(d6)}$$

$$\frac{\partial Loss}{\partial m2} = \frac{2}{n}\sum(m1 * xi + m2 * yi + b - zi) * yi \qquad \text{(d7)}$$

$$\frac{\partial Loss}{\partial b} = \frac{2}{n}\sum(m1 * xi + m2 * yi + b - zi) \qquad \text{(d8)}$$

- İlk değer atarken, m1=1, m2=2, b=0 değerlerini kullanabilirsiniz. Farklı değerler deneyerek çıktınızı gözlemleyebilirsiniz.

- Yukarıda belirtilen sınıflandırıcının imzası aşağıdaki gibi olacaktır:

  Linear Regression (learning_rate=0.000005, epoch=1000)
- Verilerin %50'sini kullanarak modeli eğitin ve kalan verilerle sınıflandırıcıyı test edin.
- Notebook üzerinden çağrılacak fonksiyon yapıları:
  - **fit(x_train, y_train,z_train) –**
    predict(x_test,y_test)

Modeli implemente ederken **herhangi bir kütüphane kullanamazsınız**. Vektörler için list yapısını ve 2D inputlar için list of list kullanabilirsiniz.

1.3 Sonuçlar

Train ve test aşamaları için ayrı ayrı;

- Loss ve Accuracy grafiklerini çizdirin
- Accuracy için hangi ölçütü kullandığınızı açıklayın (Mean Error, Rsquare…)
- Sonuçlar hakkında yorum yapın

2 Gönderme

Toplamda 3 dosya gönderilecek:

1. **Python dosyası (LR.py**):
   Lineer Regresyon modelinin implamentasyonunu içerir.
   Bu dosyada herhangi bir kitaplık kullanamazsınız.
2. **Notebook dosyası (report.ipynb)**:
   3 kısım içerir;

1 Veri setinin keşifsel veri analizi (EDA)
(2) Sınıflandırıcının eğitimi ve
(3) Sonuçların yorumlanması.

Adımları açıklamak için markdown syntax'ını kullanabilir, modeli eğitmek için python kodu yazabilirsiniz,

3. **Rapor (report.pdf):**
   İlgili rapor.ipynb dosyasının PDF dışa aktarımıdır.
   Bu dosya not defteri dosyasıyla aynı içeriğe sahip olmalıdır.
   Bu dosyayı **File > Download as > .pdf** jüpyter notebook menüsünden indirin.

Bireysel bir çalışma olmalıdır. Grup şeklinde yapılmamalıdır. Eğer çalışmanızın orijinalliğinden şüphe edilirse demoya çağrılacaksınız. Beraber yapıldığı veya büyük oranda LLM (Bard, ChatGPT...) kullanımına başvurulduğu anlaşılırsa çalışmadan 0 alınacaktır. Çalışmalarınızın Turnetin üzerinden kontrol edilmesi planlanmaktadır.