TOBB ETU

2 Temmuz 2022

Son Teslim Tarihi:

12 Temmuz 2022 23:59

BİL 470/570 Ödev 2

Amaç:

Bu ödevde "Gender-Height-Weight-Body Mass Index" veri kümesi üzerinde bir keşifsel veri analizi (exploratory data analysis - EDA) gerçekleştireceksiniz. Bunun yanında herhangi bir kütüphaneden yararlanmadan bir **lineer regresyon modeli** implamente edeceksiniz, bu modeli kullanarak bir insanın kilosuna ve boyuna bakarak BMI (Body Mass Index) değerini tahmin edeceksiniz. GÖREVLER

1.1 EDA

https://www.kaggle.com/datasets/yersever/500-person-gender-height-weight-bodymassindex üzerinden veri setini indirebilirsiniz. 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 gerçekleştirin. EDA, veri seti özetinden, özniteliklerin (features) her birinin dağılımından bahsedebilirsiniz. Verilerin boy-kilo 2 boyutlu uzayındaki dağılımlarını gösterin. Bu konuda daha fazla ayrıntı vermek istiyorsanız ek şeyler ekleyin.

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

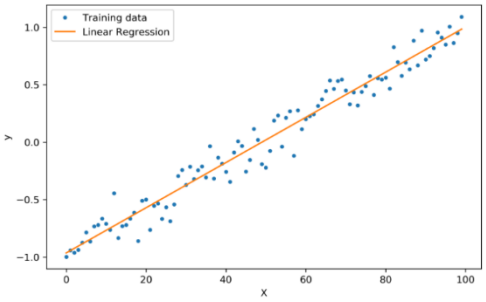
1.2 Lineer Regresyon Modeli

Derste öğrendiğiniz Lineer Regresyon modelini **herhangi bir kütüphane kullanmadan** implamente edin. Aşağıda modelin implamentasyonu için gerekli açıklamalar ve kısa bir konu özeti yer almaktadır.

Konu Tekrarı

Lineer Regresyon modeli ile temel amacımız veriler arasındaki lineer ilişkiyi en iyi şekilde yansıtan ve dağılımlarına en iyi şekilde fit edecek lineer doğruyu oluşturmaktır. Eğer elimizde sadece bir öznitelik ve ona bağlı değişen bir target değer olsaydı, model sonucunda aşağıdaki örnekteki gibi çıktının oluşmasını beklerdik.

Grafik 1: Örnek Dağılım Grafiği



Grafik 1'de y ekseni target değerlerini, x ekseni ise özniteliğimizi temsil eder. Eğitim verilerinin dağılımından x ve y arasında yaklaşık lineer bir ilişki olduğu gözle görünmektedir. Amacımız bu ilişkiyi en iyi yansıtan lineer doğruyu çizerek, test verisinde bulunan herhangi bir x değerinin bu doğru üzerinde yerine konulduğunda karşılık geldiği y değerini bulmak yani bu x değeri için target sonucu tahmin etmektir.

Bu iki değişkenli lineer ilişki ;

$$Y = mx + b \qquad \text{(d1)}$$

Denklemi ile ifade edilir. Benim bu veriler arasındaki ilişkiyi en iyi şekilde yansıtacak m ve n değerlerini bulmam lazım.

İdeal m ve n değerlerini bulmak için yapılan aşamalar:

1. Loss fonksiyonu tanımlanır.

   Bu **loss** fonksiyonunda ile seçtiğimiz m ve n değerlerinin sonucunda elde edilen Y değeri ile olması gereken (target) Yt arasındaki fark bulunur. Bunun için ***mean square error*** fonksiyonu kullanılabilir**.**

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

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

2. İlk başta m ve b için birer değer ataması yapılır.
3. Her bir epoch'da $Loss$ sonucu hesaplanır.

   (Tüm eğitim verisindeki x değerleri için seçilen m ve b değerleri kullanılarak $(Y - Yt)^2$ işlemi yapılır. Bu işlemlerin sonucu toplanır ve ortalaması alınır.)

4. Loss sonucunu minimize edecek m ve b değerleri belirlenir ve değerler güncellenir.

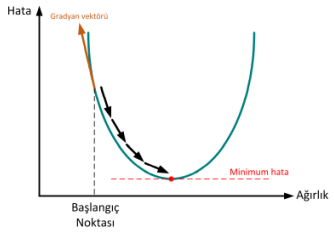


Figure 2: Gradient descent optimization algorithm methodology

Bunun için **Gradient Descent** Algoritması kullanılır.
Verilen x değerleri için Loss fonksiyon sonucunu minimize etmek gerekir. Bunun için denklem de değiştirebileceğim parametreler olan m ve b değerleri üzerinde değişiklik yapılır.

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 çıkarılır.

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

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

Bu işlemin her bir epoch da tekrarlanması, modelin her bir epoch sonucunda fonksiyonun minimasına yaklaşmasına yardımcı olur.

5. Güncellenen m ve b değerleri ile 3 ve 4. Adım belirlenen epoch sayısı kadar tekrar edilir.

Implamentasyon

- Problem tanımı 1 target değişken ve 2 bilinmeyen değişkenden oluşacaktır. Amacımız kilo ve boy özniteliklerinin target öznitelik olan BMI ' e olan etkisini gözlemlemekti. Lineer regresyon modelinizin oluşturmasını beklediği 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 implamente 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

3 dosya göndereceksiniz 2

1. **Python dosyası (LR.py**):
   Lineer Regresyon modelinin implamentasyonunu içeri. **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.

3 Geliştirme Ortamı

Bu derste Python3 kullanılacaktır.

- Önerim bu ders için Anaconda kurulumunu gerçekleştirmeniz.
  https://www.geeksforgeeks.org/how-to-install-anaconda-on-windows/
  Burada detaylı kurulum anlatılmaktadır. Bu şekilde conda environmentlarını kullanabilirsiniz. Bunun yanında Conda python içermektedir.
- Bir IDE seçin: VSCode veya Anaconda Spyder
- Jupyter Notebook indirin. Eğer Anaconda indirdiyseniz bu adımı atlayabilirsiniz.

Bireysel bir çalışma olmalıdır. Grup şeklinde yapılmamalıdır. Eğer çalışmanızın orijinalliğinden şüphe edilirse demoya çağrılacaksınız. Beraber yapıldığı anlaşılırsa çalışmadan 0 alınacaktır.