**Lojistik Regresyon**

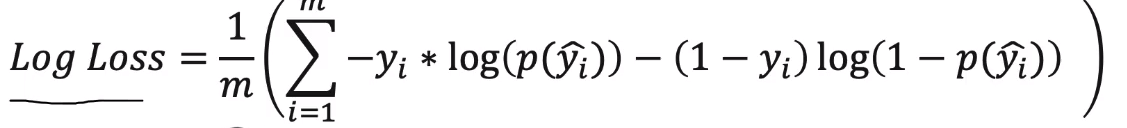
Amaç sınıflandırma problem için bağımlı ve bağımsız değişkenler arasındaki ilişkiyi doğrusal olarak modellemektir.

metin, yazı tipi, beyaz, el yazısı içeren bir resim

Açıklama otomatik olarak oluşturuldu

Yi^ 🡪 1 sınıfına ait olma olasılığını vermekte. (Sigmolit fonksiyonu olarak adlandırılır) 0 ile 1 arasında bir olasılık değeri üretir.

Log Loss 🡪 Gerçek değerler ile tahmin edilen değerler arasındaki farklara ilişkin log loss değerini minimum yapabilecek ağırlıkları bulur.



Example: b = 5, w1 = 4, w2 = -4, w3 = 3, x1 = 2, x2 = 3, x3 = 0

Verilen bias ve weight’lere göre yukarıdaki gözlem birimi için 1 sınıfına ait olma olasılığını hesapalyınız.

Z = 5 + (4\*2) + (-4\*3) + (3\*0) 🡪 1

1 / (1 + e^-z) 🡪 0.731

***Lojistik Regresyon İçin Gradient Descent***

metin, yazı tipi, el yazısı, beyaz içeren bir resim

Açıklama otomatik olarak oluşturuldu

Y^ 🡪 Gerçek değer, h0(x^) 🡪 1 sınıfının gerçekleşme olasılığı.

Entropi 🡪 Entropi ne kadar yüksekse çeşitlilik o kadar fazladır biz entropinin düşük olmasını isteriz.

***Sınıflandırma Problemlerinde Başarı Değerlendirme (Confusion Matrix)***

metin, yazı tipi, ekran görüntüsü, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Accuracy: (TP + TN) / (TP + TN + FP + FN), Doğru sınıflandırma oranı

Precision: TP / (TP + FP), Pozitif sınıf (1) tahminlerinin başarı oranı

Recall: TP / (TP + FN), Pozitif sınıfın (1) doğru tahmin edilme oranıdır

F1 Score: 2\* precision \* recall / (precision + recall)

Not: Dengeli sınıf dağılımına sahipse (40-60, 45-55, 50-50) “Accuracy” kullanabiliriz. Eğer dengeli sınıf dağılımına sahip değilse veri seti, o zaman accuracy yerine “Precision” ve “Recall” yöntemlerini kullanmalıyız.

Example: 1000 kredi kartı işlemi mevcut. 990 normal işlem ve 10 sahtekar işlem mevcut. Buna göre Confusion matrisi doldurunzu ve başarı metriklerini hesaplayınız.

metin, ekran görüntüsü, yazı tipi, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

***Classification Threshold***

CT değerini her zaman 0.5 olarak alırız. Eğer tahmin ettiğimiz değer 0.5’den büyükse 1, 0.5’den küçükse 0 sınıfıdır. Genellikle eşik değer (CT) ne kadar düşük olursa accuracy değerimiz o kadar büyük olur.

Accuracy = Doğru Sınıflandırma Sayısı / Toplam Sınıflandırılan Gözlem Sayısı

metin, sayı, numara, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, sayı, numara, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Table I. Table II.

Table I’de eşik değer 0.50 iken accuracy değerimiz 7/10 🡪 0.7 gelmiştir.

Table II’de eşik değer 0.75 iken accuracy değermiz 5/10 🡪 0.5 gelmiştir.

***ROC Curve (Receiver Operating Characteristic Curve)***

metin, çizgi, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

Threshold değerini 0.1 yaptığımızı varsayalım, tüm tahminlerimizi oluşturduk ve karmaşıklık matrisini oluşturduk daha sonra true positive rate ve false positive rate değerlerini bulduk, daha sonra eşik değerimizi değiştirdik mesela 0.2 yaptık ve tekrardan karmaşıklık matrisin oluşturup yine true positive rate ve false positive rate değerlerini bulduk ve bu işlemleri tüm Classification Threshold (eşik değeri) için tekrarladık. Bir eğri ne kadar fazla yer kaplıyorsa (eğri altında kalan alan integral ile hesaplanır) o model o kadar başarılıdır.

***Area Under Curve (AUC)***

* ROC eğrsinin tek bir sayısal değer ile ifade edilişidir.
* ROC eğrisi altında kalan alandır.
* AUC, tüm olası sınıflandırma eşikleri için toplu bir performans ölçüsüdür.

Not: Sınıflar arası dağılım dengeli de olsa dengesiz de olsa her zaman F1 Score değerine bakılmalıdır.