Makine Öğrenmesiyle Tweet Sınıflandırma

Hazırlayan: Melike Akalan

İÇİNDEKİLER

GİRİŞ

- Projede tweetler lojistik regresyon ve naive bayes algoritmalarıyla positive ve negative o.ü. sınıflandırılmıştır.
- Toplam 10.000 tweetden oluşan 5000 positive, 5000 negative olan veri seti üzerinde çalışılmıştır. Bunlar twitter_samples içerisinde gelen positive_tweets.json ve negative_tweets.json dosyalarıdır.
- Kütüphanelerin import edilmesi,

```
import nltk
import numpy as np
import pandas as pd
import string
from nltk.corpus import stopwords, twitter_samples
from nltk.tokenize import TweetTokenizer
```

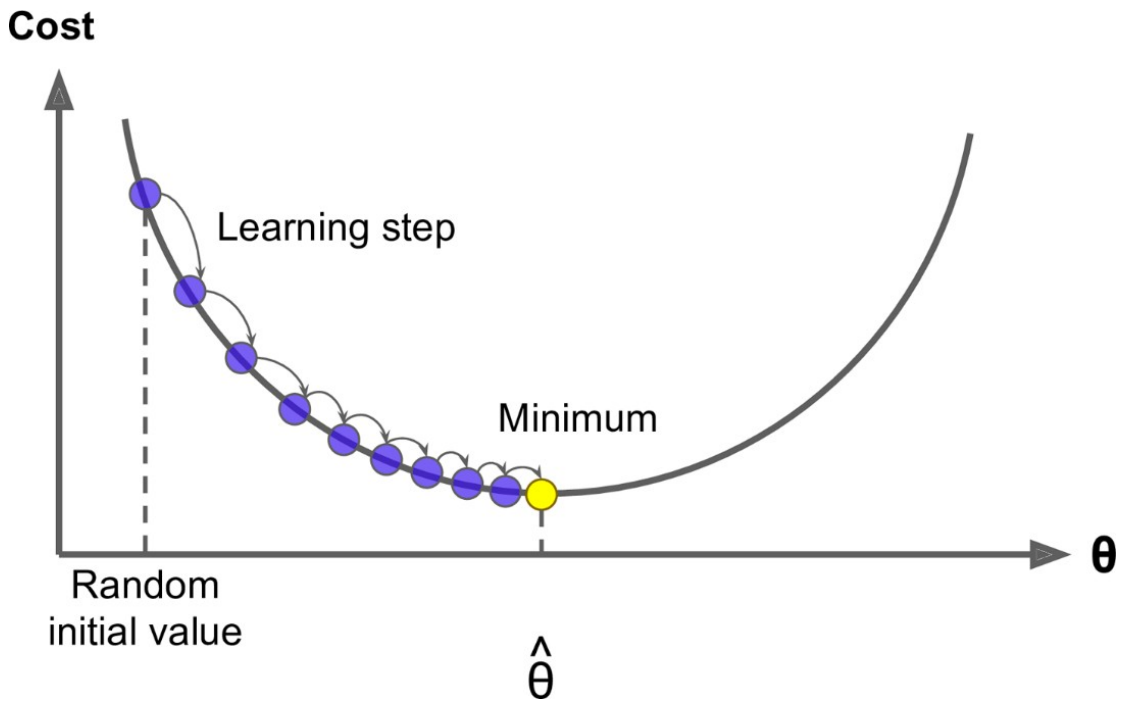
METİN MADENCİLİĞİ

1. Metin Veri Setinin Belirlenmesi: Sınıflandırma algoritmalarını uygulayacağımız veri seti belirlenir.
2. Metin Ön İşleme: Tweetde yer alan gereksiz kelimeler ve semboller temizlenir
3. Metin Dönüşümü: Tweet sayısal özellik vektörüne dönüştürülür.
4. Özellik Seçimi: Tweet için gerekli öznitelikler seçilir.
5. Veri Madenciliği: Tweetin sınıflandırıldığı aşamadır.
6. Değerlendirme: Sınıflandırma sonucu confusion matrix, roc eğrisi, accuracy değerlerine bakılarak model başarısı değerlendirilir.

LOJİSTİK REGRESYON

- Lojistik regresyon; eğitilen modelin ikili sınıflandırma yapmasını sağlayan denetimli makine öğrenme algoritmasıdır.
- Tweetin pozitif/ negatif olma olasılığını hesaplarken sigmoid fonksiyonunu kullanılırız. Sigmoid fonksiyonunun çıktısı 0.5 ten büyükse tweet pozitif, küçükse negatif şeklinde sınıflandırılır.

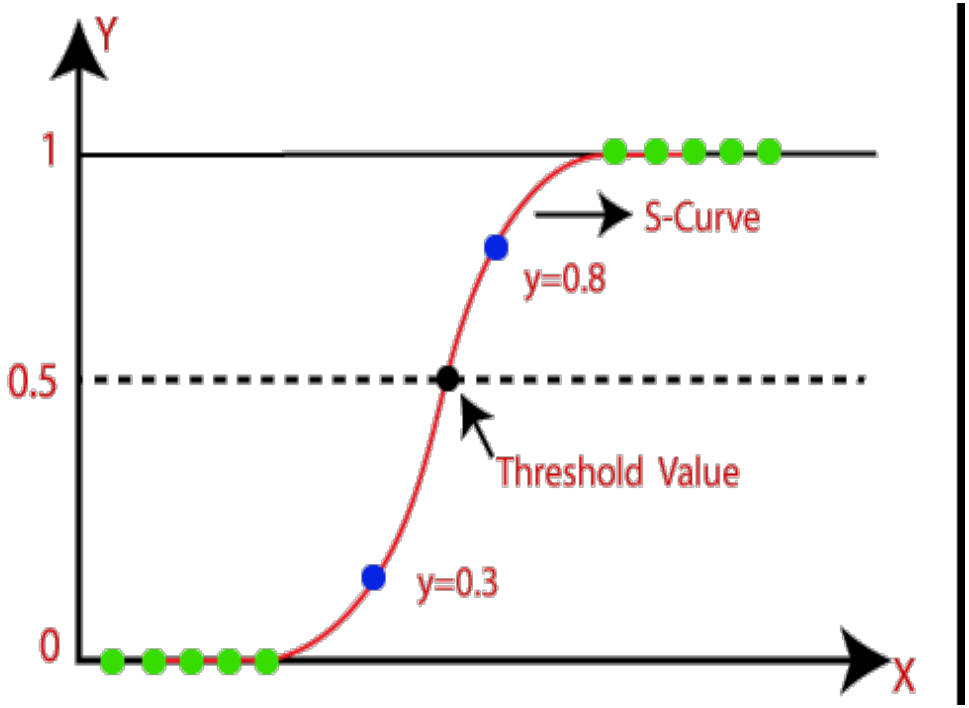
- Lojistik regresyonda modeli eğitirken amacımız doğru theta'ları yani parametreleri güncelleyerek minimum maliyet olacak şekilde veri setine eğitim yaptırmaktır. Theta'lar ilk başta random initialize edilir, her iterasyonda maliyet düşürülerek güncellenir.



SİGMOİD

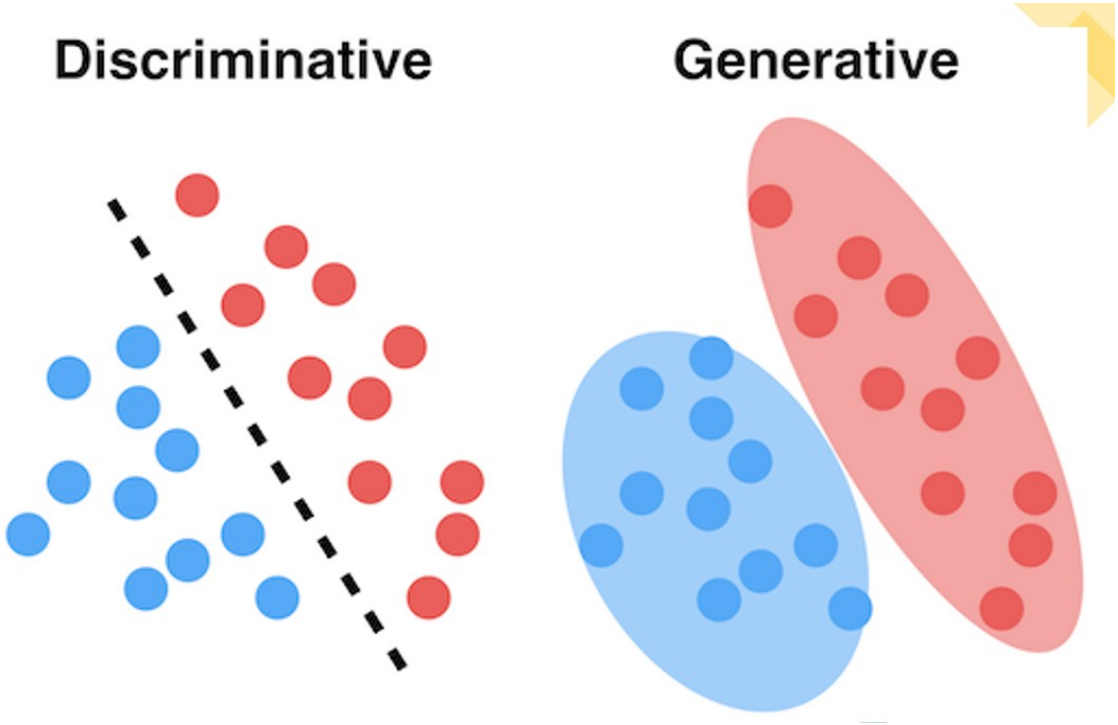
- Makine öğrenmesinde, tahminleri olasılıklara eşlemek için kullanılır.
- S(z) = 0 ile 1 arasında değer alan çıktı (olasılık tahmini)
- z = fonksiyonumuzun girdisi (x),
  e = euler sayısı

$$S(z) = \frac{1}{1 + e^{-z}}$$



NAİVE BAYES

- Naive Bayes; modelin ikili sınıflandırma yapmasını sağlayan denetimli makine öğrenme algoritmasıdır.
- Naive Bayes'i generative (gruplayıcı), lojistik regresyon ise discriminative (ayırıcı) bir algoritmadır.



SONUÇ

- Naive Bayes ve Logistic Regresyon doğrusal sınıflandırıcılardır. Gerçek hayattaki problemler doğrusal olmayabilir, değişkenler(özellikler) birbirine bağımlıdır. Böyle özellikleri seçmek ise zordur, bu da problemin çözümünü zorlaştırır.
- Naive Bayes de Lojistik regresyon gibi; hesapladığı olasılık puanına göre tweeti sınıflandırır.
- Naive Bayes ve Logistic Regression algoritmalarında tokenlaştırma işlemi için **TweetTokenizer()** kullandığımda 99.5 gibi oldukça yüksek accuracy (doğruluk) değerleri vermektedir. **nltk_tokenize()** kullandığımda ise accuracy değerleri 66.5'a düşmektedir. Bunun sebebi TweetTokenizer() emojilerin de olasılığı hesaplarken, nltk_tokenize() emojileri birer noktalama işareti kabul eder ve olasılık hesaplamalarına katılmalarını engeller.

KAYNAKLAR

- https://dergipark.org.tr/tr/pub/ngumuh/issue/35079/383709
- https://www.researchgate.net/publication/338363067_NaiveBayes_Classifier_on_Twitter_Sentiment_Analysis_BPJS_of_HEALTH
- https://dergipark.org.tr/en/pub/humder/issue/56545/772929
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8894084
- https://dergipark.org.tr/tr/pub/estudambilisim/issue/53654/676052
- https://www.youtube.com/c/inzvateam/playlists

teşekkürler.

github linki:
github.com/melikeakalan/Twitter-Sentiment-Analysis