

Bilgisayar Mühendisliđi

Makine Öğrenmesi Proje Ödevi

Atınç Yılmaz

Eren Bekman - 2003022082

Sahte Haber Makine
Öğrenimi



Proje Tanımı :

Elimizde sahte ve gerçek haberlerden oluşan bir veri seti ile çeşitli Makine Öğrenimleri sayesinde bu veri setlerinden sahte ve gerçek haber ayrımı yapabilecek bir makine öğrenmesi projesidir.

Proje Detayı :

- Anaconda arayüzünde Jupiter IDE'sini kullandım
- Proje de python kullandım
- Numpy,Pandas,Seaborn,NLTK,train_test_split,accuracy_score,mean_squared_error kütüphanelerini kullandım

Proje Detayı :

- Proje de 2 tane method kullandım ;
 - LogisticRegression (Lojistik Regresyon) modeli
 - CART modeli
- Sınıflandırma Problemleri kategorisinde Lojistik Regresyon kullandım , Doğrusal olmayan Regresyon Modellerinde 'CART' kullandım

Proje Detayı :

- İki farklı metod kullanmamın sebebi ise , 2 farklı modeli kıyaslamak ve doğruluğunu ölçmek (accuracy score)

Proje İçeriği :

- Projedeki adımlar sırasıyla :

- Veri Setini yükleme
- Veri Setini görüntüleme
- Veri Seti ön işleme
- Veri Manipulasyonu
- Verideki eksik değerleri bulup doldurulması
- İngilizce kelime kütüphanesi ekleyerek kelimelerin düzenlenmesi
- Vectorization işlemi

- Veriyi data ve result olarak ikiye bölme
- Training - Test işlemi
- Lojistik Regresyon modeli kullanımı
- CART modeli kullanımı
- KNN ($n=3$) modeli kullanımı
- Training - Test işlemi
- Üç modelin sonucu

```
[1]: import os
import sys
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import re
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.metrics import mean_squared_error
```

Kütüphanenin import edilmesi :

numpy : Sayısal verilerin işlenmesi için

pandas : Dataframe ve gerekli işlemlerin yapılması

NLTK : İngilizce kelimelerin kullanılması için

PorterStemmer : Kelimelerin düzeltilmesi için

train_test_split : Test ve Train olarak ikiye bölmek için

```
[8]: pip install nltk
```

```
Requirement already satisfied: nltk in c:\users\oguzh\anaconda3\lib\site-packages (3.6.5)
Requirement already satisfied: click in c:\users\oguzh\anaconda3\lib\site-packages (from nltk) (8.0.3)
Requirement already satisfied: joblib in c:\users\oguzh\anaconda3\lib\site-packages (from nltk) (1.1.0)
Requirement already satisfied: regex>=2021.8.3 in c:\users\oguzh\anaconda3\lib\site-packages (from nltk) (2021.8.3)
Requirement already satisfied: tqdm in c:\users\oguzh\anaconda3\lib\site-packages (from nltk) (4.62.3)
Requirement already satisfied: colorama in c:\users\oguzh\anaconda3\lib\site-packages (from click->nltk) (0.4.4)
Note: you may need to restart the kernel to use updated packages.
```

```
[9]: import nltk
from nltk.corpus import stopwords
nltk.download('stopwords')
print(stopwords.words('english')) #English Stopwords
```



```
news_data = pd.read_csv('train.csv')
news_data.head(15)
# ilk 15 degeri gösterdik
# 1 : Fake News, 0 : Real News
```

Verinin görüntülenmesi

id		title	author	
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucas	House Dem Aide: We Didn't Even See C
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life cir
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired C
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Sing
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been s
5	5	Jackie Mason: Hollywood Would Love Trump if He...	Daniel Nussbaum	In these trying times, Jackie Maso
6	6	Life: Life Of Luxury: Elton John's 6 Favorite ...	NaN	Ever wonder how Britain's most icon
7	7	Benoît Hamon Wins French Socialist Party's Pre...	Alissa J. Rubin	PARIS — France chose an idea
8	8	Excerpts From a Draft Script for Donald Trump'...	NaN	Donald J. Trump is scheduled to ma
9	9	A Back-Channel Plan for Ukraine and Russia, Co...	Megan Twohey and Scott Shane	A week before Michael T. Flynn resi
10	10	Obama's Organizing for Action Partners with So...	Aaron Klein	Organizing for Action, the activis

```
In [14]: news_data.info() # veri seti hakkında bilgi edindik
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20800 entries, 0 to 20799
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  -
 0   id          20800 non-null  int64  
 1   title       20242 non-null  object  
 2   author      18843 non-null  object  
 3   text        20761 non-null  object  
 4   label       20800 non-null  int64  
dtypes: int64(2), object(3)
memory usage: 812.6+ KB
```

Veri Seti hakkında bilgi edinildi

```
In [15]: news_data.columns # kullancagımız veri sütunlarını yazdırdık
```

```
Out[15]: Index(['id', 'title', 'author', 'text', 'label'], dtype='object')
```

```
In [18]: news_data.describe().T
```

```
Out[18]:
```

	count	mean	std	min	25%	50%	75%	max
id	20800.0	10399.500000	6004.587135	0.0	5199.75	10399.5	15599.25	20799.0

```
In [19]: print(news_data['label'].value_counts()) # Ayrı ayrı degerlerini aldık
```

```
1    10413  
0    10387  
Name: label, dtype: int64
```

```
In [20]: news_data.isnull().sum() # eksik degerleri saptamamız gerekiyor
```

```
Out[20]: id          0  
title        558  
author      1957  
text         39  
label         0  
dtype: int64
```

```
In [22]: news_data = news_data.fillna('') # Eksik degerleri doldurduk , kontrol edelim ;
```

```
In [23]: news_data.isnull().sum()
```

```
Out[23]: id          0  
title          0  
author         0  
text           0  
label          0  
dtype: int64
```

```
news_data['content'] = news_data['author']+' '+news_data['title'] # haber başlığını ve yazarı
```

```
print(news_data['content']) # şu şekilde gözükecek :
```

```
0      Darrell Lucas House Dem Aide: We Didn't Even S...
1      Daniel J. Flynn FLYNN: Hillary Clinton, Big Wo...
2      Consortiumnews.com Why the Truth Might Get You...
3      Jessica Purkiss 15 Civilians Killed In Single ...
4      Howard Portnoy Iranian woman jailed for fictio...
...
20795   Jerome Hudson Rapper T.I.: Trump a 'Poster Chi...
20796   Benjamin Hoffman N.F.L. Playoffs: Schedule, Ma...
20797   Michael J. de la Merced and Rachel Abrams Macy...
20798   Alex Ansary NATO, Russia To Hold Parallel Exer...
20799   David Swanson What Keeps the F-35 Alive
Name: content, Length: 20800, dtype: object
```

Başlık ve yazarı birleştirdim ve sonucu ayırdım

Şimdi verileri ve sonuçları(label) ayıracağız bu sayede train-test işlemini yapacağız.

```
X = news_data.drop(columns='label', axis=1)
Y = news_data['label']
```

	id	title	author	text	
0	0	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucas	House Dem Aide: We Didn't Even See Comey's Let...	Darrell Lucas House Dem Aide
1	1	FLYNN: Hillary Clinton, Big Woman on Campus - ...	Daniel J. Flynn	Ever get the feeling your life circles the rou...	Daniel J. Flynn FLYNN: Hillary
2	2	Why the Truth Might Get You Fired	Consortiumnews.com	Why the Truth Might Get You Fired October 29, ...	Consortiumnews.com Why the
3	3	15 Civilians Killed In Single US Airstrike Hav...	Jessica Purkiss	Videos 15 Civilians Killed In Single US Aistr...	Jessica Purkiss 15 Civilia
4	4	Iranian woman jailed for fictional unpublished...	Howard Portnoy	Print \nAn Iranian woman has been sentenced to...	Howard Portnoy Iranian wom
...
20795	20795	Rapper T.I.: Trump a 'Poster Child For White S...	Jerome Hudson	Rapper T. I. unloaded on black celebrities who...	Jerome Hudson Rapper T 'F
20796	20796	N.F.L. Playoffs: Schedule, Matchups and Odds -...	Benjamin Hoffman	When the Green Bay Packers lost to the Washing...	Benjamin Hoffman N.F. Sch
20797	20797	Macy's Is Said to Receive Takeover Approach by...	Michael J. de la Merced and Rachel Abrams	The Macy's of today grew from the union of sev...	Michael J. de la Merced Abra
20798	20798	NATO, Russia To Hold Parallel Exercises In Bal...	Alex Ansary	NATO, Russia To Hold Parallel Exercises In Bal...	Alex Ansary NATO, Rus Pa
20799	20799	What Keeps the F-35 Alive	David Swanson	David Swanson is an author, activist, journa...	David Swanson What Kee

```
In [31]: Y # Label'larımız (sonuçlarımız)
```

```
Out[31]: 0      1  
         1      0  
         2      1  
         3      1  
         4      1  
         ..  
        20795    0  
        20796    0  
        20797    0  
        20798    1  
        20799    1  
        Name: label, Length: 20800, dtype: int64
```

Stemming İşlemi Kısacası : Bir kelimeyi köküne indirgemek, ön eki ve son eki kaldırmak.

```
[32]: port_stem = PorterStemmer()
```

```
[33]: def stemming(content):
      stemmed_content = re.sub('[^a-zA-Z]', ' ', content)
      stemmed_content = stemmed_content.lower()
      stemmed_content = stemmed_content.split()
      stemmed_content = [port_stem.stem(word) for word in stemmed_content if not word in stopwords.words('english')]
      stemmed_content = ' '.join(stemmed_content)
      return stemmed_content
```

```
[34]: news_data['content'] = news_data['content'].apply(stemming)
```

```
[35]: news_data['content']
```

```
Out[35]: 0      darrel lucu hous dem aid even see comey letter...
1      daniel j flynn flynn hillari clinton big woman...
2      consortiumnew com truth might get fire
3      jessica purkiss civilian kill singl us airstri...
4      howard portnoy iranian woman jail fiction unpu...
      ...
20795  jerom hudson rapper trump poster child white s...
20796  benjamin hoffman n f l playoff schedul matchup...
20797  michael j de la merc rachel abram maci said re...
20798  alex ansari nato russia hold parallel exercis ...
20799  david swanson keep f aliv
Name: content, Length: 20800, dtype: object
```

Kelime düzenlenmesi ve tekrar görüntülenmesi


```
In [36]: X = news_data['content'].values  
Y = news_data['label'].values
```

```
In [37]: X
```

```
Out[37]: array(['darrel lucu hous dem aid even see comey letter jason chaffetz tweet',  
               'daniel j flynn flynn hillari clinton big woman campu breitbart',  
               'consortiumnew com truth might get fire', ...,  
               'michael j de la merc rachel abram maci said receiv takeov approach hudson bay new york time',  
               'alex ansari nato russia hold parallel exercis balkan',  
               'david swanson keep f aliv'], dtype=object)
```

```
In [38]: Y
```

```
Out[38]: array([1, 0, 1, ..., 0, 1, 1], dtype=int64)
```

Metinsel verileri sayısal verilere çevirme

```
In [53]: vectorizer = TfidfVectorizer()  
vectorizer.fit(X)  
X = vectorizer.transform(X)
```

```
In [54]: X
```

```
Out[54]: <20800x17128 sparse matrix of type '<class 'numpy.float64'>'  
         with 210687 stored elements in Compressed Sparse Row format>
```

Training - Test işlemi Verileri eğitim ve test verilerine bölüyoruz ve bu şekilde öğrenme algoritmamızı oluşturuyoruz.

```
In [55]: X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, stratify=Y, random_state=2)
```

```
In [56]: print(X_train)
print("-----")
print(X_test)
print("-----")
print(Y_train)
print("-----")
print(Y_test)
```

```
(0, 16996)    0.09995656818816077
(0, 16928)    0.3360072514082535
(0, 15295)    0.09807665903342763
(0, 13914)    0.3334797245354899
(0, 13014)    0.2680313811122545
(0, 12501)    0.3929876463935473
(0, 11936)    0.24142639024498436
(0, 10306)    0.09662001419895176
(0, 10219)    0.3019527708144002
(0, 3155)     0.3400831511004003
(0, 2794)     0.3776836172783757
(0, 336)      0.3360072514082535
(1, 16996)    0.07263181421455335
(1, 15424)    0.22579404836928033
(1, 15417)    0.26613170238131584
(1, 15295)    0.07126580880898774
(1, 13453)    0.3387500815971264
(1, 11421)    0.3084666283145136
(1, 10306)    0.07020736153621741
(1, 10000)    0.3400831511004003
```

Lojistik Regresyon Modelinin kurulması

```
In [57]: model = LogisticRegression()
```

```
In [58]: model.fit(X_train, Y_train)
```

```
Out[58]: LogisticRegression()
```

Training işlemi

Training işlemi

```
In [59]: X_train_prediction = model.predict(X_train)
```

```
In [60]: X_train_prediction
```

```
Out[60]: array([0, 0, 0, ..., 0, 0, 1], dtype=int64)
```

```
In [61]: training_data_accuracy = accuracy_score(X_train_prediction, Y_train)
```

```
In [62]: print('Training verilerinin doğruluk puanı :', training_data_accuracy)
```

```
Training verilerinin doğruluk puanı : 0.9865985576923076
```

Testing işlemi

```
In [63]: X_test_prediction = model.predict(X_test)  
testing_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

```
In [64]: print('Test verilerinin doğruluk puanı :', testing_data_accuracy)
```

```
Test verilerinin doğruluk puanı : 0.9790865384615385
```

Şimdi bu veri setini birde CART ile hesaplayalım..

```
In [67]: cart_model = DecisionTreeRegressor()
```

```
In [68]: cart_model
```

```
Out[68]: DecisionTreeRegressor()
```

```
In [72]: cart_model.fit(X_train,Y_train)
```

```
Out[72]: DecisionTreeRegressor()
```

```
In [72]: cart_model.fit(X_train,Y_train)
```

```
Out[72]: DecisionTreeRegressor()
```

```
In [73]: cart_X_train_prediction = cart_model.predict(X_train)
```

```
In [74]: cart_X_train_prediction
```

```
Out[74]: array([0., 0., 0., ..., 0., 0., 1.])
```

```
In [75]: cart_training_data_accuracy = accuracy_score(cart_X_train_prediction, Y_train)
```

```
In [76]: print('Training verilerinin doğruluk puanı :', cart_training_data_accuracy)
```

```
Training verilerinin doğruluk puanı : 1.0
```

Testing İşlemi

```
In [79]: cart_X_test_prediction = cart_model.predict(X_test)  
cart_testing_data_accuracy = accuracy_score(cart_X_test_prediction,Y_test)
```

```
In [80]: print('Test verilerinin doğruluk puanı :', cart_testing_data_accuracy)
```

```
Test verilerinin doğruluk puanı : 0.9913461538461539
```



```
### -----  
### Şimdi bu veri setini birde KNN ile hesaplayalım..
```

```
knn = KNeighborsClassifier(n_neighbors=3)  
knn.fit(X, Y)  
y_pred = knn.predict(X)  
print('Test verilerinin doğruluk puanı :', accuracy_score(y_pred, Y))
```

```
Test verilerinin doğruluk puanı : 0.5983653846153846
```

```
# Lojistik Regresyon Dogruluk Orani : %97.90
```

```
# CART Dogruluk Orani : %99.13
```


```
# KNN (n=3) Dogruluk Orani : %59.8
```

Sonuç kısmı ise ; Lojistik Regrasyon , CART Modeli ve KNN modelini kullanarak üç ayrı sonuç aldım. CART modeli bize daha iyi bir sonuç verdi.

Lojistik Regresyon :

Lojistik regresyon, bağımlı değişkenin kategorik bir değişken olduğu regresyon problemi gibidir. Doğrusal sınıflandırma problemlerinde yaygın bir biçimde kullanılır. Regresyon denilmesine rağmen burada bir sınıflandırma söz konusudur.

Lojistik regresyon, bir sonucu belirleyen bir veya daha fazla bağımsız değişken bulunan bir veri kümesini analiz etmek için kullanılan istatistiksel bir yöntemdir. Sonuç, ikili bir değişkenle ölçülür (yalnızca iki olası sonuç vardır). Lojistik regresyonda, bağımlı değişken ikili veya ikili, yani yalnızca 1 (DOĞRU, başarı, hamile vb.) Veya 0 (YANLIŞ, hata, gebe olmayan vb.) Olarak kodlanmış verileri içeriyor.

 Lojistik regresyonun amacı, iki yönlü karakteristiği (bağımlı değişken = yanıt veya sonuç değişkeni) ile ilgili bir dizi bağımsız (öngörücü veya açıklayıcı) değişken arasındaki ilişkiyi tanımlamak için en uygun (henüz biyolojik olarak makul) modeli bulmaktır.

CART Modeli :

CART, karar ağacı oluşturmada kullanılan bir algoritmadır. İkili ağaç yapısı vardır. Yani ana düğümden iki yavru düğüm oluşur. Homojen bir ağaç yapısı elde edilmeye çalışılır.

Tüm veri tipleri ile çalışabilen CART algoritmasında temel nokta, karar noktalarında ikili seçim ile birimlerin homojen sınıflar oluşacak şekilde ayrılmasıdır

KNN Modeli :

KNN, Denetimli Öğrenmede sınıflandırma ve regresyon için kullanılan algoritmalarından biridir. Diğer Denetimli Öğrenme algoritmalarının aksine, eğitim aşamasına sahip değildir. Eğitim ve test hemen hemen aynı şeydir. Tembel bir öğrenme türüdür.

Bu nedenle, kNN, geniş veri setini işlemek için gereken algoritma olarak ideal bir aday değildir. KNN, Denetimli Öğrenmede sınıflandırma ve regresyon için kullanılan algoritmalarından biridir. En basit makine öğrenmesi algoritması olarak kabul edilir. Diğer Denetimli Öğrenme algoritmalarının aksine, eğitim aşamasına sahip değildir. Eğitim ve test hemen hemen aynı şeydir. Tembel bir öğrenme türüdür. Bu nedenle, kNN, geniş veri setini işlemek için gereken algoritma olarak ideal bir aday değildir. KNN ile temelde yeni noktaya en yakın noktalar aranır. K, bilinmeyen noktanın en yakın komşularının miktarını temsil eder. Sonuçları tahmin etmek için algoritmanın k miktarını (genellikle bir tek sayı) seçeriz.

Thank
you

