Universidad Carlos III de Madrid

Universidad Carlos III de Madrid

Aprendizaje Automático, Grado en Estadística y Empresa

Práctica I: KNN y Árboles de clasificación

Marcos Álvarez Martín
Fabio Scielzo Ortiz

Índice

```
import pandas as pd
```

```
df_Fake = pd.read_csv('Fake.csv')
df_True = pd.read_csv('True.csv')
```

```
df_Fake['Fake'] = 1
df_True['Fake'] = 0
```

```
Fake_News_Data = pd.concat([df_Fake, df_True])
```

```
Fake_News_Data = Fake_News_Data.loc[: , ['Fake', 'title', 'text', 'date']
↪  ]
```

```
Fake_News_Data.index = range(0 , len(Fake_News_Data))
```

Fake = 1 yes

Fake = 0 no

```
Fake_News_Data.dtypes
```

```
Fake      int64
title    object
text     object
date     object
dtype: object
```

```
Fake_News_Data['Fake'] = Fake_News_Data['Fake'].astype('object')
```

```
Fake_News_Data.describe(include='all')
```

Fake

title

text

date

count

44898

44898

44898

44898

unique

2

38729

38646

2397

top

1

Factbox: Trump fills top jobs for his administ...

December 20, 2017

freq

23481

14

627

182

```
Fake_News_Data.isnull().sum()
```

```
Fake     0
title    0
text     0
date     0
dtype: int64
```

```
import numpy as np

import seaborn as sns
import matplotlib as mpl
import matplotlib.pyplot as plt

sns.set(rc={'figure.figsize':(8,8)})
```

```
prop_Fake_yes = len( Fake_News_Data.loc[ Fake_News_Data['Fake']== 1 , :] )
↪  / len(Fake_News_Data)

prop_Fake_no = len( Fake_News_Data.loc[ Fake_News_Data['Fake']== 0 , :] )
↪  / len(Fake_News_Data)
```

```
Fake_News_Data['proportion_Fakes'] = 0


for i in range(0, len(Fake_News_Data)):

    if Fake_News_Data['Fake'][i] == 1 :

        Fake_News_Data['proportion_Fakes'][i] = prop_Fake_yes

    else :

        Fake_News_Data['proportion_Fakes'][i] = prop_Fake_no
```

```
C:\Users\Usuario\AppData\Local\Temp\ipykernel_5228\2699169446.py:8: SettingWithCopyWarnin
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_g
  Fake_News_Data['proportion_Fakes'][i] = prop_Fake_yes
```

```
p1 = sns.barplot(x='Fake', y='proportion_Fakes', data=Fake_News_Data,
↪  palette="Spectral")
p1.set_yticks( np.arange(0, 0.85, 0.1)  )
p1.set_xticklabels(['No', 'Yes'])
p1.axes.set(xlabel='Fakes', ylabel='proportion')
```

```
[Text(0.5, 0, 'Fakes'), Text(0, 0.5, 'proportion')]
```

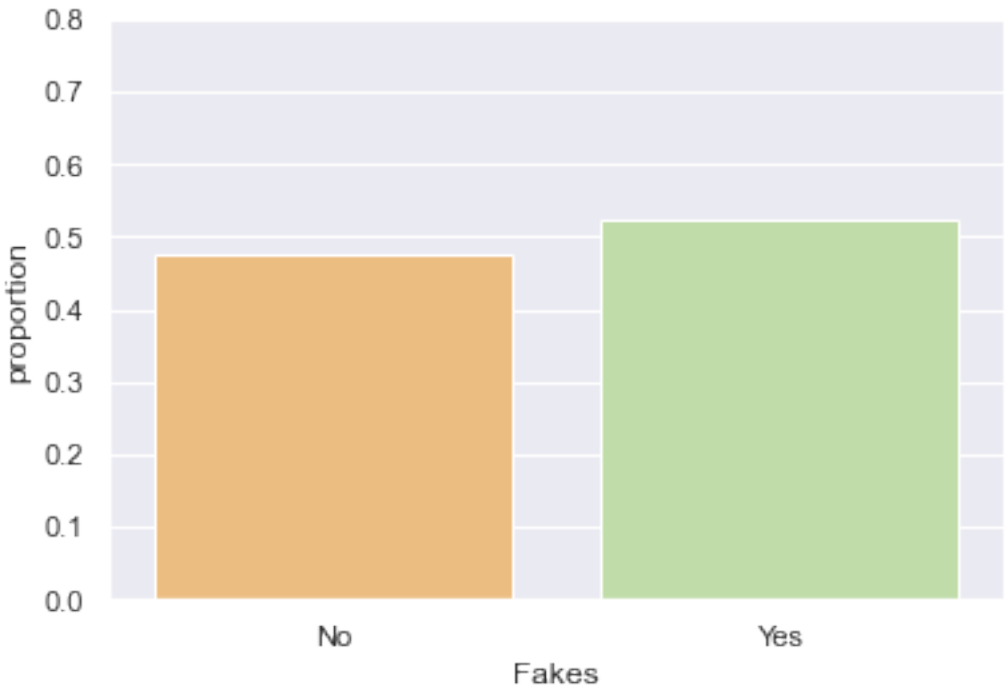

Figure 1: png

```
[prop_Fake_no , prop_Fake_yes]
```

```
[0.47701456635039424, 0.5229854336496058]
```

```
[prop_Fake_no*len(Fake_News_Data) , prop_Fake_yes*len(Fake_News_Data)]
```

```
[21417.0, 23481.0]
```

```
Fake_News_Data = Fake_News_Data.loc[ : , Fake_News_Data.columns !=
↪  'proportion_Fakes']
```

```
Fake_News_Data['word_count'] =
↪  Fake_News_Data['text'].str.split().str.len()
```

```
Fake_News_Data
```

Fake

title

text

date

word_count

0

1

Donald Trump Sends Out Embarrassing New Year'...

Donald Trump just couldn t wish all Americans ...

December 31, 2017

495

1

1

Drunk Bragging Trump Staffer Started Russian ...

House Intelligence Committee Chairman Devin Nu...

December 31, 2017

305

2

1

Sheriff David Clarke Becomes An Internet Joke...

On Friday, it was revealed that former Milwauk...

December 30, 2017

580

3

1

Trump Is So Obsessed He Even Has Obama's Name...

On Christmas day, Donald Trump announced that ...

December 29, 2017

444

4

1

Pope Francis Just Called Out Donald Trump Dur...

Pope Francis used his annual Christmas Day mes...

December 25, 2017

420

...

...

...

...

...

...

44893

0

‘Fully committed’ NATO backs new U.S. approach...

BRUSSELS (Reuters) - NATO allies on Tuesday we...

August 22, 2017

466

44894

0

LexisNexis withdrew two products from Chinese ...

LONDON (Reuters) - LexisNexis, a provider of l...

August 22, 2017

125

44895

0

Minsk cultural hub becomes haven from authorities

MINSK (Reuters) - In the shadow of disused Sov...

August 22, 2017

320

44896

0

Vatican upbeat on possibility of Pope Francis ...

MOSCOW (Reuters) - Vatican Secretary of State ...

August 22, 2017

205

44897

0

Indonesia to buy $1.14 billion worth of Russia...

JAKARTA (Reuters) - Indonesia will buy 11 Sukh...

August 22, 2017

210

44898 rows × 5 columns

```
Fake_News_Data.groupby('Fake')['word_count'].mean()
```

```
Fake
0    385.640099
1    423.197905
Name: word_count, dtype: float64
```

```
def limpiar_tokenizar(texto):

    import re

    '''
    Esta función limpia y tokeniza el texto en palabras individuales.
    El orden en el que se va limpiando el texto no es arbitrario.
    El listado de signos de puntuación se ha obtenido de:
    print(string.punctuation)
    y re.escape(string.punctuation)
    '''

    # Se convierte todo el texto a minúsculas

    nuevo_texto = texto.lower()

    # Eliminación de páginas web (palabras que empiezan por "http")

    nuevo_texto = re.sub('http\S+', ' ', nuevo_texto)

    # Eliminación de signos de puntuación

    regex =
    '[\\!\\"\\#\\$\\%\\&\\'\\(\\)\\*\\+\\,\\-\\.\\/\\:\\;\\<\\=\\>\\?\\@\\[\\\\\\]\\^_\\

    nuevo_texto = re.sub(regex , ' ', nuevo_texto)

    # Eliminación de números

    nuevo_texto = re.sub("\d+", ' ', nuevo_texto)

    # Eliminación de espacios en blanco múltiples

    nuevo_texto = re.sub("\\s+", ' ', nuevo_texto)

    # Tokenización por palabras individuales

    nuevo_texto = nuevo_texto.split(sep = ' ')

    # Eliminación de tokens con una longitud <= 1

    nuevo_texto = [token for token in nuevo_texto if len(token) >=  2]

    return(nuevo_texto)
```

```
test = "Esto es 1 ejemplo de l'limpieza de6 TEXTO  https://t.co/rnHPgyhx4Z
↪  @cienciadedatos #textmining"

print(limpiar_tokenizar(texto=test))
```

```
['esto', 'es', 'ejemplo', 'de', 'limpieza', 'de', 'texto', 'cienciadedatos', 'textmining'
```

```
Fake_News_Data['text'][0]
```

```
'Donald Trump just couldn t wish all Americans a Happy New Year and leave it at that. Ins
```

```
print(limpiar_tokenizar(texto=Fake_News_Data['text'][0]))
```

```
['donald', 'trump', 'just', 'couldn', 'wish', 'all', 'americans', 'happy', 'new', 'year',
```

```
Fake_News_Data['text_tokenizado'] = Fake_News_Data['text'].apply(
↪  limpiar_tokenizar )
```

```
Fake_News_Data['id_text'] = range(0, len(Fake_News_Data))
```

```
Fake_News_Data
```

Fake

title

text

date

word_count

text_tokenizado

id_text

0

1

Donald Trump Sends Out Embarrassing New Year'...

Donald Trump just couldn t wish all Americans ...

December 31, 2017

495

[donald, trump, just, couldn, wish, all, ameri...

0

1

1

Drunk Bragging Trump Staffer Started Russian ...

House Intelligence Committee Chairman Devin Nu...

December 31, 2017

305

[house, intelligence, committee, chairman, dev...

1

2

1

Sheriff David Clarke Becomes An Internet Joke...

On Friday, it was revealed that former Milwauk...

December 30, 2017

580

[on, friday, it, was, revealed, that, former, ...

2

3

1

Trump Is So Obsessed He Even Has Obama's Name...

On Christmas day, Donald Trump announced that ...

December 29, 2017

444

[on, christmas, day, donald, trump, announced,...

3

4

1

Pope Francis Just Called Out Donald Trump Dur...

Pope Francis used his annual Christmas Day mes...

December 25, 2017

420

[pope, francis, used, his, annual, christmas, ...

4

...

...

...

...

...

...

...

...

44893

0

‘Fully committed’ NATO backs new U.S. approach...

BRUSSELS (Reuters) - NATO allies on Tuesday we...

August 22, 2017

466

[brussels, reuters, nato, allies, on, tuesday,...

44893

44894

0

LexisNexis withdrew two products from Chinese ...

LONDON (Reuters) - LexisNexis, a provider of l...

August 22, 2017

125

[london, reuters, lexisnexis, provider, of, le...

44894

44895

0

Minsk cultural hub becomes haven from authorities

MINSK (Reuters) - In the shadow of disused Sov...

August 22, 2017

320

[minsk, reuters, in, the, shadow, of, disused,...

44895

44896

0

Vatican upbeat on possibility of Pope Francis ...

MOSCOW (Reuters) - Vatican Secretary of State ...

August 22, 2017

205

[moscow, reuters, vatican, secretary, of, stat...

44896

44897

0

Indonesia to buy $1.14 billion worth of Russia...

JAKARTA (Reuters) - Indonesia will buy 11 Sukh...

August 22, 2017

210

[jakarta, reuters, indonesia, will, buy, sukho...

44897

44898 rows × 7 columns

```
Fake_News_Tokens = Fake_News_Data.loc[:, ['id_text', 'text_tokenizado',
↪  'Fake'] ].explode(column='text_tokenizado')

Fake_News_Tokens =
↪  Fake_News_Tokens.rename(columns={'text_tokenizado':'token'})
```

```
Fake_News_Tokens
```

id_text

token

Fake

0

0

donald

1

0

0

trump

1

0

0

just

1

0

0

couldn

1

0

0

wish

1

...

...

...

...

44897

44897

technology

0

44897

44897

and

0

44897

44897

aviation

0

44897

44897

among

0

44897

44897

others

0

17503760 rows × 3 columns

```
# nº de palabras (tokens) en el conjunto de textos clasificados como fake
↪  y en los no fake

Fake_News_Tokens.groupby(by='Fake')['token'].count()
```

```
Fake
0    7891501
1    9611544
Name: token, dtype: int64
```

```
# nº de palabras (tokens) *unicas* en el conjunto de textos clasificados
↪  como fake y en los no fake

Fake_News_Tokens.groupby(by='Fake')['token'].nunique()
```

```
Fake
0    78020
1    85642
Name: token, dtype: int64
```

```
# nº de palabras (tokens) en cada texto individual clasificados como fake
↪  y en los no fake

df1 = pd.DataFrame( Fake_News_Tokens.groupby(by = ["id_text" , "Fake"]
↪  )["token"].count().rename('nº_tokens') )
```

```
df1
```

nº_tokens

id_text

Fake

0

1

447

1

1

294

2

1

563

3

1

426

4

1

415

...

...

...

44893

0

433

44894

0

120

44895

0

307

44896

0

196

44897

0

197

44898 rows × 1 columns

```
df2 = df1.loc[df1['nº_tokens'] != 0, :]
```

```
df2
```

nº_tokens

id_text

Fake

0

1

447

1

1

294

2

1

563

3

1

426

4

1

415

...

...

...

44893

0

433

44894

0

120

44895

0

307

44896

0

196

44897

0

197

44183 rows × 1 columns

```
df2.groupby("Fake")["nº_tokens"].agg(['mean'])
```

mean

Fake

0

368.486225

1

422.169983

Otra forma de hacer lo anterior (longitud media de las noticias fake y no fake)

```
m0 = (
↪  Fake_News_Tokens.loc[Fake_News_Tokens['Fake']==0].groupby('id_text')['token'].count()
↪  ).mean()
```

```
m1 = (
↪  Fake_News_Tokens.loc[Fake_News_Tokens['Fake']==1].groupby('id_text')['token'].count()
↪  ).mean()
```

```
pd.DataFrame({'fake_new': [0,1] , 'tokens_mean':[m0 , m1]})
```

fake_new

tokens_mean

0

0

368.469020

1

1

409.332822

0.1 ——————————————————————————

```
df = pd.DataFrame(  (Fake_News_Tokens.groupby(by = ["Fake", "token"]
↪  )["token"].count().unstack(fill_value=0).stack().reset_index(name='frecuencia_token')

# .unstack(fill_value=0).stack() para que tambien aparezcan los tokens con
↪  count = 0 , si no solo aprecerian los que tienen count > 0.
```

```
df # Nos da el nº de veces que sale cada token en el conjunto de las
↪  noticas fake y por otro lado en el de las no fake (solo salen tokens
↪  con count > 0 )
```

0.2 ————————————————————————-

Fake

token

frecuencia_token

0

0

aa

22

1

0

aaa

7

2

0

aaaaaaaand

0

3

0

aaaaackkk

0

4

0

aaaaapkfhk

0

. . .

. . .

. . .

. . .

251605

1

” ”it

0

251606

1

” ”when

0

251607

1

• if

0

251608

1

a

0

251609

1

⇒

0

251610 rows × 3 columns
