classification-with-logistic-regre

November 2, 2024

1 Multiclass Text Classification with

2 Logistic Regression Implemented with PyTorch and CE Loss

First, we will do some initialization.

[ ]:
```
import random
import torch
import numpy as np
import pandas as pd
from tqdm.notebook import tqdm

# enable tqdm in pandas
tqdm.pandas()

# set to True to use the gpu (if there is one available)
use_gpu = True

# select device
device = torch.device('cuda' if use_gpu and torch.cuda.is_available() else␣
 ↪'cpu')
print(f'device: {device.type}')

# random seed
seed = 1234

# set random seed
if seed is not None:
    print(f'random seed: {seed}')
    random.seed(seed)
    np.random.seed(seed)
    torch.manual_seed(seed)
```

```
device: cpu
random seed: 1234
```

2.0.1 En esta parte se importan las librerías necesarias para procesamiento de datos y aprendizaje profundo. Se configura use_gpu = True para intentar usar la GPU si está disponible, y se selecciona el dispositivo ('cuda' para GPU o 'cpu' para CPU) mediante torch.device, imprimiendo el dispositivo seleccionado. Luego, se establece una semilla aleatoria (seed = 1234) para asegurar que cada vez que se corra el codigo se tengan los mismos resultados

We will be using the AG's News Topic Classification Dataset. It is stored in two CSV files: `train.csv` and `test.csv`, as well as a `classes.txt` that stores the labels of the classes to predict.

First, we will load the training dataset using pandas and take a quick look at how the data.

[ ]:
```
train_df = pd.read_csv('/kaggle/input/ag-news-classification-dataset/train.csv')
train_df = train_df.sample(frac=0.7, random_state=42)
train_df.columns = ['class index', 'title', 'description']
train_df
```

[ ]:
```
        class index                                              title  \
71787             3            BBC set for major shake-up, claims newspaper
67218             3                                Marsh averts cash crunch
54066             2                 Jeter, Yankees Look to Take Control (AP)
7168              4                                Flying the Sun to Safety
29618             3                 Stocks Seen Flat as Nortel and Oil Weigh
…               …                                                    …
53857             1                  FDA Accused of Silencing Vioxx Warnings
111476            2        Buckeyes won #39;t play in NCAA or NIT tourneys
6343              3                           Rate hikes by Fed work in two ways
20736             4  NASA Administrator Offers Support for Kennedy …
34378             2                                Twins make it 3 straight

                                              description
71787   London - The British Broadcasting Corporation,…
67218   Embattled insurance broker #39;s banks agree t…
54066   AP - Derek Jeter turned a season that started …
7168    When the Genesis capsule comes back to Earth w…
29618    NEW YORK (Reuters) - U.S. stocks were set to …
…                                                   …
53857   WASHINGTON - The Food and Drug Administration …
111476  COLUMBUS, Ohio Ohio State has sanctioned its m…
6343    If you #39;ve noticed that the price of everyt…
20736   The following is a statement from NASA Adminis…
34378   The Minnesota Twins clinched on a bus in 1991…

[84000 rows x 3 columns]
```

The dataset consists of 120,000 examples, each consisting of a class index, a title, and a description. The class labels are distributed in a separated file. We will add the labels to the dataset so that we can interpret the data more easily. Note that the label indexes are one-based, so we need to subtract one to retrieve them from the list.

2.0.2 Se importan los datos de un dataset de kaggle y se dividen en el conjunto de entrenamiento con un valor del 70% de los datos, nos quedamos solo con las 3 columnas seleccionadas

[ ]:
```
labels = open('/kaggle/input/classes/classes.txt').read().splitlines()
classes = train_df['class index'].map(lambda i: labels[i-1])
train_df.insert(1, 'class', classes)
train_df
```

[ ]:
```
        class index     class  \
71787             3  Business
67218             3  Business
54066             2    Sports
7168              4  Sci/Tech
29618             3  Business
…               …         …
53857             1     World
111476            2    Sports
6343              3  Business
20736             4  Sci/Tech
34378             2    Sports

                                                    title  \
71787        BBC set for major shake-up, claims newspaper
67218                           Marsh averts cash crunch
54066            Jeter, Yankees Look to Take Control (AP)
7168                            Flying the Sun to Safety
29618            Stocks Seen Flat as Nortel and Oil Weigh
…                                                    …
53857             FDA Accused of Silencing Vioxx Warnings
111476   Buckeyes won #39;t play in NCAA or NIT tourneys
6343                      Rate hikes by Fed work in two ways
20736   NASA Administrator Offers Support for Kennedy …
34378                             Twins make it 3 straight

                                             description
71787   London - The British Broadcasting Corporation,…
67218   Embattled insurance broker #39;s banks agree t…
54066   AP - Derek Jeter turned a season that started …
7168    When the Genesis capsule comes back to Earth w…
29618    NEW YORK (Reuters) - U.S. stocks were set to …
…                                                    …
53857   WASHINGTON - The Food and Drug Administration …
111476  COLUMBUS, Ohio Ohio State has sanctioned its m…
6343    If you #39;ve noticed that the price of everyt…
20736   The following is a statement from NASA Adminis…
34378   The Minnesota Twins clinched on a bus in 1991…
```

```
[84000 rows x 4 columns]
```

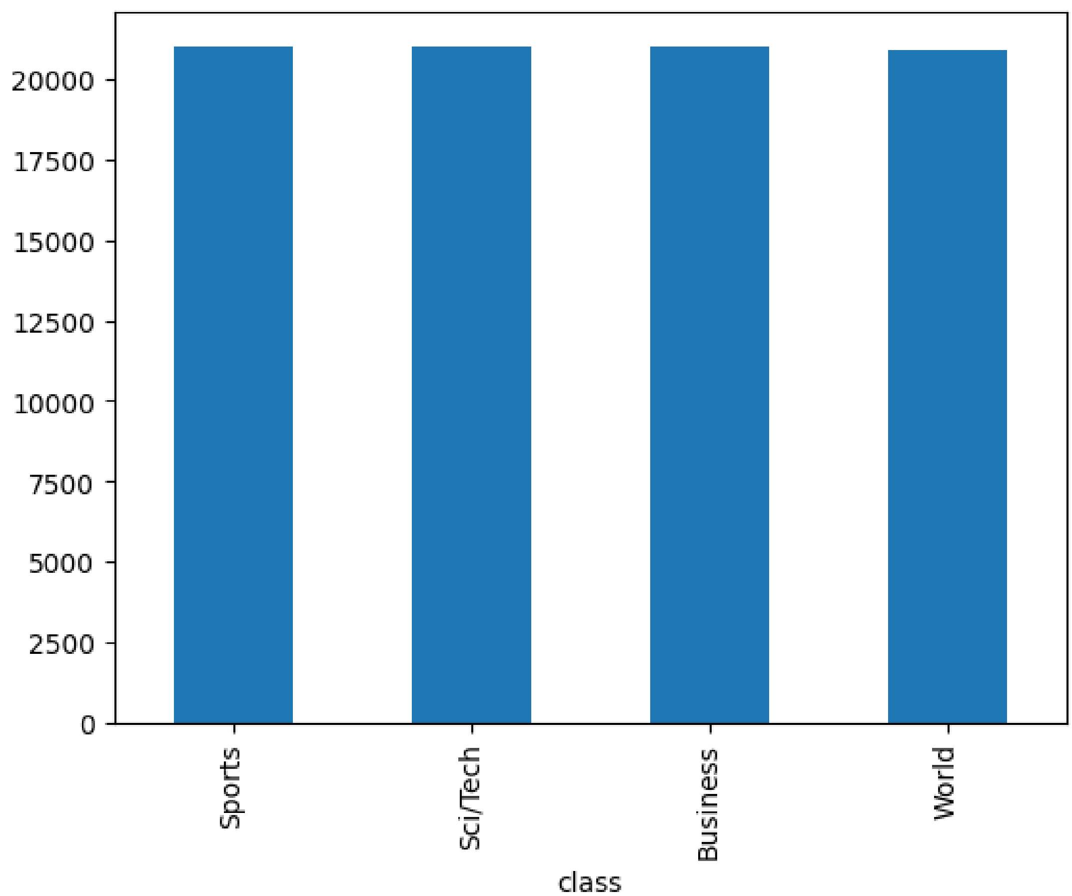
Let's inspect how balanced our examples are by using a bar plot.

En esta parte el código lee una lista de clases desde un archivo .txt, mapea los índices de clase en train_df a sus nombres correspondientes y agrega esta información al dataframe train_df como una nueva columna llamada 'class'.

```
[ ]: pd.value_counts(train_df['class']).plot.bar()
```

```
/tmp/ipykernel_311/1245903889.py:1: FutureWarning: pandas.value_counts is
deprecated and will be removed in a future version. Use
pd.Series(obj).value_counts() instead.
  pd.value_counts(train_df['class']).plot.bar()
```

```
[ ]: <Axes: xlabel='class'>
```



2.0.3 Se visualizan cuantos valores hay de cada clase para observar que las clases esten balanceadas.

The classes are evenly distributed. That's great!

However, the text contains some spurious backslashes in some parts of the text. They are meant to represent newlines in the original text. An example can be seen below, between the words "dwindling" and "band".

[ ]:
```
print(train_df.loc[0, 'description'])
```

```
Reuters - Short-sellers, Wall Street's dwindling\band of ultra-cynics, are
seeing green again.
```

We will replace the backslashes with spaces on the whole column using pandas replace method.

[ ]:
```
title = train_df['title'].str.lower()
descr = train_df['description'].str.lower()
text = title + " " + descr
train_df['text'] = text.str.replace('\\', ' ', regex=False)
train_df
```

[ ]:
```
        class index     class  \
71787             3  Business
67218             3  Business
54066             2    Sports
7168              4  Sci/Tech
29618             3  Business
…               …         …
53857             1     World
111476            2    Sports
6343              3  Business
20736             4  Sci/Tech
34378             2    Sports

                                                  title  \
71787        BBC set for major shake-up, claims newspaper
67218                              Marsh averts cash crunch
54066            Jeter, Yankees Look to Take Control (AP)
7168                               Flying the Sun to Safety
29618            Stocks Seen Flat as Nortel and Oil Weigh
…                                                     …
53857             FDA Accused of Silencing Vioxx Warnings
111476    Buckeyes won #39;t play in NCAA or NIT tourneys
6343                      Rate hikes by Fed work in two ways
20736   NASA Administrator Offers Support for Kennedy …
34378                              Twins make it 3 straight

                                            description  \
```

```
71787   London - The British Broadcasting Corporation,…
67218   Embattled insurance broker #39;s banks agree t…
54066   AP - Derek Jeter turned a season that started …
7168    When the Genesis capsule comes back to Earth w…
29618    NEW YORK (Reuters) - U.S. stocks were set to …
…                                                    …
53857   WASHINGTON - The Food and Drug Administration …
111476  COLUMBUS, Ohio Ohio State has sanctioned its m…
6343    If you #39;ve noticed that the price of everyt…
20736   The following is a statement from NASA Adminis…
34378   The Minnesota Twins clinched on a bus in 1991…

                                                      text
71787   bbc set for major shake-up, claims newspaper l…
67218   marsh averts cash crunch embattled insurance b…
54066   jeter, yankees look to take control (ap) ap - …
7168    flying the sun to safety when the genesis caps…
29618   stocks seen flat as nortel and oil weigh  new …
…                                                    …
53857   fda accused of silencing vioxx warnings washin…
111476  buckeyes won #39;t play in ncaa or nit tourney…
6343    rate hikes by fed work in two ways if you #39;…
20736   nasa administrator offers support for kennedy …
34378   twins make it 3 straight the minnesota twins c…

[84000 rows x 5 columns]
```

2.0.4 Se crea una nueva columna 'text' donde se combina el titulo y la descripcion en minusculas y reemplaza cualquier barra invertida con un espacio.

Now we will proceed to tokenize the title and description columns using NLTK's word_tokenize(). We will add a new column to our dataframe with the list of tokens.

[ ]:
```python
from nltk.tokenize import word_tokenize

train_df['tokens'] = train_df['text'].progress_map(word_tokenize)
train_df
```

```
  0%|          | 0/84000 [00:00<?, ?it/s]
```

[ ]:
```
        class index     class  \
71787             3  Business
67218             3  Business
54066             2    Sports
7168              4  Sci/Tech
29618             3  Business
…               …         …
53857             1     World
```

```
111476          2   Sports
6343            3 Business
20736           4 Sci/Tech
34378           2   Sports

                                                title  \
71787        BBC set for major shake-up, claims newspaper
67218                             Marsh averts cash crunch
54066           Jeter, Yankees Look to Take Control (AP)
7168                              Flying the Sun to Safety
29618           Stocks Seen Flat as Nortel and Oil Weigh
…                                                    …
53857            FDA Accused of Silencing Vioxx Warnings
111476   Buckeyes won #39;t play in NCAA or NIT tourneys
6343                 Rate hikes by Fed work in two ways
20736   NASA Administrator Offers Support for Kennedy …
34378                             Twins make it 3 straight

                                          description  \
71787   London - The British Broadcasting Corporation,…
67218   Embattled insurance broker #39;s banks agree t…
54066   AP - Derek Jeter turned a season that started …
7168    When the Genesis capsule comes back to Earth w…
29618    NEW YORK (Reuters) - U.S. stocks were set to …
…                                                    …
53857   WASHINGTON - The Food and Drug Administration …
111476  COLUMBUS, Ohio Ohio State has sanctioned its m…
6343    If you #39;ve noticed that the price of everyt…
20736   The following is a statement from NASA Adminis…
34378   The Minnesota Twins clinched on a bus in 1991…

                                                 text  \
71787   bbc set for major shake-up, claims newspaper l…
67218   marsh averts cash crunch embattled insurance b…
54066   jeter, yankees look to take control (ap) ap - …
7168    flying the sun to safety when the genesis caps…
29618   stocks seen flat as nortel and oil weigh  new …
…                                                    …
53857   fda accused of silencing vioxx warnings washin…
111476  buckeyes won #39;t play in ncaa or nit tourney…
6343    rate hikes by fed work in two ways if you #39;…
20736   nasa administrator offers support for kennedy …
34378   twins make it 3 straight the minnesota twins c…

                                               tokens
71787   [bbc, set, for, major, shake-up, ,, claims, ne…
67218   [marsh, averts, cash, crunch, embattled, insur…
```

```
54066    [jeter, ,, yankees, look, to, take, control, (…
7168     [flying, the, sun, to, safety, when, the, gene…
29618    [stocks, seen, flat, as, nortel, and, oil, wei…
…                                                     …
53857    [fda, accused, of, silencing, vioxx, warnings,…
111476   [buckeyes, won, #, 39, ;, t, play, in, ncaa, o…
6343     [rate, hikes, by, fed, work, in, two, ways, if…
20736    [nasa, administrator, offers, support, for, ke…
34378    [twins, make, it, 3, straight, the, minnesota,…

[84000 rows x 6 columns]
```

2.0.5 En esta parte se aplica tokenización a la columna 'text' generando listas de palabras en una nueva columna llamada 'tokens'

Now we will create a vocabulary from the training data. We will only keep the terms that repeat beyond some threshold established below.

[ ]:
```python
threshold = 10
tokens = train_df['tokens'].explode().value_counts()
tokens = tokens[tokens > threshold]
id_to_token = ['[UNK]'] + tokens.index.tolist()
token_to_id = {w:i for i,w in enumerate(id_to_token)}
vocabulary_size = len(id_to_token)
print(f'vocabulary size: {vocabulary_size:,}')
```

```
vocabulary size: 16,247
```

2.0.6 En esta parte se construye un vocabulario a partir de los tokens incluyendo aquellos que solo aparecen mas de 10 veces y agregando un token especial para palabras desconocidas. Luego crea un mapeo entre tokens y sus identificadores unicos y despues imprime el tamaño del vocabulario.

}

[ ]:
```python
from collections import defaultdict

def make_feature_vector(tokens, unk_id=0):
    vector = defaultdict(int)
    for t in tokens:
        i = token_to_id.get(t, unk_id)
        vector[i] += 1
    return vector

train_df['features'] = train_df['tokens'].progress_map(make_feature_vector)
train_df
```

```
  0%|          | 0/84000 [00:00<?, ?it/s]
```

[ ]:

```
        class index     class  \
71787             3  Business
67218             3  Business
54066             2    Sports
7168              4  Sci/Tech
29618             3  Business
…               …        …
53857             1     World
111476            2    Sports
6343              3  Business
20736             4  Sci/Tech
34378             2    Sports

                                                    title  \
71787        BBC set for major shake-up, claims newspaper
67218                               Marsh averts cash crunch
54066            Jeter, Yankees Look to Take Control (AP)
7168                              Flying the Sun to Safety
29618            Stocks Seen Flat as Nortel and Oil Weigh
…                                                     …
53857             FDA Accused of Silencing Vioxx Warnings
111476    Buckeyes won #39;t play in NCAA or NIT tourneys
6343                      Rate hikes by Fed work in two ways
20736   NASA Administrator Offers Support for Kennedy …
34378                               Twins make it 3 straight

                                              description  \
71787   London - The British Broadcasting Corporation,…
67218   Embattled insurance broker #39;s banks agree t…
54066   AP - Derek Jeter turned a season that started …
7168    When the Genesis capsule comes back to Earth w…
29618    NEW YORK (Reuters) - U.S. stocks were set to …
…                                                     …
53857   WASHINGTON - The Food and Drug Administration …
111476  COLUMBUS, Ohio Ohio State has sanctioned its m…
6343    If you #39;ve noticed that the price of everyt…
20736   The following is a statement from NASA Adminis…
34378   The Minnesota Twins clinched on a bus in 1991…

                                                     text  \
71787   bbc set for major shake-up, claims newspaper l…
67218   marsh averts cash crunch embattled insurance b…
54066   jeter, yankees look to take control (ap) ap - …
7168    flying the sun to safety when the genesis caps…
29618   stocks seen flat as nortel and oil weigh  new …
…                                                     …
53857   fda accused of silencing vioxx warnings washin…
```

```
111476   buckeyes won #39;t play in ncaa or nit tourney…
6343     rate hikes by fed work in two ways if you #39;…
20736    nasa administrator offers support for kennedy …
34378    twins make it 3 straight the minnesota twins c…

                                                  tokens  \
71787    [bbc, set, for, major, shake-up, ,, claims, ne…
67218    [marsh, averts, cash, crunch, embattled, insur…
54066    [jeter, ,, yankees, look, to, take, control, (…
7168     [flying, the, sun, to, safety, when, the, gene…
29618    [stocks, seen, flat, as, nortel, and, oil, wei…
…                                                   …
53857    [fda, accused, of, silencing, vioxx, warnings,…
111476   [buckeyes, won, #, 39, ;, t, play, in, ncaa, o…
6343     [rate, hikes, by, fed, work, in, two, ways, if…
20736    [nasa, administrator, offers, support, for, ke…
34378    [twins, make, it, 3, straight, the, minnesota,…

                                                features
71787    {2481: 1, 166: 1, 11: 1, 198: 1, 6539: 2, 2: 5…
67218    {1922: 2, 0: 2, 731: 1, 5126: 1, 2818: 1, 739:…
54066    {7031: 2, 2: 1, 507: 1, 600: 1, 4: 1, 193: 1, …
7168     {2695: 1, 1: 4, 418: 2, 4: 3, 1046: 1, 96: 1, …
29618    {156: 2, 631: 1, 1509: 1, 21: 1, 2053: 2, 9: 1…
…                                                   …
53857    {2622: 1, 616: 1, 6: 3, 0: 3, 1639: 2, 2734: 1…
111476   {7265: 2, 241: 1, 12: 2, 13: 2, 8: 2, 149: 1, …
6343     {645: 1, 3975: 1, 27: 1, 1385: 1, 364: 1, 7: 1…
20736    {421: 2, 5284: 2, 845: 1, 420: 1, 11: 1, 3687:…
34378    {1985: 2, 204: 1, 29: 1, 424: 1, 555: 1, 1: 1,…

[84000 rows x 7 columns]
```

2.0.7 Este código convierte cada conjunto de tokens en un vector de características basado en frecuencias. Cada fila en la columna 'features' contiene un diccionario en el que las claves son identificadores de tokens y los valores son sus frecuencias en el texto.

[ ]:
```
def make_dense(feats):
    x = np.zeros(vocabulary_size)
    for k,v in feats.items():
        x[k] = v
    return x

X_train = np.stack(train_df['features'].progress_map(make_dense))
y_train = train_df['class index'].to_numpy() - 1
```

```
X_train = torch.tensor(X_train, dtype=torch.float32)
y_train = torch.tensor(y_train)
```

```
  0%|          | 0/84000 [00:00<?, ?it/s]
```

2.0.8 Este código convierte los vectores de características dispersos a vectores densos en una matriz de características (X_train) y ajusta las etiquetas de clase (y_train). Ambos se convierten a tensores de PyTorch para poder ser usados para el entrenamiento del modelo.

[ ]:
```
from torch import nn
from torch import optim

# hyperparameters
lr = 1.0
n_epochs = 5
n_examples = X_train.shape[0]
n_feats = X_train.shape[1]
n_classes = len(labels)

# initialize the model, loss function, optimizer, and data-loader
model = nn.Linear(n_feats, n_classes).to(device)
loss_func = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=lr)

# train the model
indices = np.arange(n_examples)
for epoch in range(n_epochs):
    np.random.shuffle(indices)
    for i in tqdm(indices, desc=f'epoch {epoch+1}'):
        # clear gradients
        model.zero_grad()
        # send datum to right device
        x = X_train[i].unsqueeze(0).to(device)
        y_true = y_train[i].unsqueeze(0).to(device)
        # predict label scores
        y_pred = model(x)
        # compute loss
        loss = loss_func(y_pred, y_true)
        # backpropagate
        loss.backward()
        # optimize model parameters
        optimizer.step()
```

```
epoch 1:   0%|          | 0/84000 [00:00<?, ?it/s]

epoch 2:   0%|          | 0/84000 [00:00<?, ?it/s]
```

```
epoch 3:   0%|          | 0/84000 [00:00<?, ?it/s]

epoch 4:   0%|          | 0/84000 [00:00<?, ?it/s]

epoch 5:   0%|          | 0/84000 [00:00<?, ?it/s]
```

2.0.9 En esta parte se entrena el modelo para cada ejemplo realiza una prediccion, calcula una perdida y ajusta los parametros usando SGD, y repite para cada epoca

Next, we evaluate on the test dataset

[ ]:

```
# repeat all preprocessing done above, this time on the test set
test_df = pd.read_csv('/kaggle/input/ag-news-classification-dataset/test.csv')
test_df = test_df.sample(frac=0.7, random_state = 42)
test_df.columns = ['class index', 'title', 'description']
test_df['text'] = test_df['title'].str.lower() + " " + test_df['description'].
 ↪str.lower()
test_df['text'] = test_df['text'].str.replace('\\', ' ', regex=False)
test_df['tokens'] = test_df['text'].progress_map(word_tokenize)
test_df['features'] = test_df['tokens'].progress_map(make_feature_vector)

X_test = np.stack(test_df['features'].progress_map(make_dense))
y_test = test_df['class index'].to_numpy() - 1
X_test = torch.tensor(X_test, dtype=torch.float32)
y_test = torch.tensor(y_test)
```

```
  0%|          | 0/5320 [00:00<?, ?it/s]

  0%|          | 0/5320 [00:00<?, ?it/s]

  0%|          | 0/5320 [00:00<?, ?it/s]
```

2.0.10 En esta parte se prepara el conjunto de prueba como se hizo con el de entrenamiento (tokenizacion, vectores, tensores)

[ ]:

```
from sklearn.metrics import classification_report

# set model to evaluation mode
model.eval()

# don't store gradients
with torch.no_grad():
    X_test = X_test.to(device)
    y_pred = torch.argmax(model(X_test), dim=1)
    y_pred = y_pred.cpu().numpy()
    print(classification_report(y_test, y_pred, target_names=labels))
```

```
              precision    recall  f1-score   support
```

```
       World       0.79      0.93      0.86      1330
      Sports       0.96      0.93      0.95      1334
    Business       0.88      0.77      0.82      1314
    Sci/Tech       0.86      0.85      0.86      1342

    accuracy                           0.87      5320
   macro avg       0.88      0.87      0.87      5320
weighted avg       0.88      0.87      0.87      5320
```

2.0.11 En esta parte se evalua el modelo obteniendo los resultados de precision, recall, f1-score