NLP

February 19, 2021

1 Questão 5

Para este exercício, foi implementada uma extração de features com o vetorizador de textos TfidfVectorizer e para o classificador foi ultiliada uma LGBM. Tanto o vetorizador quanto o classificador tiveram seus parâmetros tunados com o scikit-optimize.

[1]:
```
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold

from skopt import forest_minimize

from lightgbm import LGBMClassifier
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics import roc_auc_score, average_precision_score

from sklearn.metrics import plot_confusion_matrix
```

[2]:
```
df = pd.read_excel('../data/teste_smarkio_lbs.xls', sheet_name='NLP')
```

[3]:
```
df.head()
```

[3]:

	letra	artista
0	Jay-z Uh-uh-uh You ready b? Let's go get 'em. ...	Beyoncé
1	Your challengers are a young group from Housto...	Beyoncé
2	Dum-da-de-da Do, do, do, do, do, do (Coming do...	Beyoncé
3	If I ain't got nothing I got you If I ain't go...	Beyoncé
4	Six inch heels She walked in the club like nob...	Beyoncé

[4]:
```
df.artista.unique()
```

[4]:
```
array(['Beyoncé', 'Rihanna'], dtype=object)
```

Como temos somente duas classes, os artistas foram substituidos por 0 (‘Beyoncé’) e 1 (‘Rihanna’).

[5]:
```
df.artista = df.artista.apply(lambda x: 0 if x == 'Beyoncé' else 1)
```

O dataset foi então dividido em treino e validação, com 33% do dataset para teste.

[6]:
```
df_train, df_test, ytrain, ytest = train_test_split(df.letra, df.artista,␣
 ↪test_size=0.2, random_state=42)

X = df_train
y = ytrain
```

Em seguida foi implementado um modelo Naive Bayes simples, para ser usado como baseline.

[7]:
```
from sklearn.naive_bayes import GaussianNB

letra_vec = TfidfVectorizer()
letra_bow_train = letra_vec.fit_transform(df_train)
letra_bow_val = letra_vec.transform(df_test)

mdl_nb = GaussianNB()

mdl_nb.fit(letra_bow_train.toarray(), ytrain)

p_nb = mdl_nb.predict_proba(letra_bow_val.toarray())[:, 1]
```

[8]:
```
average_precision_score(ytest, p_nb), roc_auc_score(ytest, p_nb)
```

[8]:
```
(0.528657616892911, 0.5643729189789123)
```

Com o Naive Bayes avaliado, o próximo passo foi implementar a função de tune dos parâmetros e executar classificador proposto.

[9]:
```
def tune_lgbm(params):
    print(params)
    lr = params[0]
    max_depth = params[1]
    min_child_samples = params[2]
    subsample = params[3]
    colsample_bytree = params[4]
    n_estimators = params[5]

    min_df = params[6]
    ngram_range = (1, params[7])

    average = []

    for (train, val) in KFold(5).split(X, y):
        X_train, X_val = X.iloc[train], X.iloc[val]
        y_train, y_val = y.iloc[train], y.iloc[val]

        letra_vec = TfidfVectorizer(min_df=min_df, ngram_range=ngram_range)
```

```
        letra_bow_train = letra_vec.fit_transform(X_train)
        letra_bow_val = letra_vec.transform(X_val)

        mdl = LGBMClassifier(learning_rate=lr, num_leaves=2 ** max_depth,␣
 ↪max_depth=max_depth,
                             min_child_samples=min_child_samples,␣
 ↪subsample=subsample,
                             colsample_bytree=colsample_bytree,␣
 ↪n_estimators=n_estimators, random_state=0,
                             class_weight="balanced", n_jobs=6)

        mdl.fit(letra_bow_train, y_train)

        p = mdl.predict_proba(letra_bow_val)[:, 1]

        average.append(-average_precision_score(y_val, p))

    return np.mean(average)


space = [(1e-3, 1e-1, 'log-uniform'), # lr
          (1, 10), # max_depth
          (1, 20), # min_child_samples
          (0.05, 1.), # subsample
          (0.05, 1.), # colsample_bytree
          (100,1000), # n_estimators
          (1,5), # min_df
          (1,5)] # ngram_range

res = forest_minimize(tune_lgbm, space, random_state=160745, n_random_starts=20,␣
 ↪n_calls=50, verbose=1)
```

```
Iteration No: 1 started. Evaluating function at random point.
[0.009944912110647982, 5, 1, 0.4677107511929402, 0.49263223036174764, 272, 3, 1]
Iteration No: 1 ended. Evaluation done at random point.
Time taken: 3.3330
Function value obtained: -0.7480
Current minimum: -0.7480
Iteration No: 2 started. Evaluating function at random point.
[0.053887464791860025, 1, 15, 0.7437489153990157, 0.8675167974293533, 549, 3, 4]
Iteration No: 2 ended. Evaluation done at random point.
Time taken: 5.1339
Function value obtained: -0.7578
Current minimum: -0.7578
Iteration No: 3 started. Evaluating function at random point.
[0.004151454520895999, 6, 20, 0.8682075103820793, 0.9491436163200662, 411, 4, 3]
Iteration No: 3 ended. Evaluation done at random point.
```

```
Time taken: 4.0346
Function value obtained: -0.7650
Current minimum: -0.7650
Iteration No: 4 started. Evaluating function at random point.
[0.0014099928811969545, 9, 9, 0.6502182010234373, 0.6866210554187129, 828, 5, 2]
Iteration No: 4 ended. Evaluation done at random point.
Time taken: 9.5960
Function value obtained: -0.7800
Current minimum: -0.7800
Iteration No: 5 started. Evaluating function at random point.
[0.08530558241838007, 8, 19, 0.2137736299768322, 0.1313765544201984, 961, 4, 1]
Iteration No: 5 ended. Evaluation done at random point.
Time taken: 2.6424
Function value obtained: -0.7226
Current minimum: -0.7800
Iteration No: 6 started. Evaluating function at random point.
[0.003567949451535685, 10, 19, 0.7232951768944309, 0.7298538828427115, 939, 4,
3]
Iteration No: 6 ended. Evaluation done at random point.
Time taken: 6.5351
Function value obtained: -0.7803
Current minimum: -0.7803
Iteration No: 7 started. Evaluating function at random point.
[0.014828577273549474, 7, 1, 0.18428087097824575, 0.3261556557915816, 274, 1, 2]
Iteration No: 7 ended. Evaluation done at random point.
Time taken: 41.6620
Function value obtained: -0.7703
Current minimum: -0.7803
Iteration No: 8 started. Evaluating function at random point.
[0.0015212976972079912, 3, 12, 0.44234694306528044, 0.399351303640462, 272, 3,
5]
Iteration No: 8 ended. Evaluation done at random point.
Time taken: 6.8598
Function value obtained: -0.7686
Current minimum: -0.7803
Iteration No: 9 started. Evaluating function at random point.
[0.01946212855369041, 9, 18, 0.5235636153223084, 0.6728679300083596, 747, 4, 5]
Iteration No: 9 ended. Evaluation done at random point.
Time taken: 8.1060
Function value obtained: -0.7725
Current minimum: -0.7803
Iteration No: 10 started. Evaluating function at random point.
[0.0012116790683302117, 3, 2, 0.06616307483844217, 0.23025600705315752, 677, 2,
5]
Iteration No: 10 ended. Evaluation done at random point.
Time taken: 31.8602
Function value obtained: -0.7583
Current minimum: -0.7803
```

```
Iteration No: 11 started. Evaluating function at random point.
[0.0053139776214487944, 6, 9, 0.14251441334450304, 0.8175761405215897, 297, 1,
5]
Iteration No: 11 ended. Evaluation done at random point.
Time taken: 11.0490
Function value obtained: -0.7708
Current minimum: -0.7803
Iteration No: 12 started. Evaluating function at random point.
[0.0068572961982704935, 10, 5, 0.2390386584472456, 0.49053406102209746, 176, 2,
4]
Iteration No: 12 ended. Evaluation done at random point.
Time taken: 9.7771
Function value obtained: -0.7926
Current minimum: -0.7926
Iteration No: 13 started. Evaluating function at random point.
[0.00781968225875022, 3, 4, 0.7078936710077383, 0.31818755505678337, 275, 4, 4]
Iteration No: 13 ended. Evaluation done at random point.
Time taken: 6.7621
Function value obtained: -0.7701
Current minimum: -0.7926
Iteration No: 14 started. Evaluating function at random point.
[0.017293945600511968, 2, 15, 0.9007557574888567, 0.41026441194439994, 316, 5,
1]
Iteration No: 14 ended. Evaluation done at random point.
Time taken: 1.2056
Function value obtained: -0.7342
Current minimum: -0.7926
Iteration No: 15 started. Evaluating function at random point.
[0.012250750764764855, 8, 6, 0.5976582413192033, 0.2474882432951916, 516, 4, 4]
Iteration No: 15 ended. Evaluation done at random point.
Time taken: 10.2254
Function value obtained: -0.7739
Current minimum: -0.7926
Iteration No: 16 started. Evaluating function at random point.
[0.018353598126553926, 4, 3, 0.47305622526323254, 0.1404164811277527, 133, 4, 1]
Iteration No: 16 ended. Evaluation done at random point.
Time taken: 1.6024
Function value obtained: -0.7216
Current minimum: -0.7926
Iteration No: 17 started. Evaluating function at random point.
[0.0010383234748454694, 9, 19, 0.9256771571832196, 0.9321438677645206, 312, 4,
3]
Iteration No: 17 ended. Evaluation done at random point.
Time taken: 4.4954
Function value obtained: -0.7602
Current minimum: -0.7926
Iteration No: 18 started. Evaluating function at random point.
[0.004955229758078229, 5, 5, 0.06939551310802591, 0.4193273080472823, 725, 4, 1]
```

```
Iteration No: 18 ended. Evaluation done at random point.
Time taken: 4.2496
Function value obtained: -0.7513
Current minimum: -0.7926
Iteration No: 19 started. Evaluating function at random point.
[0.0699516121742407, 9, 10, 0.6477856515609233, 0.8594430701440198, 616, 1, 1]
Iteration No: 19 ended. Evaluation done at random point.
Time taken: 3.4947
Function value obtained: -0.7603
Current minimum: -0.7926
Iteration No: 20 started. Evaluating function at random point.
[0.0014752743467850462, 5, 4, 0.9747950537021096, 0.982207187458162, 909, 2, 4]
Iteration No: 20 ended. Evaluation done at random point.
Time taken: 12.6882
Function value obtained: -0.7653
Current minimum: -0.7926
Iteration No: 21 started. Searching for the next optimal point.
[0.01125398986068822, 9, 7, 0.14473876182828932, 0.0879360872417403, 179, 5, 2]
Iteration No: 21 ended. Search finished for the next optimal point.
Time taken: 3.1844
Function value obtained: -0.7733
Current minimum: -0.7926
Iteration No: 22 started. Searching for the next optimal point.
[0.0797549993390731, 9, 2, 0.05315153331975511, 0.12966303282390224, 164, 3, 4]
Iteration No: 22 ended. Search finished for the next optimal point.
Time taken: 11.8603
Function value obtained: -0.7640
Current minimum: -0.7926
Iteration No: 23 started. Searching for the next optimal point.
[0.08410044886036924, 8, 7, 0.5731286012446287, 0.09025587140282179, 820, 4, 2]
Iteration No: 23 ended. Search finished for the next optimal point.
Time taken: 5.9770
Function value obtained: -0.7568
Current minimum: -0.7926
Iteration No: 24 started. Searching for the next optimal point.
[0.005787721824929793, 10, 8, 0.5939121340018981, 0.6172307431544003, 829, 2, 1]
Iteration No: 24 ended. Search finished for the next optimal point.
Time taken: 7.0781
Function value obtained: -0.7719
Current minimum: -0.7926
Iteration No: 25 started. Searching for the next optimal point.
[0.0014391963055247722, 10, 3, 0.6293515708711286, 0.36310585870701184, 754, 5,
1]
Iteration No: 25 ended. Search finished for the next optimal point.
Time taken: 11.0055
Function value obtained: -0.7595
Current minimum: -0.7926
Iteration No: 26 started. Searching for the next optimal point.
```

```
[0.006601489646809119, 10, 12, 0.4500443401698899, 0.12740119212035794, 625, 3,
4]
Iteration No: 26 ended. Search finished for the next optimal point.
Time taken: 8.9421
Function value obtained: -0.7590
Current minimum: -0.7926
Iteration No: 27 started. Searching for the next optimal point.
[0.017751614487899404, 10, 14, 0.18857044456717498, 0.2197144981552267, 901, 4,
2]
Iteration No: 27 ended. Search finished for the next optimal point.
Time taken: 5.8772
Function value obtained: -0.7653
Current minimum: -0.7926
Iteration No: 28 started. Searching for the next optimal point.
[0.018765373851852736, 10, 1, 0.3659523375704048, 0.5485214926450781, 122, 2, 4]
Iteration No: 28 ended. Search finished for the next optimal point.
Time taken: 49.6902
Function value obtained: -0.7623
Current minimum: -0.7926
Iteration No: 29 started. Searching for the next optimal point.
[0.01598567296800897, 10, 5, 0.4900965530905445, 0.63930087274586, 997, 2, 1]
Iteration No: 29 ended. Search finished for the next optimal point.
Time taken: 11.8982
Function value obtained: -0.7586
Current minimum: -0.7926
Iteration No: 30 started. Searching for the next optimal point.
[0.005777704365766577, 5, 5, 0.21229049285961343, 0.1412571065913224, 109, 2, 2]
Iteration No: 30 ended. Search finished for the next optimal point.
Time taken: 3.0050
Function value obtained: -0.7601
Current minimum: -0.7926
Iteration No: 31 started. Searching for the next optimal point.
[0.006917152621544271, 10, 13, 0.2638264040251437, 0.516738006247261, 192, 1, 4]
Iteration No: 31 ended. Search finished for the next optimal point.
Time taken: 8.6987
Function value obtained: -0.7766
Current minimum: -0.7926
Iteration No: 32 started. Searching for the next optimal point.
[0.00490070620162511, 10, 11, 0.15438068838870028, 0.19601658148691992, 170, 5,
1]
Iteration No: 32 ended. Search finished for the next optimal point.
Time taken: 1.8082
Function value obtained: -0.7487
Current minimum: -0.7926
Iteration No: 33 started. Searching for the next optimal point.
[0.0035147112822723975, 10, 6, 0.2511365985246428, 0.5489751927269599, 645, 1,
4]
Iteration No: 33 ended. Search finished for the next optimal point.
```

```
Time taken: 16.2775
Function value obtained: -0.7778
Current minimum: -0.7926
Iteration No: 34 started. Searching for the next optimal point.
[0.007884705802235688, 10, 3, 0.15663531483107163, 0.48411372271159103, 271, 2,
3]
Iteration No: 34 ended. Search finished for the next optimal point.
Time taken: 15.6840
Function value obtained: -0.7865
Current minimum: -0.7926
Iteration No: 35 started. Searching for the next optimal point.
[0.06645946279161963, 10, 6, 0.27386422030614416, 0.4330076439576368, 344, 2, 5]
Iteration No: 35 ended. Search finished for the next optimal point.
Time taken: 11.1731
Function value obtained: -0.7778
Current minimum: -0.7926
Iteration No: 36 started. Searching for the next optimal point.
[0.001963504599242526, 10, 19, 0.2501013649142373, 0.4155794752835497, 641, 2,
5]
Iteration No: 36 ended. Search finished for the next optimal point.
Time taken: 9.5535
Function value obtained: -0.7886
Current minimum: -0.7926
Iteration No: 37 started. Searching for the next optimal point.
[0.007239808599756794, 10, 2, 0.22207997441718702, 0.2604457347307627, 949, 2,
5]
Iteration No: 37 ended. Search finished for the next optimal point.
Time taken: 116.7873
Function value obtained: -0.7757
Current minimum: -0.7926
Iteration No: 38 started. Searching for the next optimal point.
[0.0041032579276086166, 10, 17, 0.05265727139908982, 0.12321172859649512, 286,
2, 5]
Iteration No: 38 ended. Search finished for the next optimal point.
Time taken: 6.7330
Function value obtained: -0.7500
Current minimum: -0.7926
Iteration No: 39 started. Searching for the next optimal point.
[0.015049930619097533, 10, 20, 0.10633981401933146, 0.48623669746437254, 355, 2,
4]
Iteration No: 39 ended. Search finished for the next optimal point.
Time taken: 5.8364
Function value obtained: -0.7683
Current minimum: -0.7926
Iteration No: 40 started. Searching for the next optimal point.
[0.004245651859976361, 10, 5, 0.10419601528411987, 0.6454676347373791, 347, 2,
5]
Iteration No: 40 ended. Search finished for the next optimal point.
```

```
Time taken: 11.8652
Function value obtained: -0.7915
Current minimum: -0.7926
Iteration No: 41 started. Searching for the next optimal point.
[0.005168899976786656, 10, 10, 0.1372880435892768, 0.39936720940702214, 718, 2,
5]
Iteration No: 41 ended. Search finished for the next optimal point.
Time taken: 13.0412
Function value obtained: -0.8055
Current minimum: -0.8055
Iteration No: 42 started. Searching for the next optimal point.
[0.0027249931916062467, 10, 17, 0.0538808101063164, 0.6169919999377859, 900, 2,
5]
Iteration No: 42 ended. Search finished for the next optimal point.
Time taken: 11.1177
Function value obtained: -0.7843
Current minimum: -0.8055
Iteration No: 43 started. Searching for the next optimal point.
[0.001216551559043194, 10, 4, 0.13346544053930245, 0.3465591709390513, 133, 2,
5]
Iteration No: 43 ended. Search finished for the next optimal point.
Time taken: 11.5710
Function value obtained: -0.7681
Current minimum: -0.8055
Iteration No: 44 started. Searching for the next optimal point.
[0.006753668932956609, 9, 17, 0.07192117948310067, 0.3510473204746125, 701, 2,
5]
Iteration No: 44 ended. Search finished for the next optimal point.
Time taken: 9.9926
Function value obtained: -0.7871
Current minimum: -0.8055
Iteration No: 45 started. Searching for the next optimal point.
[0.001233092021225729, 10, 9, 0.09010194400494967, 0.7664672928834364, 697, 2,
5]
Iteration No: 45 ended. Search finished for the next optimal point.
Time taken: 13.3443
Function value obtained: -0.7901
Current minimum: -0.8055
Iteration No: 46 started. Searching for the next optimal point.
[0.027122789097250747, 10, 9, 0.19311196467559588, 0.3961016684782185, 726, 5,
5]
Iteration No: 46 ended. Search finished for the next optimal point.
Time taken: 12.3095
Function value obtained: -0.7829
Current minimum: -0.8055
Iteration No: 47 started. Searching for the next optimal point.
[0.018068301715971867, 10, 9, 0.10869187076810655, 0.16092986247464472, 739, 1,
5]
```

```
Iteration No: 47 ended. Search finished for the next optimal point.
Time taken: 15.9067
Function value obtained: -0.7780
Current minimum: -0.8055
Iteration No: 48 started. Searching for the next optimal point.
[0.0015259815528993771, 10, 7, 0.11853950552638917, 0.4393854693873235, 644, 2,
5]
Iteration No: 48 ended. Search finished for the next optimal point.
Time taken: 16.0972
Function value obtained: -0.7909
Current minimum: -0.8055
Iteration No: 49 started. Searching for the next optimal point.
[0.004214919547592913, 10, 10, 0.05201856657430508, 0.9012319407803342, 677, 3,
5]
Iteration No: 49 ended. Search finished for the next optimal point.
Time taken: 12.2144
Function value obtained: -0.7794
Current minimum: -0.8055
Iteration No: 50 started. Searching for the next optimal point.
[0.003399250014190765, 10, 14, 0.1892657684805284, 0.2947769051799051, 662, 1,
5]
Iteration No: 50 ended. Search finished for the next optimal point.
Time taken: 13.2845
Function value obtained: -0.7868
Current minimum: -0.8055
```

Com os parâmetros otimizados, o classificador foi treinado e avaliado.

[10]:
```
params = res.x
lr = params[0]
max_depth = params[1]
min_child_samples = params[2]
subsample = params[3]
colsample_bytree = params[4]
n_estimators = params[5]

min_df = params[6]
ngram_range = (1, params[7])

letra_vec = TfidfVectorizer(min_df=min_df, ngram_range=ngram_range)
letra_bow_train = letra_vec.fit_transform(df_train)
letra_bow_val = letra_vec.transform(df_test)

mdl_lgbm = LGBMClassifier(learning_rate=lr, num_leaves=2 ** max_depth,␣
 ↪max_depth=max_depth,
                      min_child_samples=min_child_samples, subsample=subsample,
                      colsample_bytree=colsample_bytree,␣
 ↪n_estimators=n_estimators, random_state=0,
```

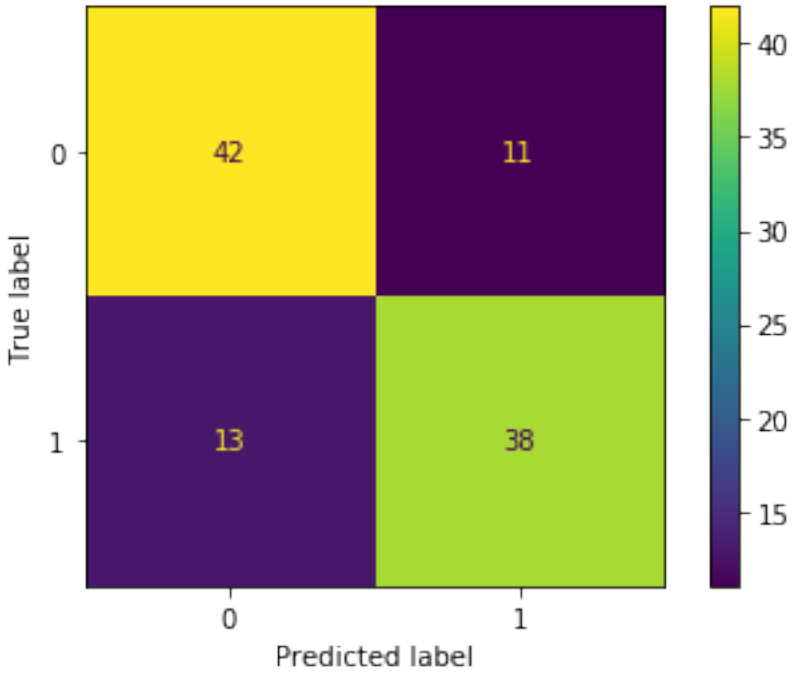
```
                      class_weight="balanced", n_jobs=6)
mdl_lgbm.fit(letra_bow_train, ytrain)

p_lgbm = mdl_lgbm.predict_proba(letra_bow_val)[:, 1]
```

[11]:
```
average_precision_score(ytest, p_lgbm), roc_auc_score(ytest, p_lgbm)
```

[11]:
```
(0.8033082763989248, 0.8305586385497595)
```

[12]:
```
plot_confusion_matrix(mdl_lgbm, letra_bow_val, ytest)
```

[12]:
```
<sklearn.metrics._plot.confusion_matrix.ConfusionMatrixDisplay at 0x221112d4e08>
```



Os resultados obtidos pela lgbm se mostrou bem superior aos do modelo Naive Bayes.

1.0.1 Salvar os Modelos

[13]:
```
import joblib as jb
```

[14]:
```
jb.dump(mdl_lgbm, "../model/lgbm_model.pkl.z")
jb.dump(letra_vec, "../model/letra_vectorizer.pkl.z")
```

[14]:
```
['../model/letra_vectorizer.pkl.z']
```