# Laboratorul 6

November 5, 2021

## 1 Laboratorul 6

[1]:
```
import os
import nltk

from sklearn.naive_bayes import GaussianNB
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.metrics import confusion_matrix, classification_report

import pandas as pd
import numpy as np
```

### 1.1 Prepare data

[2]:
```
training_reviews = []
testing_reviews = []
for folder in os.listdir('aclImdb'):
    if folder == 'train':
        new_folder_path = os.path.join('aclImdb', 'train')
        for folder in os.listdir(new_folder_path):
            if folder == 'neg':
                for file in os.listdir(os.path.join(new_folder_path, 'neg')):
                    review_file = open(os.path.join(new_folder_path, 'neg',␣
↪file), 'r', encoding="utf8")
                    training_reviews.append([review_file.read(), -1])
            elif folder == 'pos':
                for file in os.listdir(os.path.join(new_folder_path, 'pos')):
                    review_file = open(os.path.join(new_folder_path, 'pos',␣
↪file), 'r', encoding="utf8")
                    training_reviews.append([review_file.read(), 1])
    elif folder == 'test':
        new_folder_path = os.path.join('aclImdb', 'test')
        for folder in os.listdir(new_folder_path):
            if folder == 'neg':
                for file in os.listdir(os.path.join(new_folder_path, 'neg')):
                    review_file = open(os.path.join(new_folder_path, 'neg',␣
↪file), 'r', encoding="utf8")
```

```
                    testing_reviews.append([review_file.read(), -1])
            elif folder == 'pos':
                for file in os.listdir(os.path.join(new_folder_path, 'pos')):
                    review_file = open(os.path.join(new_folder_path, 'pos',␣
↪file), 'r', encoding="utf8")
                    testing_reviews.append([review_file.read(), 1])
```

[3]:
```
training_reviews = pd.DataFrame(training_reviews, columns=['review', 'label'])
testing_reviews = pd.DataFrame(testing_reviews, columns=['review', 'label'])
```

[4]:
```
training_reviews
```

[4]:

| | review | label |
|---|---|---|
| 0 | Story of a man who has unnatural feelings for … | -1 |
| 1 | Airport '77 starts as a brand new luxury 747 p… | -1 |
| 2 | This film lacked something I couldn't put my f… | -1 |
| 3 | Sorry everyone,,, I know this is supposed to b… | -1 |
| 4 | When I was little my parents took me along to … | -1 |
| … | … | … |
| 24995 | Seeing as the vote average was pretty low, and… | 1 |
| 24996 | The plot had some wretched, unbelievable twist… | 1 |
| 24997 | I am amazed at how this movie(and most others … | 1 |
| 24998 | A Christmas Together actually came before my t… | 1 |
| 24999 | Working-class romantic drama from director Mar… | 1 |

[25000 rows x 2 columns]

[5]:
```
testing_reviews
```

[5]:

| | review | label |
|---|---|---|
| 0 | Once again Mr. Costner has dragged out a movie… | -1 |
| 1 | This is an example of why the majority of acti… | -1 |
| 2 | First of all I hate those moronic rappers, who… | -1 |
| 3 | Not even the Beatles could write songs everyon… | -1 |
| 4 | Brass pictures (movies is not a fitting word f… | -1 |
| … | … | … |
| 24995 | I was extraordinarily impressed by this film. … | 1 |
| 24996 | Although I'm not a golf fan, I attended a snea… | 1 |
| 24997 | From the start of "The Edge Of Love", the view… | 1 |
| 24998 | This movie, with all its complexity and subtle… | 1 |
| 24999 | I've seen this story before but my kids haven'… | 1 |

[25000 rows x 2 columns]

### 1.2 Encode the training data and divide the number of occurences by the sum of frequencies for each word per class and add Laplace add-one rule for smoothing

[6]:
```
countvec = CountVectorizer(ngram_range=(1,1), stop_words='english')
encoded_data = countvec.fit_transform(training_reviews['review'])
```

[7]:
```
encoded_data_df = pd.DataFrame(encoded_data.toarray(),columns=countvec.
↪get_feature_names())
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[8]:
```
encoded_data_df
```

[8]:

| | 00 | 000 | 0000000000001 | 00001 | 00015 | 000s | 001 | 003830 | 006 | 007 | … | \ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| … | .. | … | … | … | … | … | … | … | … | … | … | |
| 24995 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 24996 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 24997 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 24998 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |
| 24999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | … | |

| | était | état | étc | évery | êxtase | ís | ísnt | østbye | über | üvegtigris |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| … | … | … | … | … | … | .. | … | … | … | … |
| 24995 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24996 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24997 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24998 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24999 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

[25000 rows x 74538 columns]

[9]:
```
# The length of entire vocabulary is
len(countvec.get_feature_names())
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[9]:
```
74538
```

[10]:
```
nominator = encoded_data_df.iloc[training_reviews[training_reviews['label'] ==␣
 ↪-1].index] + 1
denominator = np.sum(encoded_data_df.
 ↪iloc[training_reviews[training_reviews['label'] == -1].index], axis=0) +␣
 ↪len(countvec.get_feature_names())
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[11]:
```
# Update the encoded data
temp_negative = nominator / denominator.values
```

[12]:
```
nominator = encoded_data_df.iloc[training_reviews[training_reviews['label'] ==␣
 ↪1].index] + 1
denominator = np.sum(encoded_data_df.
 ↪iloc[training_reviews[training_reviews['label'] == 1].index], axis=0) +␣
 ↪len(countvec.get_feature_names())

# Update the encoded data
temp_positive = nominator / denominator.values
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[13]:
```
encoded_data_df = temp_negative.append(temp_positive)
```

[14]:
```
del temp_negative
del temp_positive
```

### 1.3 Train Gaussian Naive Bayes

[15]:
```
model = GaussianNB()
model.fit(encoded_data_df, training_reviews['label'])
```

[15]:
```
GaussianNB()
```

[16]:
```
del encoded_data_df
```

### 1.4 Encode testing data with Laplace add-one smoothing

[17]:
```
testing_encoded_data = countvec.transform(testing_reviews['review'])
testing_encoded_data_df = pd.DataFrame(testing_encoded_data.
 ↪toarray(),columns=countvec.get_feature_names())
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[18]:
```
nominator = testing_encoded_data_df.
 ↪iloc[testing_reviews[testing_reviews['label'] == -1].index] + 1
denominator = np.sum(testing_encoded_data_df.
 ↪iloc[testing_reviews[testing_reviews['label'] == -1].index], axis=0) +␣
 ↪len(countvec.get_feature_names())

# Update the encoded data
temp_negative = nominator / denominator.values
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[19]:
```
nominator = testing_encoded_data_df.
 ↪iloc[testing_reviews[testing_reviews['label'] == 1].index] + 1
denominator = np.sum(testing_encoded_data_df.
 ↪iloc[testing_reviews[testing_reviews['label'] == 1].index], axis=0) +␣
 ↪len(countvec.get_feature_names())

# Update the encoded data
temp_positive = nominator / denominator.values
```

```
C:\Users\abuinoschi\Anaconda3\envs\rn4nlp\lib\site-
packages\sklearn\utils\deprecation.py:87: FutureWarning: Function
get_feature_names is deprecated; get_feature_names is deprecated in 1.0 and will
be removed in 1.2. Please use get_feature_names_out instead.
  warnings.warn(msg, category=FutureWarning)
```

[20]:
```
testing_encoded_data_df = temp_negative.append(temp_positive)
```

[21]:
```
del temp_negative
del temp_positive
```

### 1.5 Predict for the test data

[22]:
```
predictions = model.predict(testing_encoded_data_df)
```

[23]:
```
del testing_encoded_data_df
```

[24]:
```
testing_reviews['predicted'] = predictions
```

[25]:
```
testing_reviews
```

[25]:

| | review | label | predicted |
|---|---|---|---|
| 0 | Once again Mr. Costner has dragged out a movie… | -1 | -1 |
| 1 | This is an example of why the majority of acti… | -1 | -1 |
| 2 | First of all I hate those moronic rappers, who… | -1 | -1 |
| 3 | Not even the Beatles could write songs everyon… | -1 | 1 |
| 4 | Brass pictures (movies is not a fitting word f… | -1 | -1 |
| … | … | … | … |
| 24995 | I was extraordinarily impressed by this film. … | 1 | -1 |
| 24996 | Although I'm not a golf fan, I attended a snea… | 1 | -1 |
| 24997 | From the start of "The Edge Of Love", the view… | 1 | 1 |
| 24998 | This movie, with all its complexity and subtle… | 1 | -1 |
| 24999 | I've seen this story before but my kids haven'… | 1 | -1 |

[25000 rows x 3 columns]

### 1.6 Print results metrics

[26]:
```
confusion_mat = confusion_matrix(testing_reviews['label'],␣
↪testing_reviews['predicted'])
confusion_mat_df = pd.DataFrame(confusion_mat, columns = ['Predicted negative',␣
↪'Predicted positive'], index = ['Actual negative', 'Actual positive'])
confusion_mat_df
```

[26]:

| | Predicted negative | Predicted positive |
|---|---|---|
| Actual negative | 9472 | 3028 |
| Actual positive | 7799 | 4701 |

[27]:
```
print(classification_report(testing_reviews['label'],␣
↪testing_reviews['predicted'], target_names=['Negative sentiment', 'Positive␣
↪sentiment']))
```

```
                    precision    recall  f1-score   support

Negative sentiment       0.55      0.76      0.64     12500
Positive sentiment       0.61      0.38      0.46     12500
```

```
              accuracy                           0.57     25000
             macro avg       0.58      0.57      0.55     25000
          weighted avg       0.58      0.57      0.55     25000
```

[ ]: