# Sci-kit learn

July 17, 2020

[1]:

```python
import json
import numpy as np
import random
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
from sklearn.metrics import f1_score
from sklearn import svm
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression
```

### 0.0.1 Load Data

[69]:

```python
file_name = 'Books_small_10000.json'
with open(file_name) as f:
    for line in f:
        review  = json.loads(line)
        print(review['reviewText'])
        print(review['overall'])
        break
```

```
I bought both boxed sets, books 1-5.  Really a great series!  Start book 1 three
weeks ago and just finished book 5.  Sloane Monroe is a great character and
being able to follow her through both private life and her PI life gets a reader
very involved!  Although clues may be right in front of the reader, there are
twists and turns that keep one guessing until the last page!  These are books
you won't be disappointed with.
5.0
```

[55]:

```python
class Sentiment:
    NEGATIVE = 'NEGATIVE'
    NEUTRAL  = 'NEUTRAL'
    POSITIVE = 'POSITIVE'

class Review:
    def __init__(self,text,score):
        self.text = text
        self.score = score
        self.sentiment = self.get_sentiment()
    def get_sentiment(self):
        if self.score <= 2:
            return Sentiment.NEGATIVE
        elif self.score == 3:
            return Sentiment.NEUTRAL
        else:
            return Sentiment.POSITIVE

class ReviewContainer:
    def _init__(self, reviews):
        self.reviews = reviews

    def evenly_distribute(self):
        negative = filter(lambda x: x.sentiment == Sentiment.NEGATIVE, self.
 ↪reviews)
        positive = filter(lambda x: x.sentiment == Sentiment.POSITIVE, self.
 ↪reviews)
```

[56]:

```
reviews = []
with open(file_name) as f:
    for line in f:
        review  = json.loads(line)
        reviews.append(Review(review['reviewText'],review['overall']))
reviews[5].score
```

[56]:

```
5.0
```

### 0.0.2 Prep Data

[57]:

```
training, test = train_test_split(reviews, test_size=0.33, random_state =42)
```

[58]:

```
print(training[0].sentiment)
```

```
POSITIVE
```

[59]:

```
train_x = [x.text for x in training]
train_y = [x.sentiment for x in training]

test_x = [x.text for x in test]
test_y = [x.sentiment for x in test]
print(train_y[0:10])
```

```
['POSITIVE', 'POSITIVE', 'POSITIVE', 'NEGATIVE', 'POSITIVE', 'POSITIVE',
'POSITIVE', 'POSITIVE', 'POSITIVE', 'NEGATIVE']
```

### 0.0.3 Bag of worlds

[60]:
```
vectorizer = CountVectorizer()
train_x_vectors = vectorizer.fit_transform(train_x)
test_x_vectors = vectorizer.transform(test_x)
print(train_x[0])
print(train_x_vectors[0].toarray())
```

```
Olivia Hampton arrives at the Dunraven family home as cataloger of their
extensive library. What she doesn't expect is a broken carriage wheel on the
way. Nor a young girl whose mind is clearly gone, an old man in need of care
himself (and doesn&#8217;t quite seem all there in Olivia&#8217;s opinion).
Furthermore, Marion Dunraven, the only sane one of the bunch and the one Olivia
is inexplicable drawn to, seems captive to everyone in the dusty old house. More
importantly, she doesn't expect to fall in love with Dunraven's daughter
Marion.Can Olivia truly believe the stories of sadness and death that surround
the house, or are they all just local neighborhood rumor?Was that carriage
trouble just a coincidence or a supernatural sign to stay away? If she remains,
will the Castle&#8217;s dark shadows take Olivia down with them or will she and
Marion long enough to declare their love?Patty G. Henderson has created an
atmospheric and intriguing story in her Gothic tale. I found this to be an
enjoyable read, even if it isn&#8217;t my usual preferred genre. I think, with
this tale, I got hooked on the old Gothic romantic style. So I think fans of the
genre (and of lesbian romances) will enjoy it.
[[0 0 0 … 0 0 0]]
```

### 0.0.4 Clasification, linear SVM

[61]:
```
clf_svm = svm.SVC(kernel= 'linear')
clf_svm.fit(train_x_vectors, train_y)
test_x[0]
clf_svm.predict(test_x_vectors[0])
```

[61]:
```
array(['POSITIVE'], dtype='<U8')
```

# 1 Decision Tree

[62]:
```
clf_dec = DecisionTreeClassifier()
clf_dec.fit(train_x_vectors, train_y)
clf_dec.predict(test_x_vectors[0])
```

[62]:
```
array(['POSITIVE'], dtype='<U8')
```

## 2 Logistic regression

[65]:

```
clf_log = LogisticRegression()
clf_log.fit(train_x_vectors, train_y)
clf_log.predict(test_x_vectors[0])
```

```
/home/antonio/anaconda3/lib/python3.7/site-
packages/sklearn/linear_model/_logistic.py:940: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
```

[65]:

```
array(['POSITIVE'], dtype='<U8')
```

### 2.0.1 Evaluation

[66]:

```
# Mean acuracy
print('SVM',clf_svm.score(test_x_vectors,test_y))
print('Dec',clf_dec.score(test_x_vectors,test_y))
print('Log',clf_log.score(test_x_vectors,test_y))
```

```
SVM 0.8124242424242424
Dec 0.7660606060606061
Log 0.8409090909090909
```

### 2.0.2 F1 scores

[67]:

```
print('SVM',f1_score(test_y,clf_svm.
 ↪predict(test_x_vectors),average=None,labels=[Sentiment.POSITIVE,Sentiment.
 ↪NEUTRAL,Sentiment.NEGATIVE]))
print('Dec',f1_score(test_y,clf_dec.
 ↪predict(test_x_vectors),average=None,labels=[Sentiment.POSITIVE,Sentiment.
 ↪NEUTRAL,Sentiment.NEGATIVE]))
print('Log',f1_score(test_y,clf_log.
 ↪predict(test_x_vectors),average=None,labels=[Sentiment.POSITIVE,Sentiment.
 ↪NEUTRAL,Sentiment.NEGATIVE]))
```

```
SVM [0.90738061 0.2656     0.40268456]
Dec [0.87178578 0.13735343 0.15934066]
Log [0.92139968 0.29250457 0.40983607]
```

[ ]: