200010021

October 14, 2022

0.1 *LAB 9 : Naive Bayes Classifier*

1. Binary Classification using Naive Bayes Classifier
2. Sentiment Analysis using Naive Bayes

[1]:
```
import numpy as np
import matplotlib.pyplot as plt
```

0.2 Binary Classification using Naive Bayes Classifier

Useful References : 1. https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/

2. https://www.analyticsvidhya.com/blog/2021/01/a-guide-to-the-naive-bayes-algorithm/
3. https://towardsdatascience.com/implementing-naive-bayes-algorithm-from-scratch-python-c6880cfc9c41

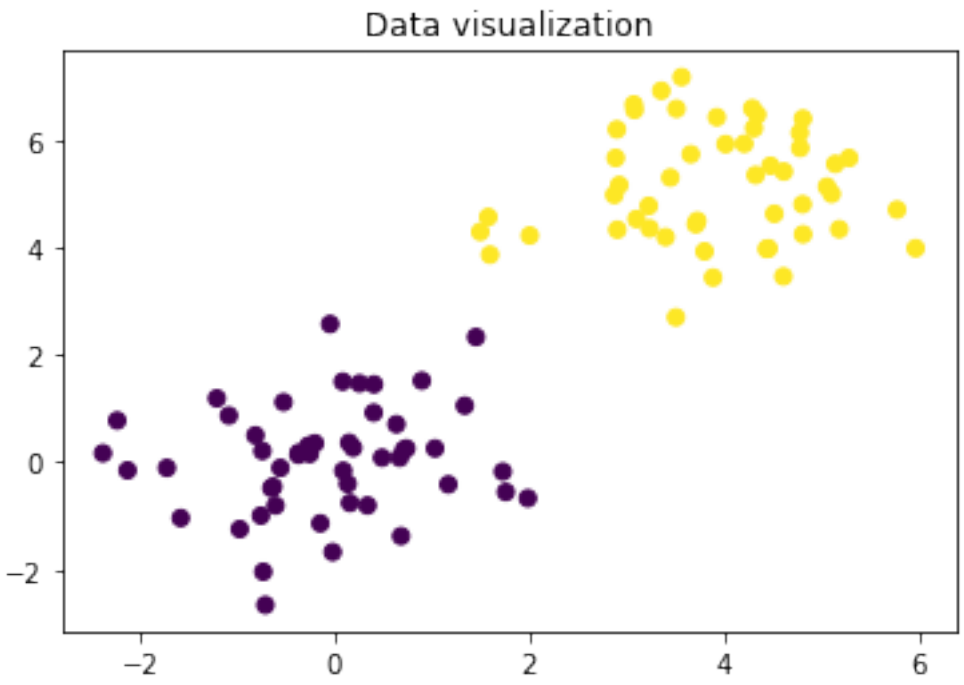
Note : The goal of this experiment is to perform and understand Naive Bayes classification by applying it on the below dataset, you can either fill in the below functions to get the result or you can create a class of your own using the above references to perform classification

1. Generation of 2D training data

[2]:
```
mean1=np.array([0,0])
mean2=np.array([4,5])
var=np.array([[1,0.1],[0.1,1]])
np.random.seed(0)
data1=np.random.multivariate_normal(mean1,var,50)
data2=np.random.multivariate_normal(mean2,var,50)
data=np.concatenate((data1,data2))
label=np.concatenate((np.zeros(data1.shape[0]),np.ones(data2.shape[0])))

plt.figure()
plt.scatter(data[:,0],data[:,1],c=label)
plt.title('Data visualization')
```

[2]:
```
Text(0.5, 1.0, 'Data visualization')
```



2. Split the Dataset by Class Values (Create a Dictionary)

[3]:
```
import pandas
def class_dictionary(data: np.ndarray, label: np.ndarray) -> pandas.DataFrame:
    data1 = data[np.where(label == 0)]
    data2 = data[np.where(label == 1)]
    dataset = pandas.DataFrame({'Lable 0': dict(enumerate(data1)), 'Lable 1':␣
 ↪dict(enumerate(data2))})
    return dataset

df = class_dictionary(data, label)
print(df['Lable 0'])
```

```
0      [-1.5766898498109918, -1.0398226180535974]
1       [-2.2290883728182354, 0.7773853414206259]
2      [-0.7294401424182014, -2.0405960058045194]
3      [-0.6030709264745643, -0.8061379304575134]
4     [-0.19888869928702047, 0.35198699778924736]
5        [-1.082381897285784, 0.8687307543819514]
6     [-0.6460227649981973, -0.48277860017337426]
7     [-0.5530133270230154, -0.10534224000978795]
8       [-0.9704143490174237, -1.245663043373122]
9       [0.3407676181772856, -0.8051220612456715]
10       [1.4548872377471582, 2.3318086042047077]
11     [-0.14322361217186214, -1.138942473862656]
```

```
12      [-0.7076769270859744, -2.658913234613081]
13    [0.09163131920609444, -0.15950216856263935]
14      [-2.12241531690744, -0.15106366120793535]
15    [-0.36859121651520094, 0.13876704370863302]
16      [1.9871581976060144, -0.6703591339181066]
17      [0.1531364786089271, 0.36290063248603555]
18       [-1.71898891150744, -0.1058270655127238]
19     [0.49004010574872753, 0.08445840570575938]
20     [1.7302046823573887, -0.17494929972333928]
21     [-0.04321607714272613, 2.5740237584006547]
22       [0.6718373492046692, 0.0840989986456544]
23      [0.40754151913707726, 1.4506542916314682]
24       [1.3396091959718586, 1.0541881591053868]
25       [0.4045536898735961, 0.9236378614002625]
26     [1.1708153748804253, -0.41316891719263105]
27     [-0.26643325391388417, 0.3082342537475074]
28      [-0.2522348097892404, 0.1535738249884692]
29      [0.7137600267642441, 0.22709168353337916]
30      [0.7399056087616519, 0.25751442232388433]
31      [1.7610709968252143, -0.5549801514881049]
32     [0.13794009685220396, -0.4011055944029041]
33        [0.8985436768026501, 1.519431226084436]
34       [0.6380244448928473, 0.7077165065638268]
35       [-0.6272323993344852, -0.45418366550087]
36    [-0.016655820885765338, -1.673348505695392]
37     [0.16100002744645364, -0.7577691215657584]
38       [1.034104609782848, 0.25749629028312987]
39      [0.1933766847798092, 0.26873039870295956]
40       [0.25980547041716073, 1.468391026874295]
41       [0.6851990856535891, -1.375888101858851]
42      [-2.3755184894275074, 0.16808375581876 09]
43     [-0.7535090781102758, -0.9949035763404194]
44      [0.08674367504699954, 1.5014391192930403]
45      [-0.5210370564229527, 1.1190451092578209]
46      [-0.8096102402485931, 0.5006885247858939]
47     [-0.7382720841040296, 0.20969529698872302]
48      [-1.2057853707937938, 1.1902113232369487]
49     [-0.36378318985853964, 0.1755421358779934]
Name: Lable 0, dtype: object
```

3. Calculate Mean, Std deviation and count for each column in a dataset

[4]:
```
def get_variables(dataset: pandas.DataFrame) -> dict:
  Lable_0 = {}
  Lable_1 = {}

  Lable_0['mean'] = np.mean(dataset['Lable 0'])
  Lable_0['std']  = np.array(dataset['Lable 0']).std()
```

```
  Lable_0['count'] = dataset['Lable 0'].count()

  Lable_1['mean'] = np.mean(dataset['Lable 1'])
  Lable_1['std']  = np.array(dataset['Lable 1']).std()
  Lable_1['count'] = dataset['Lable 1'].count()

  return Lable_0 , Lable_1

var0, var1 = get_variables(df)
#print(var1['mean'])
```

[5]:
```
print(var1)
```

```
{'mean': array([3.89129353, 5.13647498]), 'std': array([1.02082994,
1.03546734]), 'count': 50}
```

[6]:
```
x = list(enumerate(df['Lable 0'][1]))
print(df['Lable 0'][1])
print(x)
```

```
[-2.22908837  0.77738534]
[(0, -2.2290883728182354), (1, 0.7773853414206259)]
```

3. Calculate Class Probabilities

[7]:
```
import numpy
## calculate P(xi/ Y=y)
def guassian_dim(x: float, mean: float, std: float) -> float:

    exp_pow = -(1/2)*(x-mean)**2 / (std)**2
    A = np.exp(exp_pow)
    B = ( 1 / (np.sqrt(2*numpy.pi)*std) )

    return B * A

def calculate_class_prob(model: dict, data: numpy.ndarray, other_model: dict)␣
 ↪-> numpy.ndarray: ## calculate prob of data belonging to given model
    ## Calculation of P(data/Y) for each point of data
    Prob = []

    for x in data:
        class_prob = 1

        likelihood_model = 1
        likelihood_other_model = 1
        priori_model = model['count']/(model['count'] + other_model['count'])

        for i, xi in enumerate(x):
```

```
            likelihood_model *= guassian_dim(xi, model['mean'][i],␣
 ↪model['std'][i])
            likelihood_other_model *= guassian_dim(xi, other_model['mean'][i],␣
 ↪other_model['std'][i])

        class_prob = (likelihood_model * priori_model) / (likelihood_model +␣
 ↪likelihood_other_model)

        Prob.append(class_prob)

    return Prob

def predict(model0, model1, data):
    pred = []
    prob0 = calculate_class_prob(model0, data, model1)
    prob1 = calculate_class_prob(model1, data, model0)

    if len(prob0) != len(prob1):
        return "error"

    for i in range(len(prob0)):
        if prob0[i] - prob1[i] >= 0:
            pred.append(0)
        else:
            pred.append(1)

    return pred
```

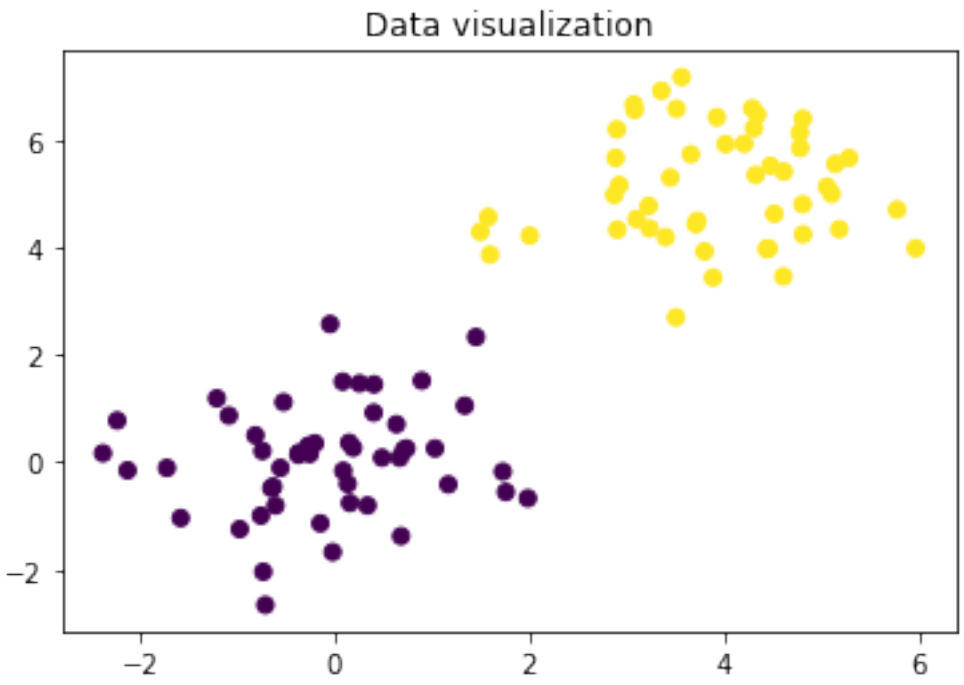
4. Test the model using some samples

[8]:
```
mean1=np.array([0,0])
mean2=np.array([4,5])
var=np.array([[1,0.1],[0.1,1]])

data1=np.random.multivariate_normal(mean1,var,10)
data2=np.random.multivariate_normal(mean2,var,10)
test_data=np.concatenate((data1,data2))
y_test=np.concatenate((np.zeros(data1.shape[0]),np.ones(data2.shape[0])))


plt.figure()
plt.scatter(data[:,0],data[:,1],c=label)
plt.title('Data visualization')
```

[8]:
```
Text(0.5, 1.0, 'Data visualization')
```



Testing for a sample point

[9]:
```
class_dict = class_dictionary(data,label)
var0, var1 = get_variables(class_dict)
out = predict(var0, var1, numpy.array(df['Lable 0'][:5]))
print(out)
```

```
[0, 0, 0, 0, 0]
```

As seen above the class probability for the 1st sample is given, we can observe that probability is higher for class 0 than 1 and hence imply that this datapoint belongs to class 0

Now Calculate the class probabilities for all the data points in the test dataset and calculate the accuracy by comparing the predicted labels with the true test labels

[10]:
```
y_pred = np.array(predict(var0, var1, test_data), dtype=float)
from sklearn.metrics import accuracy_score

print(f'Accuracy of model is {accuracy_score(y_test, y_pred)}')
```

```
Accuracy of model is 1.0
```

[11]:
```
print(y_pred)
print(y_test)
```

```
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
[0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

5. Use the Sci-kit Learn library to perform Gaussian Naive Bayes classifier on the above dataset, also report the accuracy and confusion matrix for the same

[12]:
```
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import confusion_matrix
nb=GaussianNB();
nb.fit(data,label)
y_predict=nb.predict(test_data)
print(confusion_matrix(y_test,y_predict))
```

```
[[10  0]
 [ 0 10]]
```

0.3 Sentiment Analysis using Naive Bayes Classifier

Go through the following article and implement the same

Keypoints :

1. The link to the dataset is given in the above article, download the same to perform sentiment analysis
2. Understanding how to deal with text data is very important since it requires a lot of preprocessing, you can go through this article if you are interested in learning more about it
3. Split the dataset into train-test and train the model
4. Report the accuracy metrics and try some sample prediction outside of those present in the dataset

Note : The goal of this experiment is to explore a practical use case of Naive bayes classifier as well as to understand how to deal with textual data, you can follow any other open source implemetations of sentiment analysis using naive bayes also

Other References :

1. https://towardsdatascience.com/sentiment-analysis-introduction-to-naive-bayes-algorithm-96831d77ac91
2. https://gist.github.com/CateGitau/6608912ca92733036c090676c61c13cd

[13]:
```
%pip install nltk
```

```
Requirement already satisfied: nltk in
/home/abhishekj/anaconda3/lib/python3.9/site-packages (3.7)
Requirement already satisfied: joblib in
/home/abhishekj/anaconda3/lib/python3.9/site-packages (from nltk) (1.1.0)
Requirement already satisfied: tqdm in
/home/abhishekj/anaconda3/lib/python3.9/site-packages (from nltk) (4.64.0)
Requirement already satisfied: regex>=2021.8.3 in
/home/abhishekj/anaconda3/lib/python3.9/site-packages (from nltk) (2022.3.15)
Requirement already satisfied: click in
/home/abhishekj/anaconda3/lib/python3.9/site-packages (from nltk) (8.0.4)
Note: you may need to restart the kernel to use updated packages.
```

[14]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
import joblib
from sklearn.feature_extraction.text import CountVectorizer

data = pd.read_csv('./train.csv')
```

[15]:
```
def preprocess_data(data):
    data = data.drop('package_name', axis=1)
    data['review'] = data['review'].str.strip().str.lower()
    return data

data = preprocess_data(data)
```

[16]:
```
x = data['review']
y = data['polarity']
x, x_test, y, y_test = train_test_split(x,y, stratify=y, test_size=0.25,␣
 ↪random_state=42)



vec = CountVectorizer(stop_words='english')
x = vec.fit_transform(x).toarray()
x_test = vec.transform(x_test).toarray()

from sklearn.naive_bayes import MultinomialNB

model = MultinomialNB()
model.fit(x, y)

model.score(x_test, y_test)

model.predict(vec.transform(['I love this app! this is so amazing!!!']))
```

[16]:
```
array([1])
```