Poisoning_Attack_Spam_Filter

April 20, 2022

1 Poisoning Attack Spam Filter

Code based on [Build a machine learning email spam detector with Python](https://blog.logrocket.com/email-spam-detector-python-machine-learning/

1.1 Import Libraries

[1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn import svm
```

1.2 Download Data and Process CSV

[2]:
```
!wget https://raw.githubusercontent.com/SmallLion/Python-Projects/main/
↪Spam-detection/spam.csv
```

```
--2022-04-20 20:42:11--  https://raw.githubusercontent.com/SmallLion/Python-
Projects/main/Spam-detection/spam.csv
Resolving raw.githubusercontent.com (raw.githubusercontent.com)…
185.199.109.133, 185.199.110.133, 185.199.108.133, …
Connecting to raw.githubusercontent.com
(raw.githubusercontent.com)|185.199.109.133|:443… connected.
HTTP request sent, awaiting response… 200 OK
Length: 499362 (488K) [text/plain]
Saving to: ‘spam.csv.8’

spam.csv.8          100%[===================>] 487.66K  --.-KB/s    in 0.03s

2022-04-20 20:42:11 (19.0 MB/s) - ‘spam.csv.8’ saved [499362/499362]
```

[3]:
```
spam = pd.read_csv('spam.csv')
```

1.3 Split Dataset into Train and Test

[4]:
```
x = spam['v2']
y = spam["v1"]
x_train, x_test, y_train, y_test = train_test_split(x, y,test_size = 0.2)
x_train, x_poisoning, y_train, y_poisoning = train_test_split(x_train,␣
 ↪y_train,test_size = 0.01)
```

[5]:
```
data_train_dic = {'x': x_train, 'y': y_train}
data_train = pd.DataFrame(data_train_dic)
data_test_dic = {'x': x_test, 'y': y_test}
data_test = pd.DataFrame(data_test_dic)
data_poisoning_dic = {'x': x_poisoning, 'y': y_poisoning}
data_poisoning = pd.DataFrame(data_poisoning_dic)
```

[6]:
```
print("TRAIN DATA. size:", len(data_train))
print(data_train.head(10))
```

```
TRAIN DATA. size: 4412
                                                      x    y
1541  Do u konw waht is rael FRIENDSHIP Im gving yuo…  ham
5209  I know you are thinkin malaria. But relax, chi…  ham
2239  Every day i use to sleep after  &lt;#&gt;  so …  ham
5439       Hey i've booked the 2 lessons on sun liao…  ham
618   I come n pick Ì_ up… Come out immediately af…  ham
2536                          You do what all you like  ham
4312  I wasn't well babe, i have swollen glands at m…  ham
2537          That's y we haf to combine n c how lor…  ham
4869  Dip's cell dead. So i m coming with him. U bet…  ham
4223  Double eviction this week - Spiral and Michael…  ham
```

[7]:
```
print("TEST DATA. size:", len(data_test))
print(data_test.head(10))
```

```
TEST DATA. size: 1115
                                                      x     y
2895  pdate_Now - Double mins and 1000 txts on Orang…  spam
827                             Hey. You got any mail?   ham
1270  Sorry chikku, my cell got some problem thts y …   ham
3902  Waiting in e car 4 my mum lor. U leh? Reach ho…   ham
4573                          :( but your not here…   ham
3875                            did u get that message   ham
2574  Your next amazing xxx PICSFREE1 video will be …  spam
3671   came to look at the flat, seems ok, in his 50…   ham
4964  A few people are at the game, I'm at the mall …   ham
1286  Night has ended for another day, morning has c…   ham
```

[8]:
```
print("POISONING DATA. size:", len(data_poisoning))
print(data_poisoning.head(10))
```

```
POISONING DATA. size: 45
                                                      x     y
4221  Yo you around? A friend of mine's lookin to pi…   ham
4901  * FREE* POLYPHONIC RINGTONE Text SUPER to 8713…  spam
3561  Still chance there. If you search hard you wil…   ham
4894  Want the latest Video handset? 750 anytime any…  spam
5474  Where's mummy's boy ? Is he being good or bad …   ham
443                             Sorry, I'll call later   ham
3046  Ok. Not much to do here though. H&M Friday, ca…   ham
1392  Thk shld b can… Ya, i wana go 4 lessons… H…   ham
4511       Now project pa. After that only i can come.   ham
4832  New Mobiles from 2004, MUST GO! Txt: NOKIA to …  spam
```

1.4 Obtain Features and Train Model

[9]:
```
cv = CountVectorizer()
features = cv.fit_transform(x_train)
```

[10]:
```
model = svm.SVC()
model.fit(features,y_train)
```

[10]:
```
SVC()
```

1.5 Evaluate Model Accuracy

[11]:
```
features_test = cv.transform(x_test)
print("Accuracy: {}".format(model.score(features_test,y_test)))
```

```
Accuracy: 0.9838565022421525
```

1.6 Poison 50% of Training Data and Re-train model

[12]:
```
import random

for i, y in enumerate(y_poisoning):
  if y == 'ham':
    y_poisoning.iloc[i] = 'spam'
  else:
    y_poisoning.iloc[i] = 'ham'
```

[13]:
```
data_poisoning_dic = {'x': x_poisoning, 'y': y_poisoning}
data_poisoning = pd.DataFrame(data_poisoning_dic)
print("POISONED DATA. size:", len(data_poisoning))
print(data_poisoning.head(10))
```

```
POISONED DATA. size: 45
```

	x	y
4221	Yo you around? A friend of mine's lookin to pi…	spam
4901	* FREE* POLYPHONIC RINGTONE Text SUPER to 8713…	ham
3561	Still chance there. If you search hard you wil…	spam
4894	Want the latest Video handset? 750 anytime any…	ham
5474	Where's mummy's boy ? Is he being good or bad …	spam
443	Sorry, I'll call later	spam
3046	Ok. Not much to do here though. H&M Friday, ca…	spam
1392	Thk shld b can… Ya, i wana go 4 lessons… H…	spam
4511	Now project pa. After that only i can come.	spam
4832	New Mobiles from 2004, MUST GO! Txt: NOKIA to …	ham

[14]:
```
cv = CountVectorizer()
features = cv.fit_transform(x_poisoning)
```

[15]:
```
model = svm.SVC()
model.fit(features,y_poisoning)
```

[15]:
```
SVC()
```

1.7 Evaluate Poisoned Model Accuracy

[16]:
```
features_test = cv.transform(x_test)
print("Accuracy: {}".format(model.score(features_test,y_test)))
```

```
Accuracy: 0.14349775784753363
```