Bayes Naïf-Classification des emails

November 13, 2023

Département Génie Informatique, FST de Tanger, MST IASD 2023-2024¶

Module “Apprentissage automatique”.

Bayes Naïf: Classification des emails (spam ou non spam) en utilisant la fréquence d’apparition des mots.

Livre de Peter Harrington, chapitre 4 “Classifying with probability theory: naïve Bayes“ qui concerne le prochain cours”, page 74.

[1]:

```python
def createVocabList(dataSet):
    vocabSet = set([])  # ensemble vide (pas de doublon )
    for document in dataSet:
        vocabSet = vocabSet | set(document) # Union
    return list(vocabSet)
```

Pour une entrée du dataset cacluler la fréquence d’appartion des mots du catalogue

[2]:

```python
import numpy as np
def bagOfWords2VecMN(vocabList, inputSet):
    returnVec = [0]*len(vocabList)
    for word in inputSet:
        if word in vocabList:
            returnVec[vocabList.index(word)] += 1
    return returnVec
```

0.0.1 Apprentissage :

Calcul des probabiltés conditionnelles pour chaque caractéristique et pour chaque classe + probabilité à priori de chaque classe

[3]:

```python
from numpy import *
def trainNB0(trainMatrix,trainCategory):
    numTrainDocs = len(trainMatrix)
    numWords = len(trainMatrix[0])
    pSpam = sum(trainCategory)/float(numTrainDocs)
    p0Num = zeros(numWords); p1Num = zeros(numWords)      #change to ones()
    p0Denom = 0.0; p1Denom = 0.0                          #change to 2.0
    for i in range(numTrainDocs):
        if trainCategory[i] == 1:
            p1Num += trainMatrix[i]
            p1Denom += sum(trainMatrix[i])
        else:
            p0Num += trainMatrix[i]
            p0Denom += sum(trainMatrix[i])

    # lissage
    p1Vect = (1+p1Num)/(2+p1Denom)
    p0Vect = (1+p0Num)/(2+p0Denom)
    return p0Vect,p1Vect,pSpam
```

0.0.2 Généralisation

- Entrée : vecteur des fréquences d'apparition des mots du vocabulaire

[4]:
```
def classifyNB(vec2Classify, p0Vect, p1Vect, pC1):
    p1 = sum(vec2Classify*log(p1Vect))+log(pC1)      # modèle multinomial
    p0 = sum(vec2Classify*log(p0Vect))+log(1.0 - pC1)

    if p1 > p0:
        return 1
    else:
        return 0
```

Transformer une chaine de caractères en une liste de mots

[5]:
```
import re
def textParse(bigString):   # input is big string, #output is word list
    listOfTokens = re.split("[^A-Za-z]",bigString)
    tokRet=[tok for tok in listOfTokens if len(tok)>4 ]
    return tokRet
```

Construire la base des exemples d'apprentissage et de test + calcul du taux de la mauvaise classification

[6]:
```
docList=[]; classList = []; fullText =[]
for i in range(1,26):
    doc = open('./email/spam/%d.txt' % i).read()
    fullText.append(doc)
    wordList = textParse(doc)
    docList.append(wordList)
    classList.append(1)
    doc = open('./email/nspam/%d.txt' % i).read()
    fullText.append(doc)
    wordList = textParse(doc)
    docList.append(wordList)
    classList.append(0)
vocabList = createVocabList(docList) # Create vocabulary
print('Feature vector size : ',len(vocabList))
print(vocabList)
```

```
Feature vector size :  464
['create', 'shape', 'forward', 'Wallets', 'faster', 'couple', 'address',
'proven', 'Learn', 'attaching', 'Quality', 'favorite', 'Shipment', 'enabled',
'Watches', 'Thanks', 'prices', 'pages', 'lists', 'pills', 'Required', 'Vicodin',
'check', 'mathematician', 'another', 'Experience', 'spaying', 'Ambiem',
'ViagraNoPrescription', 'would', 'website', 'FemaleViagra', 'suggest',
'service', 'Drugs', 'length', 'pictures', 'Jewerly', 'others', 'tickets',
'magazine', 'looking', 'creative', 'comment', 'derivatives', 'enough', 'group',
'LinkedIn', 'close', 'sliding', 'store', 'OrderCializViagra', 'Microsoft',
'Extended', 'Monte', 'should', 'items', 'Vuitton', 'control', 'parallel',
'These', 'Sorry', 'Office', 'reply', 'having', 'running', 'Genuine', 'ferguson',
'father', 'Certified', 'WARRANTY', 'plane', 'reputable', 'inside',
'reservation', 'BuyVIAGRA', 'generates', 'level', 'riding', 'while',
'prototype', 'doors', 'endorsed', 'Safest', 'chapter', 'Acrobat', 'assistance',
'major', 'Pharmacy', 'party', 'withoutPrescription', 'WatchesStore', 'increase',
'titles', 'drugs', 'Explosive', 'plugin', 'windows', 'museum', 'place',
'HardErecetions', 'borders', 'holiday', 'Design', 'Millions', 'ofEjacu',
'because', 'serial', 'location', 'Ultimate', 'Natural', 'interesting',
'Supplement', 'using', 'Where', 'inform', 'treat', 'Famous', 'discussions',
'Percocet', 'bathroom', 'Hello', 'Shipping', 'SeverePain', 'BetterEjacu',
'intenseOrgasns', 'china', 'download', 'BetterErections', 'hotels', 'Effective',
'Hermes', 'Major', 'stuff', 'thickness', 'opportunity', 'price', 'Looking',
'please', 'computer', 'about', 'Fermi', 'modelling', 'groups', 'Tesla',
'strategic', 'members', 'today', 'changes', 'Discreet', 'Cheap', 'differ',
'Finder', 'articles', 'thing', 'release', 'retirement', 'Hydrocodone', 'either',
'Accept', 'drunk', 'EXPRESS', 'February', 'yesterday', 'information', 'PenisEn',
'Codeine', 'Germany', 'hotel', 'notification', 'Magazine', 'starting', 'credit',
'Thank', 'cards', 'financial', 'Cartier', 'eEnhancement', 'based', 'things',
'aRolexBvlgari', 'needed', 'thank', 'nline', 'update', 'mathematics', 'There',
'welcome', 'includes', 'thought', 'business', 'fundamental', 'doing', 'leaves',
'CHECK', 'Vivek', 'announcement', 'exhibit', 'requested', 'changing', 'easily',
'upload', 'creation', 'mailing', 'Carlo', 'specifications', 'contact',
'Thailand', 'dusty', 'Groups', 'Cards', 'Mandarin', 'Wilmott', 'least',
'programming', 'answer', 'located', 'moderately', 'Julius', 'working',
'reliever', 'hangzhou', 'ideas', 'launch', 'Series', 'through', 'Chinese',
'Perhaps', 'share', 'Online', 'assigning', 'pretty', 'status', 'selected',
'required', 'prepared', 'Louis', 'Tiffany', 'StoreDetailView', 'FreeViagra',
'example', 'Check', 'quantitative', 'BrandViagra', 'Since', 'mandatory',
'listed', 'going', 'Discount', 'Doctor', 'works', 'color', 'featured',
'automatically', 'Brands', 'Prices', 'Mandelbrot', 'expertise', 'province',
'NVIDIA', 'below', 'possible', 'scenic', 'herbal', 'money', 'received',
'Sounds', 'extended', 'network', 'Could', 'approach', 'volume', 'there',
'improving', 'encourage', 'capabilities', 'Speedpost', 'those', 'owner',
'jqplot', 'Guaranteeed', 'automatic', 'train', 'softwares', 'Today', 'concise',
'ready', 'hours', 'Please', 'Benoit', 'window', 'survive', 'Sites',
'Phentermin', 'grounds', 'individual', 'being', 'behind', 'finance', 'WILSON',
'Kerry', 'effective', 'ation', 'heard', 'incoming', 'signed', 'decision',
'storage', 'access', 'competitive', 'development', 'pricing', 'fractal',
'focusing', 'Incredib', 'Haloney', 'lined', 'wrote', 'Superb', 'Methods',
'might', 'Hommies', 'Zolpidem', 'longer', 'Whybrew', 'Accepted', 'Jocelyn',
'inches', 'talked', 'products', 'chance', 'Wholesale', 'roofer',
'NoPrescription', 'school', 'Express', 'NaturalPenisEnhancement', 'style',
'recieve', 'focus', 'Enjoy', 'issues', 'order', 'model', 'enjoy', 'source',
'income', 'questions', 'great', 'management', 'narcotic', 'Cheers', 'Tokyo',
'pavilion', 'SciFinance', 'Worldwide', 'yourPenis', 'Bargains', 'number',
'dozen', 'Brand', 'argement', 'Credit', 'Increase', 'winter', 'glimpse',
'Regards', 'accept', 'generation', 'definitely', 'Trusted', 'logged', 'Windows',
'Canadian', 'follow', 'class', 'doggy', 'BiggerPenis', 'PERMANANTLY',
'customized', 'Approved', 'inconvenience', 'online', 'night', 'Watson',
'google', 'message', 'these', 'brands', 'right', 'supporting', 'commented',
'latest', 'Giants', 'computing', 'connection', 'email', 'program', 'courier',
'foaming', 'Everything', 'Professional', 'Order', 'moderate', 'thread',
'inspired', 'nature', 'writing', 'sites', 'sophisticated', 'Reply', 'jquery',
'Peter', 'cannot', 'files', 'coast', 'Amazing', 'advocate', 'placed',
'professional', 'Strategy', 'October', 'Experts', 'analgesic', 'significantly',
'station', 'opioid', 'important', 'specifically', 'Delivery', 'Photoshop',
'forum', 'Stepp', 'support', 'think', 'often', 'wednesday', 'lunch',
'Methylmorphine', 'could', 'gains', 'Instead', 'Adobe', 'MoneyBack',
'Thirumalai', 'invitation', 'designed', 'features', 'Python', 'Success',
'transformed', 'Eugene', 'insights', 'Private', 'knocking', 'Google', 'Gucci',
'FEDEX', 'functionalities', 'FedEx', 'thousand', 'brained', 'phone', 'Arvind']
```

[7]:

```
 # Réduire la taille en gardant les mots qui apparaissent plus de CST fois
 def vocabListFr(vocabList, dataset):
    vocabFr = [0]*len(vocabList)
    for doc in dataset:
        for word in doc:
            if word in vocabList:
                vocabFr[vocabList.index(word)] += 1
    return vocabFr

vocabfr = vocabListFr(vocabList,docList) # Create vocabulary
print('Feature vector size : ',len(vocabfr))
print('max : ',np.max(vocabfr),', min : ', np.min(vocabfr),'mediane : ',np.
  ↪median(vocabfr))
CST = 3;
count = len([val for val in vocabfr if val > CST])
print('Nomre de valeurs supérieures à ',CST,' : ', count)

vocabListbis=[]
for word in vocabList:
    if vocabfr[vocabList.index(word)] > CST:
        vocabListbis.append(word)
```

```

print('Feature vector size : ',len(vocabListbis))
print(vocabListbis)
```

```
Feature vector size :  464
max :  14 , min :  1 mediane :  1.0
Nomre de valeurs supérieures à  3  :  54
Feature vector size :  54
['Quality', 'Watches', 'pills', 'Experience', 'Drugs', 'length', 'group',
'LinkedIn', 'store', 'control', 'endorsed', 'increase', 'Explosive',
'HardErecetions', 'ofEjacu', 'Natural', 'Percocet', 'BetterEjacu',
'intenseOrgasns', 'thickness', 'about', 'Codeine', 'answer', 'Online', 'status',
'going', 'Doctor', 'Mandelbrot', 'herbal', 'volume', 'there', 'Phentermin',
'WILSON', 'ation', 'Incredib', 'wrote', 'inches', 'order', 'model', 'narcotic',
'yourPenis', 'Increase', 'PERMANANTLY', 'google', 'commented', 'email',
'Everything', 'Peter', 'files', 'Amazing', 'support', 'gains', 'designed',
'Google']
```

Test (le résultat est différent d'un test à l'autre, du fait du choix alétoire des 10 exemples de la base Test)

[8]:
```
#Example
doc = open('./email/nspam/17.txt').read()
print(doc)
print(textParse(doc))
print(bagOfWords2VecMN(vocabListbis, textParse(doc)))
```

```
Benoit Mandelbrot 1924-2010

Benoit Mandelbrot 1924-2010

Wilmott Team

Benoit Mandelbrot, the mathematician, the father of fractal mathematics, and
advocate of more sophisticated modelling in quantitative finance, died on 14th
October 2010 aged 85.

Wilmott magazine has often featured Mandelbrot, his ideas, and the work of
others inspired by his fundamental insights.

You must be logged on to view these articles from past issues of Wilmott
Magazine.
['Benoit', 'Mandelbrot', 'Benoit', 'Mandelbrot', 'Wilmott', 'Benoit',
'Mandelbrot', 'mathematician', 'father', 'fractal', 'mathematics', 'advocate',
'sophisticated', 'modelling', 'quantitative', 'finance', 'October', 'Wilmott',
'magazine', 'often', 'featured', 'Mandelbrot', 'ideas', 'others', 'inspired',
'fundamental', 'insights', 'logged', 'these', 'articles', 'issues', 'Wilmott',
'Magazine']
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0]
```

[9]:

```python
# Create test set
trainingSet = list(range(50)); testSet=[]
for i in range(10):
    randIndex = int(random.uniform(0,len(trainingSet)))
    testSet.append(trainingSet[randIndex])
    del(trainingSet[randIndex])

# Préparer les entrées de trainNB0
trainMat=[]; trainClasses = []
for docIndex in trainingSet:
    x=bagOfWords2VecMN(vocabListbis, docList[docIndex])
    trainMat.append(x)
    trainClasses.append(classList[docIndex])

# Apprentissage
p0V,p1V,pSpam = trainNB0(array(trainMat),array(trainClasses))
```

[10]:

```python
# Classer les sxemples Tests
errorCount = 0
for docIndex in testSet:
    wordVector = bagOfWords2VecMN(vocabListbis, docList[docIndex])
    docClass = classifyNB(array(wordVector),p0V,p1V,pSpam)
    if  docClass != classList[docIndex] :
        errorCount += 1
        print("\n\nError : ", errorCount)
        if not classList[docIndex] :
            print("-- ( Non spam classé comme spam.) -- \n")
        else:
            print("-- ( Spam classé comme non spam.) -- \n")
        print(fullText[docIndex])

print('\n Misslassification rate : ',float(errorCount)/len(testSet));
```

```


Error :  1
-- ( Non spam classé comme spam.) --

Yay to you both doing fine!

I'm working on an MBA in Design Strategy at CCA (top art school.)  It's a new
program focusing on more of a right-brained creative and strategic approach to
management.  I'm an 1/8 of the way done today!
```

```
Error :  2
-- ( Spam classé comme non spam.) --

Percocet 10/625 mg withoutPrescription 30 tabs - $225!
Percocet, a narcotic analgesic, is used to treat moderate to moderately
SeverePain
Top Quality, EXPRESS Shipping, 100% Safe & Discreet & Private.
Buy Cheap Percocet Online

 Misslassification rate :  0.2
```