CS777

Big Data Analytics

Spring 2021

Professor Kia Teymourian

Assignment 6

Kunfei Chen

U15575304

Introduction

This project aims to apply the MCMC algorithm to achieve the clustering of text documents.

Task 1 – MCMC Algorithm

Gvien k document clusters, the generative process for n documents is shown in Figure1.

Figure 1

$\pi \sim Dirichlet(\alpha)$
for $j = 1$ to k **do**
  $\mu_j \sim Dirichlet(\beta)$
end for
for $i = 1$ to n **do**
  $c_i \sim Categorical(\pi)$
  $x_i \sim Multinomial(\mu_{c_i}, l_i)$
end for

According to this process, the correspond probability dense function is derived as shown in Figure 2.

Figure 2

$$\text{Dirichlet}(\pi|\alpha) \times \prod_j \text{Dirichlet}(\mu_j|\beta) \times \prod_i \text{Categorical}(c_i|\pi)\ \text{Multinominal}(x_i|\mu_{c_i})\ \text{Multinominal}(x_i|l_i)$$

Task 2

In this task, the input documents are transofrmed into count vectors in which each vector has 20000 entries. All words are converted into lowercase so that the same word with different capitalization are matched and can be count up as same word (See Figure 3).

Figure 3

```
sc = SparkContext(appName="A6")
lines = sc.textFile(input_dir)
header_text = lines.map(lambda x: (x[x.index('id=')+4:x.index(' url=')], x[x.index('">')+2:][:-8].lower()))
regex = re.compile('[^a-zA-Z]')
def get_words(input_val):
    result = []
    words = regex.sub(' ', input_val).split()
    for w in words:
        if len(w) > 2:
            result.append(w)
    return result
header_words = header_text.map(lambda x: (x[0], get_words(x[1])))
top_size = 20000
sorted_words = header_words.flatMap(lambda x: x[1]).map(lambda x: (x,1)).reduceByKey(add).takeOrdered(top_size, key = lambda x: -x[1])
top_words = []
for each in sorted_words:
    top_words.append(each[0])
def countWords(input_val):
    header = input_val[0]
    words = input_val[1]
    numwords = np.zeros(top_size)
    count = 0
    for w in words:
        if w in top_words:
            count += 1
            idx = top_words.index(w)
            numwords[idx] += 1
    return (header, numwords, count)
result = header_words.map(countWords)
result.cache()
```

In addition, for the document “20 newsgroups/comp.graphics/37261”, its number of times that each of the 100 most common dictionary words are identified (See Figure 4). The words list is shown in Figure 5.

Figure 4

```
target = result.filter(lambda x: x[0].split("/")[-1][0:] == "37261").collect()
sort_index = list(np.argsort(target[0][1]))
list.reverse(sort_index)

target_100_words = []
for i in range(177):
    index = sort_index[i]
    target_100_words.append((top_words[index], target[0][1][index]))
print("The 100 most common words appear in document 20 newsgroups/comp.graphics/37261:")
for w in target_100_words:
    print(w)
```

Figure 5

```
The 100 most common words appear in document 20 newsgroups/comp.graphics/37261:
(u'and', 12.0)
(u'navy', 11.0)
(u'the', 8.0)
(u'lipman', 7.0)
(u'for', 6.0)
(u'presentations', 6.0)
(u'seminar', 6.0)
(u'scientific', 5.0)
(u'virtual', 5.0)
(u'visualization', 5.0)
(u'reality', 5.0)
(u'will', 4.0)
(u'robert', 4.0)
(u'mil', 4.0)
(u'presentation', 4.0)
(u'one', 3.0)
(u'oasys', 3.0)
(u'bethesda', 3.0)
(u'should', 3.0)
(u'maryland', 3.0)
(u'center', 3.0)
(u'authors', 2.0)
(u'length', 2.0)
(u'related', 2.0)
(u'work', 2.0)
(u'information', 2.0)
(u'david', 2.0)
(u'warfare', 2.0)
(u'surface', 2.0)
(u'call', 2.0)
(u'carderock', 2.0)
```

Task 3

Then the MCMC algorithm derived in Task 1 is implemented for 200 iterations on the 20 newsgroups data, with the goal of learning the π, μj , and ci values. As shown in Figure 6, after 200 iterations, the 50 most important words with the highest probability in each of the 20 learned mixture components are identified.

Figure 6

```
50 important_words for category 0 : [u'the', u'from', u'edu', u'that', u'for', u'subject', u'you', u'and', u'date', u'lines', u'apr', u'gmt', u'about
50 important_words for category 1 : [u'the', u'and', u'that', u'for', u'you', u'from', u'this', u'are', u'with', u'have', u'edu', u'subject', u'apr',
50 important_words for category 2 : [u'the', u'and', u'that', u'scx', u'chz', u'for', u'kuwait', u'from', u'not', u'this', u'r1k', u'was', u'uww', u
50 important_words for category 3 : [u'the', u'and', u'that', u'you', u'for', u'not', u'from', u'are', u'have', u'this', u'they', u'was', u'with', u
50 important_words for category 4 : [u'the', u'for', u'and', u'from', u'edu', u'apr', u'subject', u'lines', u'date', u'gmt', u'with', u'sale', u'you'
50 important_words for category 5 : [u'the', u'that', u'and', u'this', u'you', u'was', u'stephanopoulos', u'president', u'for', u'with', u'have', u'no
50 important_words for category 6 : [u'the', u'that', u'and', u'you', u'not', u'are', u'this', u'for', u'from', u'have', u'but', u'god', u'with', u'
50 important_words for category 7 : [u'the', u'and', u'pit', u'period', u'tor', u'from', u'edu', u'det', u'bos', u'ch1', u'pts', u'for', u'st1', u'
50 important_words for category 8 : [u'max', u'giz', u'bhj', u'the', u'bxn', u'and', u'qax', u'you', u'from', u'edu', u'that', u'com', u'appears', u
50 important_words for category 9 : [u'edu', u'from', u'subject', u'lines', u'date', u'apr', u'for', u'gmt', u'the', u'and', u'distribution', u'this',
50 important_words for category 10 : [u'the', u'and', u'for', u'from', u'edu', u'subject', u'date', u'lines', u'that', u'apr', u'gmt', u'you', u'have
50 important_words for category 11 : [u'the', u'and', u'for', u'that', u'you', u'this', u'from', u'are', u'with', u'not', u'have', u'can', u'will',
50 important_words for category 12 : [u'the', u'for', u'edu', u'and', u'from', u'you', u'subject', u'are', u'apr', u'gmt', u'lines', u'date', u'com',
50 important_words for category 13 : [u'the', u'and', u'edu', u'from', u'for', u'that', u'lines', u'subject', u'apr', u'date', u'you', u'gmt', u'this
50 important_words for category 14 : [u'the', u'and', u'that', u'was', u'for', u'you', u'they', u'have', u'from', u'but', u'with', u'this', u'not',
50 important_words for category 15 : [u'the', u'and', u'that', u'for', u'from', u'this', u'you', u'apr', u'subject', u'com', u'date', u'lines', u'edu
50 important_words for category 16 : [u'air', u'the', u'subject', u'lines', u'from', u'date', u'ahf', u'kjz', u'khf', u'gmt', u'region', u'apr', u'ne
50 important_words for category 17 : [u'the', u'and', u'for', u'that', u'from', u'with', u'you', u'have', u'this', u'date', u'edu', u'lines', u'subje
50 important_words for category 18 : [u'from', u'edu', u'date', u'for', u'subject', u'lines', u'gmt', u'and', u'apr', u'ndet', u'loop', u'the', u'has
50 important_words for category 19 : [u'the', u'that', u'you', u'and', u'frank', u'not', u'can', u'are', u'for', u'objective', u'from', u'writes', u'
```

Task 4

Finally, in order to evaluate the accuracy f the learning algorithm, the number of documents to each of the 20 learned mixture components are printed. In addition, the percentage of the top three real categories for documents assigned to each cluster are also identified (See Figure 7). It can be seen that generally each component has a high prevalence of several related categories with the lowest percentage of 0.16 and the highest percentage of 0.61. In addition, the top categories in the 20 clusters are roughly correspond to the 20 categories present in the data.

Figure 7

```
Total count of cluster 0: 125
Top 3 topics in cluster 0 : [[u'misc.forsale', 0.176], [u'rec.sport.baseball', 0.128], [u'talk.religion.misc', 0.104]]
Total count of cluster 1: 3369
Top 3 topics in cluster 1 : [[u'rec.motorcycles', 0.2716], [u'rec.autos', 0.247], [u'sci.space', 0.2188]]
Total count of cluster 2: 54
Top 3 topics in cluster 2 : [[u'sci.med', 0.2778], [u'rec.sport.hockey', 0.2222], [u'talk.politics.misc', 0.1111]]
Total count of cluster 3: 3036
Top 3 topics in cluster 3 : [[u'talk.politics.mideast', 0.2955], [u'talk.politics.guns', 0.2915], [u'talk.politics.misc', 0.2516]]
Total count of cluster 4: 636
Top 3 topics in cluster 4 : [[u'misc.forsale', 0.6148], [u'sci.med', 0.1164], [u'sci.electronics', 0.0456]]
Total count of cluster 5: 54
Top 3 topics in cluster 5 : [[u'talk.politics.misc', 0.3704], [u'sci.med', 0.1852], [u'talk.politics.guns', 0.1481]]
Total count of cluster 6: 2485
Top 3 topics in cluster 6 : [[u'soc.religion.christian', 0.3646], [u'alt.atheism', 0.334], [u'talk.religion.misc', 0.229]]
Total count of cluster 7: 153
Top 3 topics in cluster 7 : [[u'rec.sport.hockey', 0.5621], [u'alt.atheism', 0.1176], [u'talk.religion.misc', 0.0784]]
Total count of cluster 8: 99
Top 3 topics in cluster 8 : [[u'talk.politics.misc', 0.2828], [u'talk.politics.mideast', 0.1818], [u'talk.religion.misc', 0.1515]]
Total count of cluster 9: 361
Top 3 topics in cluster 9 : [[u'misc.forsale', 0.2188], [u'comp.windows.x', 0.1108], [u'comp.graphics', 0.0859]]
Total count of cluster 10: 1760
Top 3 topics in cluster 10 : [[u'comp.graphics', 0.2938], [u'comp.os.ms-windows.misc', 0.2665], [u'comp.windows.x', 0.1994]]
Total count of cluster 11: 2329
Top 3 topics in cluster 11 : [[u'sci.crypt', 0.3366], [u'comp.windows.x', 0.2001], [u'sci.med', 0.1498]]
Total count of cluster 12: 44
Top 3 topics in cluster 12 : [[u'misc.forsale', 0.1591], [u'talk.politics.misc', 0.1136], [u'comp.sys.ibm.pc.hardware', 0.1136]]
Total count of cluster 13: 737
Top 3 topics in cluster 13 : [[u'rec.sport.hockey', 0.46], [u'rec.sport.baseball', 0.1506], [u'sci.electronics', 0.0651]]
Total count of cluster 14: 1551
Top 3 topics in cluster 14 : [[u'rec.sport.baseball', 0.5016], [u'rec.sport.hockey', 0.3217], [u'sci.med', 0.109]]
Total count of cluster 15: 255
Top 3 topics in cluster 15 : [[u'sci.crypt', 0.2863], [u'comp.windows.x', 0.1059], [u'rec.autos', 0.1059]]
Total count of cluster 16: 34
Top 3 topics in cluster 16 : [[u'sci.med', 0.1765], [u'soc.religion.christian', 0.1765], [u'talk.politics.guns', 0.0882]]
Total count of cluster 17: 2659
Top 3 topics in cluster 17 : [[u'comp.sys.ibm.pc.hardware', 0.3238], [u'comp.sys.mac.hardware', 0.3129], [u'comp.os.ms-windows.misc', 0.1392]]
Total count of cluster 18: 101
Top 3 topics in cluster 18 : [[u'misc.forsale', 0.2277], [u'comp.windows.x', 0.1089], [u'soc.religion.christian', 0.0891]]
Total count of cluster 19: 155
Top 3 topics in cluster 19 : [[u'talk.religion.misc', 0.4258], [u'alt.atheism', 0.3742], [u'sci.electronics', 0.0581]]
```