Programming Assignent 2

2018.05.11

Introduction

- **You are required to implement a text classification model by PLSA.**
- **We will give you a subset of 20 News Groups and some seed words of each class, and your task is to classify documents according to the given seed words (It's an unsupervised learning).**

Kaggle

- https://www.kaggle.com/t/

  d78b100c24ef4de18eb6391b6feea9f3

Data

- **doc.csv**
- **group.csv**
- **sample_submission.csv**
- **stop.txt**

doc.csv

```
doc_id,content
0,"From: Mamatha Devineni Ratnam <mr47+@andrew.cmu.edu>\nSubject: Pens fans reactions\nOrganization: Post Office,
Carnegie Mellon, Pittsburgh, PA\nLines: 12\nNNTP-Posting-Host: po4.andrew.cmu.edu\n\n\n\nI am sure some bashers of
Pens fans are pretty confused about the lack\nof any kind of posts about the recent Pens massacre of the Devils.
Actually,\nI am  bit puzzled too and a bit relieved. However, I am going to put an end\nto non-PIttsburghers' relief
with a bit of praise for the Pens. Man, they\nare killing those Devils worse than I thought. Jagr just showed you
why\nhe is much better than his regular season stats. He is also a lot\nfo fun to watch in the playoffs. Bowman should
let JAgr have a lot of\nfun in the next couple of games since the Pens are going to beat the pulp out of Jersey anyway.
 I was very disappointed not to see the Islanders lose the final\nregular season game.          PENS RULE!!!\n\n"
1,"From: mblawson@midway.ecn.uoknor.edu (Matthew B Lawson)\nSubject: Which high-performance VLB video card?\nSummary:
Seek recommendations for VLB video card\nNntp-Posting-Host: midway.ecn.uoknor.edu\nOrganization: Engineering Computer
Network, University of Oklahoma, Norman, OK, USA\nKeywords: orchid, stealth, vlb\nLines: 21\n\n  My brother is in the
market for a high-performance video card that supports\nVESA local bus with 1-2MB RAM.  Does anyone have
suggestions/ideas on:\n\n  - Diamond Stealth Pro Local Bus\n\n  - Orchid Farenheit 1280\n\n  - ATI Graphics Ultra
Pro\n\n  - Any other high-performance VLB card\n\n\nPlease post or email.  Thank you!\n\n  - Matt\n\n-- \n    |
Matthew B. Lawson <------------> (mblawson@essex.ecn.uoknor.edu)  |   \n  --+-- ""Now I, Nebuchadnezzar, praise and
exalt and glorify the King  --+-- \n    |   of heaven, because everything he does is right and all his ways  |   \n
|   are just."" - Nebuchadnezzar, king of Babylon, 562 B.C.           |   \n"
2,"From: hilmi-er@dsv.su.se (Hilmi Eren)\nSubject: Re: ARMENIA SAYS IT COULD SHOOT DOWN TURKISH PLANES (Henrik)
\nLines: 95\nNntp-Posting-Host: viktoria.dsv.su.se\nReply-To: hilmi-er@dsv.su.se (Hilmi Eren)\nOrganization: Dept. of
Computer and Systems Sciences, Stockholm University\n\n\n\n\n|>The student of ""regional killings"" alias Davidian (
not the Davidian religios sect) writes:\n\n\n|>Greater Armenia would stretch from Karabakh, to the Black Sea, to the\n|
>Mediterranean, so if you use the term ""Greater Armenia"" use it with care.\n\n\n  Finally you said what you dream
about. Mediterranean???? That was new....\n The area will be ""greater"" after some years, like your ""holocaust""
numbers......\n\n\n\n\n|>It has always been up to the Azeris to end their announced winning of Karabakh \n|>by
removing the Armenians! When the president of Azerbaijan, Elchibey, came to \n|>power last year, he announced he would
be be ""swimming in Lake Sevan [in \n|>Armeniaxn] by July"".\n     *****\n Is't July in USA now????? Here in Sweden
it's April and still cold.\n   Or have you changed your calendar???\n\n\n|>Well, he was wrong! If Elchibey is going to
shell the \n|>Armenians of Karabakh from Aghdam, his people will pay the price! If Elchibey \n
    **********\n|>is going to shell Karabakh from Fizuli his people will pay the price! If \n
        **********\n|>Elchibey thinks he can get away with bombing Armenia from the hills of \n|>Kelbajar, his
people will pay the price. \n                **********\n\n\n   NOTHING OF THE MENTIONED IS TRUE, BUT LET SAY
IT's TRUE.\n  \n  SHALL THE AZERI WOMEN AND CHILDREN GOING TO PAY THE PRICE WITH\n
```

group.csv

```
class_id,class_name,relevant_words
0,alt.atheism,atheists
1,comp.graphics,image
2,comp.os.ms-windows.misc,windows
3,comp.sys.ibm.pc.hardware,drive
4,comp.sys.mac.hardware,mac
5,misc.forsale,sale
6,rec.autos,car
7,rec.motorcycles,bike
8,rec.sport.baseball,baseball
9,rec.sport.hockey,hockey
10,sci.crypt,encryption
11,sci.med,medical
12,sci.space,space
13,soc.religion.christian,christians
14,talk.politics.guns,gun
15,talk.politics.mideast,israel
16,talk.politics.misc,president
```

sample_submission.csv

```
doc_id,class_id
0,0
1,0
2,0
3,0
4,0
5,0
6,0
7,0
8,0
9,0
10,0
11,0
12,0
13,0
14,0
15,0
16,0
17,0
18,0
19,0
20,0
21,0
22,0
23,0
24,0
```

Information

- 我們提供*16245*個文件，您在實作完*PLSA*之後，需要對其標上*label*
- 一樣的規定，我們不限制您的程式語言，一樣要求在實作*PLSA*時，不得使用相關套件

Program I/O

- 您一樣需要交上兩個*shell script*

  1. compile.sh

  2. execute.sh -option_1 value_1 -option_2 value_2 …

Program execution detail

```
SYNOPSIS:
    ./execute.sh [-e] [-b] -d doc.csv -g group.csv -o output.csv

OPTIONS:
    -e
        If specified, use the dictionary you made to classify the document
    -b
        If specified, run your best version of your program
    -d doc.csv
        The doc.csv
    -g group.csv
        The group.csv
    -o output.csv
        The output path of your classification
```

Evaluation

- 簡簡單單，就是*accuracy*
- 每個文章皆只有一類

Score(15%)

- **(3%)Programming - PLSA**
- **(2%)Programming - make a dictionary and use it**
- **(2%)Programming - baseline (0.4699)**
- **(8%)Report**

Report

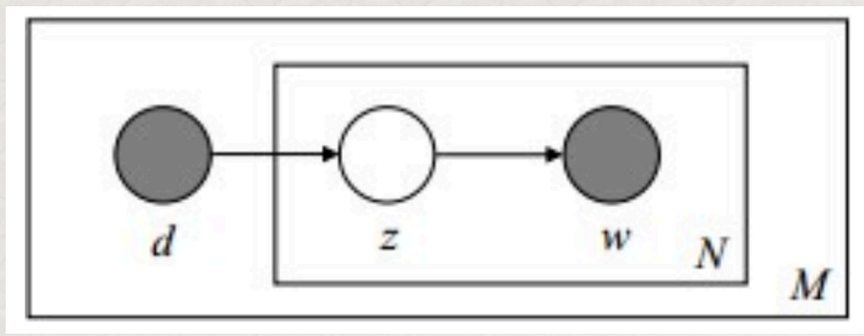
- (4%)推導PLSA公式(**detail later**)
- (2%)說明你分類的方法(**map hidden topic to class**)
- (1%)比較不同**topic number**的影響
- (1%)比較是否使用字典的影響

PLSA

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$

$$P(d_i, w_j) = P(d_i)P(w_j|d_j)$$

$$\theta = P(z|d), P(w|z)$$



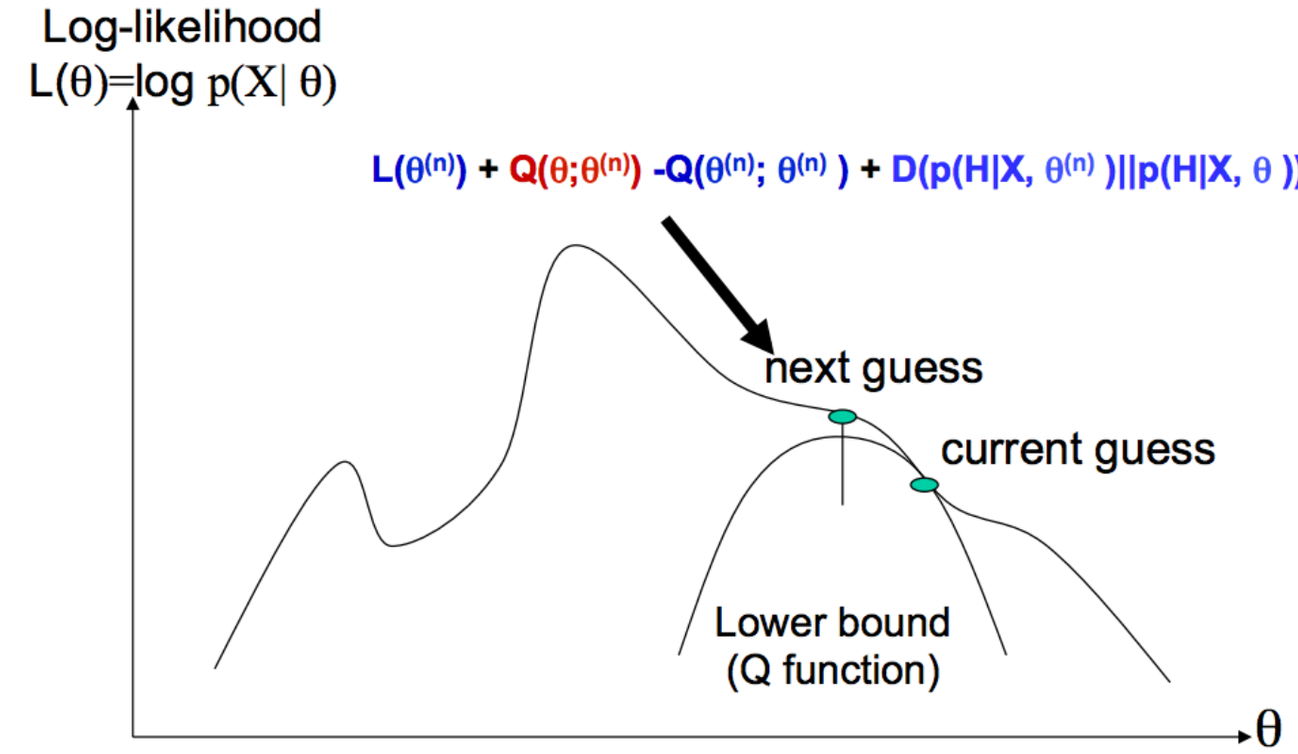
Log likelihood

$$L(\theta) = P(D, W; \theta)$$
$$= \Pi_{d \in D} \Pi_{w \in W} P(d, w)^{n(d,w)}$$

- **(1%)** $\log(L(\theta)) = ?$
- **請用** $n(), P(), d, w, z, i, j, k, D, W, K$ **表示**

EM algorithm

Another way of looking at EM



E-step = computing the lower bound
M-step = maximizing the lower bound

尋找Q function

$$\log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) = \log \sum_{k=1}^{K} Q(z_k) \frac{P(w_j|z_k)P(z_k|d_i)}{Q(z_k)}$$

- 根據 *Jessen* 不等式

$$\log \sum_{k=1}^{K} Q(z_k) \frac{P(w_j|z_k)P(z_k|d_i)}{Q(z_k)} \geq \sum_{k=1}^{K} Q(z) \log(\frac{P(w_j|z_k)P(z_k|d_i)}{Q(z_k)})$$

$$Q(z_k) \propto P(w_j|z_k)P(z_k|d_i) \qquad \sum_{k=1}^{K} Q(z_k) = 1$$

尋找Q function

- **(1 %)** $Q(z_k) = ?$

請用 $P(), w, z, d, i, j, k, K$ 表示

推導EM

- **(2 %)**試問*M step*為何**(P(w | z), P(z | d))***?***

  Hint: Apply Lagrange Multiplier

Baseline

- **Preprocess**
  1. 移除所有標點符號
  2. 用nltk.tokenize.wordpunct_tokenize()
  3. 把字變小寫
  4.用stop.txt去除stop words
  5. 只用word count > 200的字來實作

- **Train PLSA**
  1. initialize probability tables uniformly
  2. topic size = 50, iterations = 100

- **Classification**
  find predicted class c = argmax_c score(d, c)
  score(d, c) = mean of [p(w | d) for w in c (w是class c的seed word或是該類別字典的字)]
  p(w | d) = sum_z (p(w | z) * p(z | d))

Bonus

- **Top-3 ranking at public scoreboard**

  1% - 1~3 place

- **Top-5 ranging at private scoreboard**

  2% - 1~2 place

  1% - 3~5 place

Submission

- **R059XXXXXX.zip**
  - **R059XXXXXX (directory)**
    - **report.pdf**
    - **compile.sh**
    - **execute.sh**
    - **source/ (directory)**

Rules

- **Kaggle:**

  display name: 學號_ID(R05922032_AABDDAACCD)

  5 times submissions a day

  2 entries for private score

- **Deadline:**

  Kaggle: 2018/06/01 24:00:00 (GMT +8)

  Report: 2018/06/02 24:00:00 (GMT +8)

- **Late policy 10% a day**