EE412 Foundation of Big Data Analytics, Fall 2021

HW3

20170616 정희진

1. Link Analysis

(a) Solve the following problems, which are based on the exercises in the Mining of Massive Datasets 3rd edition (MMDS) textbook.

Exercise 5.1.2

문제를 풀기 위한 식과 코드는 다음과 같다.

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0.8 \begin{bmatrix} \frac{1}{3} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} + 0.2 \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

```
import numpy as np
beta = 0.8
M = np.array([[1/3, 1/2, 0], [1/3, 0, 1/2], [1/3, 1/2, 1/2]])
v = np.array([1/3, 1/3, 1/3])
w = np.array([1/3, 1/3, 1/3])

for i in range(40):
  v = beta*M@v+(1-beta)*w
print(v)

for i in range(10):
  v = beta*M@v+(1-beta)*w
print(v)
```

위의 코드에서 두 개의 for loop문이 있는데, 첫 번째는 PageRanks equation을 40번 계산하여 v를 구한 것이고, 두 번째는 첫 번째에서 구한 v를 가지고 10번 더 계산하여 v를 구한 것이다. 두 결과값이 차이가 없다면 마지막으로 나온 v가 PageRank라고 할 수 있다.

코드의 결과는 다음과 같다.

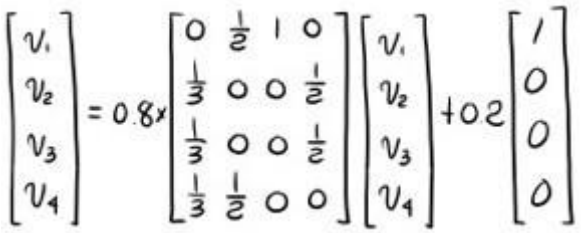
```
[0.25925926 0.30864198 0.43209877]
[0.25925926 0.30864198 0.43209877]
```

Print된 두 개의 결과에 차이가 없는 것으로 보아, 각 page의 PageRank는 아래와 같다.

Page	PageRank
a	0.25925926
b	0.30864198
c	0.43209877

Exercise 5.3.1

(a) 문제를 풀기 위한 식과 코드는 다음과 같다.



```
beta = 0.8
M = np.array([[0, 1/2, 1, 0], [1/3, 0, 0, 1/2], [1/3, 0, 0, 1/2], [1/3,
 1/2, 0, 0]])
v = np.array([1, 0, 0, 0])
w = np.array([1, 0, 0, 0])

for i in range(40):
  v = beta*M@v+(1-beta)*w
print(v)

for i in range(10):
  v = beta*M@v+(1-beta)*w
print(v)
```

마찬가지로 PageRanks equation을 40번, 50번 계산하여 얻은 v를 프린트 하였다. 두 결과값이 차이가 없다면 마지막으로 나온 v가 PageRank라고 할 수 있다.

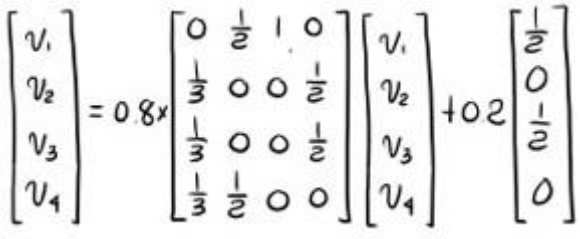
코드의 결과는 다음과 같다.

```
[0.42857143 0.19047619 0.19047619 0.19047619]
[0.42857143 0.19047619 0.19047619 0.19047619]
```

Print된 두 개의 결과에 차이가 없는 것으로 보아, 각 page의 PageRank는 아래와 같다.

Page	PageRank
a	0.42857143
b	0.19047619
c	0.19047619
d	0.19047619

(b) 문제를 풀기 위한 식과 코드는 다음과 같다.



```
beta = 0.8
M = np.array([[0, 1/2, 1, 0], [1/3, 0, 0, 1/2], [1/3, 0, 0, 1/2], [1/3,
 1/2, 0, 0]])
v = np.array([1/2, 0, 1/2, 0])
w = np.array([1/2, 0, 1/2, 0])

for i in range(40):
  v = beta*M@v+(1-beta)*w
print(v)

for i in range(10):
  v = beta*M@v+(1-beta)*w
print(v)
```

마찬가지로 PageRanks equation을 40번, 50번 계산하여 얻은 v를 프린트 하였다. 두 결과값이 차이가 없다면 마지막으로 나온 v가 PageRank라고 할 수 있다.

코드의 결과는 다음과 같다.

```
[0.38571429 0.17142857 0.27142857 0.17142857]
[0.38571429 0.17142857 0.27142857 0.17142857]
```

Print된 두 개의 결과에 차이가 없는 것으로 보아, 각 page의 PageRank는 아래와 같다.

Page	PageRank
a	0.38571429
b	0.17142857
c	0.27142857
d	0.17142857

(b) Implement the PageRank algorithm using Spark.

```
263     0.00216
537     0.00212
965     0.00206
243     0.00197
255     0.00194
285     0.00193
16      0.00191
126     0.00190
747     0.00190
736     0.00189
```

2. Mining Social-Network Graphs

(a) Solve the following problems, which are based on the exercises in the MMDS textbook.

Exercise 10.3.2

다음 두가지 조건을 만족하면서 maximal pair을 찾아야 한다

① $S \times t \leq n \times d$ (nd는 edge의 총수)

② $t \leq s \leq n$

(a) $St \leq 100$, $t \leq s \leq 20$

t	S
5	20
6	16
7	14
8	12
9	11
10	10

(b) $St \leq 30000$, $t \leq s \leq 200$

S	200	198	197	196	194	193	192	191	189	188	187	186	185	184	182	181
t	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165

S	180	179	178	177	176	175	174	173
t	166	167	168	169	170	171	172	173

Exercise 10.5.2

(a) $P_{wx} = P_c$ $P_{xy} = \epsilon$

$P_{wy} = \epsilon$ $P_{xz} = \epsilon$ (이때 ϵ는 매우 작은수)

$P_{wz} = \epsilon$ $P_{yz} = P_D$

likelihood $= P_c \epsilon^2 (1-\epsilon)^3 P_D = P_c P_D (\epsilon - \epsilon^3)^2$

$P_c = 1$, P_D이 클 때 likelihood가 maximum이 된다

따라서 C 커뮤니티에 있는 멤버끼리 100% edge가 있고

D 커뮤니티에 있는 멤버끼리 100% edge가 있고

maximum likelihood는 0이다 ($\because \epsilon$가 매우 작으므로)

(b) $P_{wx} = P_c$ $P_{xy} = 1-(1-P_c)(1-P_D) = P_c + P_D - P_c P_D$

$P_{wy} = P_c$ $P_{xz} = P_c + P_D - P_c P_D$

$P_{wz} = P_c$ $P_{yz} = P_c + P_D - P_c P_D$

likelihood $= P_c^2 (1-P_c)(P_c + P_D - P_c P_D)^2 (1 - P_c - P_D + P_c P_D)$

$k^2(1-k)$라는 식에서 미분을 취함

$2k(1-k) - k^2 = 2k - 3k^2 = k(2-3k)$가 된다

$k(2-3k) = 0 \Rightarrow k = 0$ 또는 $\frac{2}{3}$인데

$k=0$일때 0, $k=\frac{2}{3}$일때 $\frac{4}{27}$, $k=1$일 때 0이므로

$k=\frac{2}{3}$일 때 최댓값 $\frac{4}{27}$를 갖는다

위 식에서 $P_c^2(1-P_c)$는 $k = P_c$일때 $k^2(1-k)$와 같으므로

$P_c = \frac{2}{3}$이어야 한다

또한 $(P_c + P_D - P_c P_D)^2 (1 - P_c - P_D + P_c P_D)$는 $k = P_c + P_D - P_c P_D$일 때 $k^2(1-k)$과 같으므로

$P_c + P_D - P_c P_D = \frac{2}{3} + P_D - \frac{2}{3} P_D = \frac{2}{3} + \frac{1}{3} P_D = \frac{2}{3} \Rightarrow P_D = 0$이어야 한다

$P_c = \frac{2}{3}$, $P_D = 0$일 때 위식은 $\frac{4}{27} \times \frac{4}{27} = \frac{16}{729}$이된다

따라서 C 커뮤니티에 있는 멤버끼리 약 66.66%의 확률로 edge가 있고

D 커뮤니티에 있는 멤버끼리 0%의 확률로 edge가 있고

maximum likelihood는 $\frac{16}{729}$ 이다

(b) Implement the algorithm for finding triangles in MMDS Chapter 10.7.2. You will analyze part of the Facebook (now Meta) social network to identify communities

```
3501542
```

3. Large-Scale Machine Learning

(a) Solve the following problems, which are based on the exercises in the MMDS textbook.

Exercise 12.5.3

(a) $f(p) = 1 - \sum_{i=1}^{n} (p_i)^2$

$\frac{\partial f(p)}{\partial p_i} = -2p_i$

$\frac{\partial^2 f(p)}{\partial p_i^2} = -2 < 0$

모든 p_i에 대해, 이차 미분이 음수이므로

GINI impurity는 concave 하다

(b) $f(p) = \sum_{i=1}^{n} p_i \log_2(1/p_i) = \sum_{i=1}^{n} -p_i \log_2 p_i$

$\frac{\partial f(p)}{\partial p_i} = -\log_2 p_i - p_i \times \frac{1}{p_i \ln 2} = -\log_2 p_i - \frac{1}{\ln 2}$

$\frac{\partial^2 f(p)}{\partial p_i^2} = -\frac{1}{p_i \ln 2} < 0$ ($0 \le p_i \le 1$이므로)

모든 p_i에 대해, 이차 미분이 음수이므로

Entropy measure of impurity는 concave 하다

(b) Implement the gradient descent SVM algorithm described in MMDS Chapter 12.3.4 using Python