EE412 Foundation of Big Data Analytics, Fall 2021
HW3

Name: Cao Viet Hai Nam

Student ID: 20200817

Discussion Group (People with whom you discussed ideas used in your answers):

On-line or hardcopy documents used as part of your answers:

Answer to Problem 1:

<u>Exercise 5.1.2.</u>

Code:

```
1.  import numpy as np
2.  beta = 0.8
3.  M = np.array([[1/3,1/2,0],
4.                [1/3,0,1/2],
5.                [1/3,1/2,1/2]])
6.  v = np.array([[1/3], [1/3], [1/3]])
7.  teleport_fac = (1-beta)*np.ones((3, 1)) / 3
8.  n = 50
9.  for i in range(n):
10.     v = beta*M@v + teleport_fac
11. print(v)
```

<u>RESULT:</u>

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

Explanation:

Line 2: initialize beta
Line 3-5: Transition matrix
Line 6: initial vector v will have 1/n for each component
Line 7: Using taxation allow each suffer a small probability (1-beta) to teleport to a random page.
Line 8: Number of iterations (about 50 will ensure for v to converge)
Line 9-10: Iterate and update v.
Line 11: print the result

Exercise 5.3.1:

Code:

```
1.  import numpy as np
2.  beta = 0.8
3.  N = np.array([[0, 1/2, 1, 0 ],
4.                [1/3, 0, 0, 1/2],
5.                [1/3, 0, 0, 1/2],
6.                [1/3,1/2,0, 0 ]])
7.  v1 = np.array([[1/4], [1/4], [1/4], [1/4]])
8.  v2 = np.array([[1/4], [1/4], [1/4], [1/4]])
9.  teleport_fac_A_only = (1-beta)*np.array([[1], [0], [0], [0]])
10. teleport_fac_A_C = (1-beta)*np.array([[1/2], [0], [1/2], [0]])
11. n = 50
12. for i in range(n):
13.     v1 = beta*N@v1 + teleport_fac_A_only
14.     v2 = beta*N@v2 + teleport_fac_A_C
15. print("A only: ")
16. print(v1)
17. print("A and C: ")
18. print(v2)
```

RESULT: (Page tank vector for both case)

```
A only =
[[0.42857143]
 [0.19047619]
 [0.19047619]
 [0.19047619]]
A and C =
[[0.38571429]
 [0.17142857]
 [0.27142857]
 [0.17142857]]
```

Explanation:

Line 2: initialize beta
Line 3-6: Transition matrix
Line 7-8 : initial vector v1, v2 will have 1/n for each component
Line 9, 10: Using taxation allow each suffer a small probability (1-beta) to teleport to teleport set (only A or A and C)
Line 11: Number of iterations (about 50 will ensure for both v to converge)
Line 12-14: Iterate and update v1 and v2
Line 15-18: print the result

Answer to Problem 2

② Mining Social-Network Graphs

Exercise 10.3.2

a) let G be the bipartite graph.

let the "item" be the nodes on one side of G, we call it left side. Assume instance of $K_{s,t}$ has t nodes on the left side. "baskets" corresponds to the nodes on the right side.

let support threshold be s = # nodes that $K_{s,t}$ has on the right side

⇒ frequent itemset of size t and s of the baskets in which all those items appear form $K_{s,t}$

Suppose the degree of i^{th} node on the right side is d_i (size of i^{th} basket)

⇒ this basket contributes to $\binom{d_i}{t}$ itemsets of size t ⇒ total contribution of n nodes on the right is $\sum_i \binom{d_i}{t}$.

Average of d_i is d so this sum is minimized when each d_i is d

⇒ we shall assume that all nodes have the average degree of d.

⇒ total contribution of n nodes on the

Right to the counts of itemsets size t is $n\binom{d}{t}$

\# itemsets size t is $\binom{n}{t}$

$\Rightarrow$ Average count of an itemset of size t is $n\binom{d}{t} / \binom{n}{t}$, this expression must

be at least s if we are to argue that $K_{s,t}$ exists

1) for $n = 20$, $d = 5$

$\Rightarrow s \leq 20\binom{5}{t} / \binom{20}{t}$ ↙ guaranted constrain

- if $t = 1 \Rightarrow s \leq 5 \Rightarrow s$ can be as large as 5.

Note: For this example, we can reconfirm $s = 5$ is maximal by consider the following G where node i^{th} on the left is connect to 5 node (i^{th}; $(i+1)^{th}$; $(i+2)^{th}$; $(i+3)^{th}$; $(i+4)^{th}$) on the Right (if $(i+j)^{th} > 20$ we take $(i+j) \bmod 20$).
We can see that G does not have $K_{6,1}$.

- if $t \geq 2 \Rightarrow s \leq 1{,}053 < t$ (not considered)

Thus, our (t, s) set is $\boxed{\{(1, 5)\}}$

2) FOR $n = 200$, $d = 150$

$\Rightarrow s \leq 200\binom{150}{t} / \binom{200}{t}$ ↙ guaranted constrain

- $t = 1 \Rightarrow s \leq 150 \Rightarrow S_{max} = 150$ (same arguement in part (a).1)
- $t = 2 \Rightarrow s \leq 112,3$. However if the average support for an item set of 2 is 112,3, then it is impossible that all those itemsets have support ≤ 112. Thus, we can be sure that at least one itemset of size 2 has support 113 or more $\Rightarrow K_{113,2}$ exists (*)

  $\Rightarrow S_{max} = 113$.
- $t = 3 \Rightarrow S \leq 83.95 \Rightarrow S_{max} = 84$ (same arguement with (*)).
- $t = 4 \Rightarrow S \leq 62.64 \Rightarrow S_{max} = 63$.
- $t = 5 \Rightarrow S \leq 46.66 \Rightarrow S_{max} = 47$
- $t = 6 \Rightarrow S \leq 34.69 \Rightarrow S_{max} = 35$
- $t = 7 \Rightarrow S \leq 25.75 \Rightarrow S_{max} = 26$
- $t = 8 \Rightarrow S \leq 19.08 \Rightarrow S_{max} = 20$
- $t = 9 \Rightarrow S \leq 14.11 \Rightarrow S_{max} = 15$
- $t = 10 \Rightarrow S \leq 10.41 \Rightarrow S_{max} = 11$

• $t \geqslant 11 \Rightarrow s \leq 7.67 < t \Rightarrow$ not considered

Thus, our (s,t) maximal set is
$\{(1, 150)$; $(2, 113)$; $(3, 84)$; $(4, 63)$; $(5, 47)$
$(6, 35)$; $(7, 26)$; $(8, 20)$; $(9, 15)$; $(10, 11)\}$

* Exercise 10.5.2,

Supposed communities C, D with associated probabilities p_C and p_D.

a) $C = \{w, x\}$; $D = \{y, z\}$.

$p_{wx} = p_C$; $p_{yz} = p_D$

$p_{wy} = \epsilon_1$; $p_{wz} = \epsilon_2$; $p_{xy} = \epsilon_3$; $p_{xz} = \epsilon_4$

The likelihood of the graph is:

$L = p_C \cdot \epsilon_1 \cdot \epsilon_3 \cdot p_D \, (1 - \epsilon_2)(1 - \epsilon_4)$

Note that, ϵ_i is very small so

$L \approx p_C \, \epsilon_1 \, \epsilon_3 \, p_D \leq \boxed{\epsilon_1 \epsilon_3}$. $\leftarrow$ MLE
(a very small number)

$\Rightarrow$ MLE when $p_C = p_D = 1$

b) $C = \{w, x, y, z\}$; $D = \{x, y, z\}$

$p_{wx} = p_{wy} = p_{wz} = p_C$

$$p_{xy} = p_{yz} = p_{xz} = 1 - (1-p_C)(1-p_D)$$
$$= p_C + p_D - p_C p_D$$

The likelihood of the graph is:

$$L = p_C^2 \cdot (p_C + p_D - p_C p_D)^2 (1-p_C)^2 (1-p_D)$$

let $x = 1 - p_C$; $y = 1 - p_D$ $(x, y \in [0, 1])$

$$\Rightarrow L = (1-x)^2 (1-xy)^2 x^2 y$$

$$= 16 \left[\frac{(1-x)}{2} \cdot \frac{(1-x)}{2} \cdot x\right] \left[\frac{(1-xy)}{2} \cdot \frac{(1-xy)}{2} \cdot xy\right]$$

$$\leq 16 \left(\frac{\frac{1-x}{2} + \frac{1-x}{2} + x}{3}\right)^3 \cdot \left(\frac{\frac{1-xy}{2} + \frac{1-xy}{2} + xy}{3}\right)^3$$

(Cauchy inequality)

$$= 16 \left(\frac{1}{3}\right)^3 \left(\frac{1}{3}\right)^3 = \boxed{\frac{16}{729}} \leftarrow \text{MLE}$$

Equality holds when: $\begin{cases} \frac{1-x}{2} = x \\ \frac{1-xy}{2} = xy \end{cases}$

$$\Leftrightarrow \begin{cases} x = \frac{1}{3} \\ y = 1 \end{cases} \Leftrightarrow \begin{cases} p_C = \frac{2}{3} \\ p_D = 0 \end{cases}$$

Answer to Problem 3

③. Large-Scale Machine Learning

ⓐ Exercise 12.5.3

a) Gini impurity :

$G = f(x,y) = 1 - x^2 - y^2$

(where x is the fraction of examples in 1^{st} class

y " " " 2^{nd} ")

$$\Rightarrow G = 1 - x^2 - (1-x)^2$$

$$= 1 - x^2 - 1 + 2x - x^2$$

$$= 2x - 2x^2$$

$$\Rightarrow G'' = -4 < 0$$

$\Rightarrow$ G is concave

b) Entropy measure :

$$E = g(x,y) = x \log_2\left(\frac{1}{x}\right) + y \log_2\left(\frac{1}{y}\right)$$

$$= x \log_2\left(\frac{1}{x}\right) + (1-x) \log_2\left(\frac{1}{1-x}\right)$$

$$E' = \log_2\left(\frac{1}{x}\right) + x \cdot \frac{-1}{x^2} \cdot (\ln 2\, x)$$

$$- \log_2\left(\frac{1}{1-x}\right) + (1-x) \cdot \frac{1}{(1-x)^2} \cdot \ln 2\,(1-x)$$

$$= \log_2\left(\frac{1}{x}\right) - \log_2\left(\frac{1}{1-x}\right)$$

$$\Rightarrow E'' = \frac{-1}{x^2}(\ln 2\, x) - \frac{1}{(1-x)^2}\ln 2\,(1-x)$$

$$= -\ln 2\left(\frac{1}{x} + \frac{1}{1-x}\right) < 0$$

(since $0 < x < 1$)

$\Rightarrow$ E is concave