EE412 Foundation of Big Data Analytics, Fall 2018
HW3

Name: Park Jaeyoung

Student ID: 20170273

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

(a) Solving Textbook
*Exercise 5.1.2
(a, b, c) = (0.26, 0.31, 0.43)

In [1]:
```
import numpy as np
from numpy import linalg as LA
import math
```

In [2]:
```
beta = 0.8
n = 3
ep = 10**-8
```

In [3]:
```
v = np.ones((n, 1))
e = np.ones((n, 1))

M = np.array([[1/3, 1/2, 0]
              ,[1/3, 0, 1/2]
              ,[1/3, 1/2, 1/2]], dtype=float);
```

In [4]:
```
delta = 1

while(delta > ep):
    nextv = beta * np.matmul(M, v) + (1 - beta) * e / n
    delta = LA.norm(nextv - v)
    v = nextv
```

In [5]:
```
v
```

Out[5]:
```
array([[0.25925927],
       [0.30864199],
       [0.43209879]])
```

*Exercise 5.3.1
(a) - (a, b, c, d) = (0.43, 0.19, 0.19, 0.19)

```
import numpy as np
from numpy import linalg as LA
import math
```

```
beta = 0.8
n = 4
ep = 10**-8
```

```
v = np.ones((n, 1))
e = np.array([[1],
              [0],
              [0],
              [0]])
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]], dtype=float);
```

```
delta = 1

while(delta > ep):
    nextv = beta * np.matmul(M, v) + (1 - beta) * e / 1
    delta = LA.norm(nextv - v)
    v = nextv
```

```
v
```

```
array([[0.42857145],
       [0.19047621],
       [0.19047621],
       [0.19047621]])
```

(b) – (a, b, c, d) = (0.39, 0.17, 0.27, 0.17)

```
import numpy as np
from numpy import linalg as LA
import math
```

```
beta = 0.8
n = 4
ep = 10**-8
```

```
v = np.ones((n, 1))
e = np.array([[1],
              [0],
              [1],
              [0]])
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]], dtype=float);
```

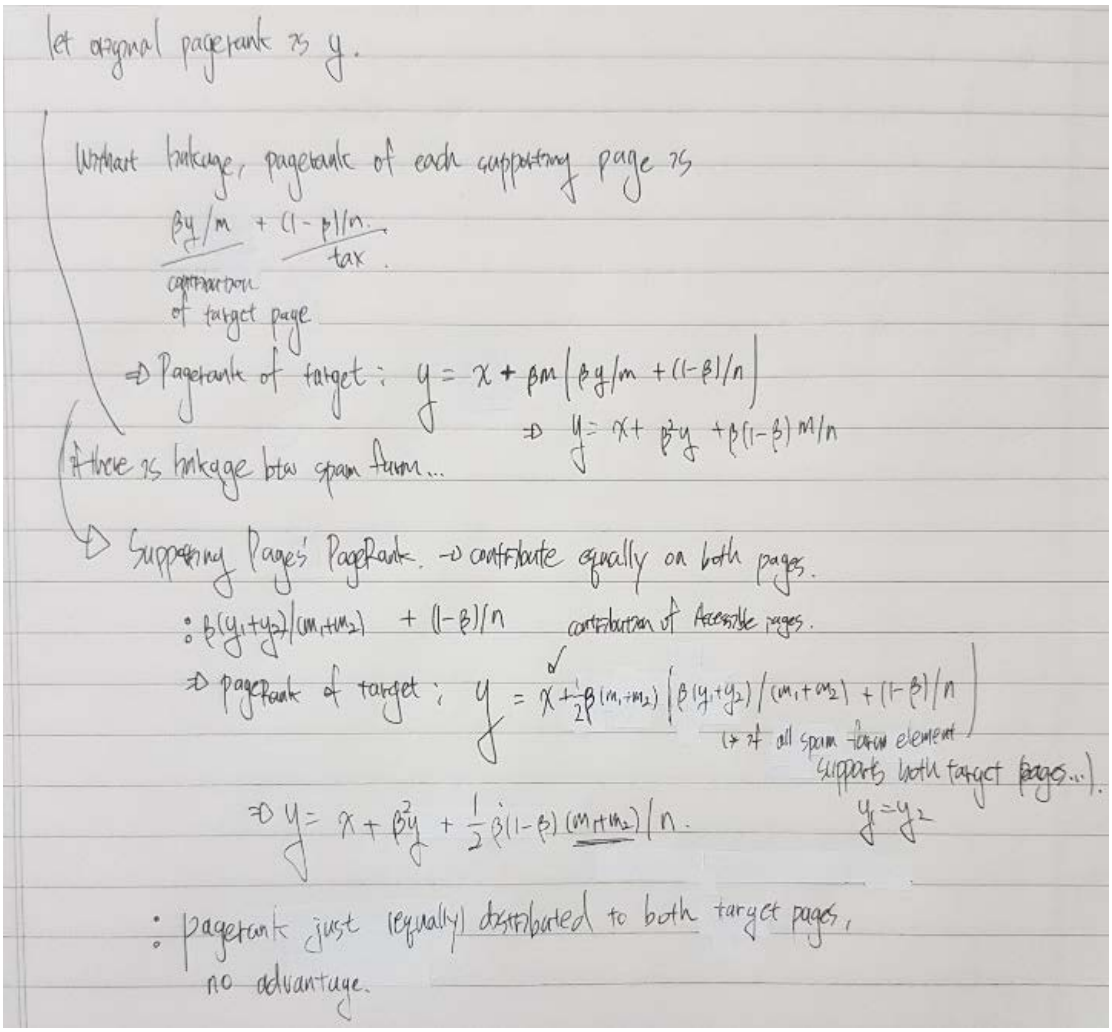
```
delta = 1

while(delta > ep):
    nextv = beta * np.matmul(M, v) + (1 - beta) * e / 2
    delta = LA.norm(nextv - v)
    v = nextv
```

```
v
```

```
array([[0.38571431],
       [0.17142859],
       [0.27142859],
       [0.17142859]])
```

*Exercise 5.4.3

Suppose first spam farmer has m1 supporting pages and second has m2 supporting pages. Let x be amount of PageRank contributed by the accessible pages. It would be same to both target pages, n be number of pages on the Web.



(b) Implement the PageRank algorithm using Spark.

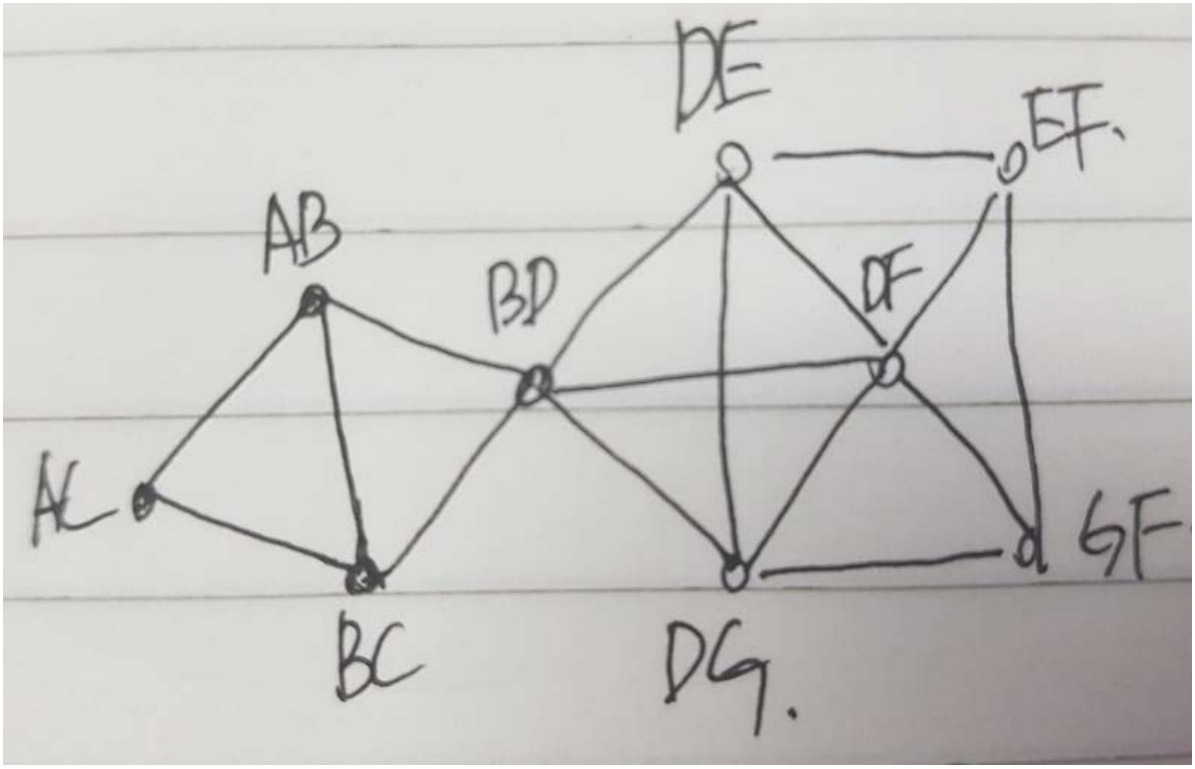
537 0.002317768147392944
263 0.0022954975147433778
965 0.002189010552384875
243 0.0020974242493302965
255 0.002078731413507323
285 0.002034967645645389
16 0.0020290537137209836
736 0.0020093037888945747
747 0.0020080350790072105
126 0.001991457813337111

(c) Implement the HITs algorithm using Spark.

893 1.0
16 0.9635572849634396
799 0.9510158161074016
146 0.9246703586198441
473 0.8998661973604047
624 0.8922197517765468
533 0.8832413304913613
780 0.8800357843384586
494 0.8749884615072088
130 0.8465465351844075
840 1.0
155 0.9499618624906542
234 0.8986645288972261
389 0.8634171101843788
472 0.8632841092495215
444 0.8229716669865107
666 0.8007139982829948
499 0.7966145570824411
737 0.7746877622644929
137 0.7715148677313686

Answer to Problem 2

*Exercise 10.1.1
(a) If node AB and BC are connected in G ′, it means that A – B – C are connected and B is between A and C in G.
(b)



(c)
$\deg_{G'}(XY) = \deg_G(X) + \deg_G(Y) - 2$

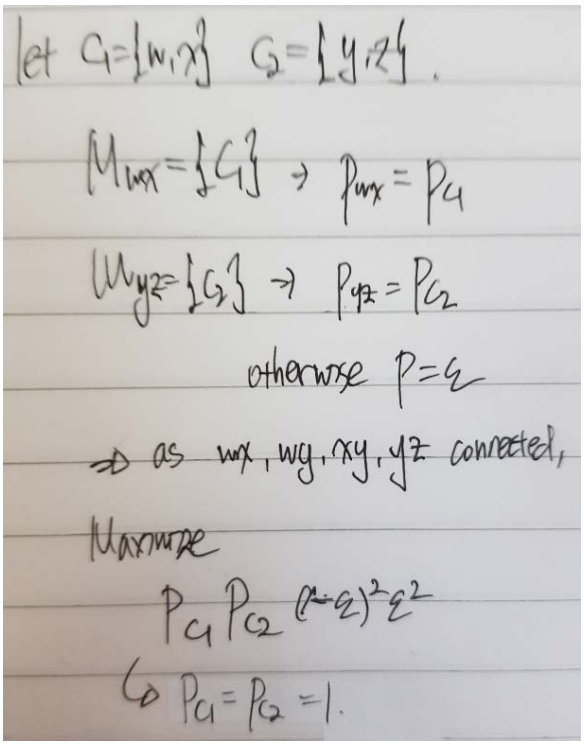
(d)
To be isomorphic, number of edge and vertex should be same. (It doesn't mean that converse is also true.)
If graph G shapes as polygon, then it is isomorphic with G′.
Example of (b) – between fig. 10.1(G) and photo above (G′) are not isomorphic.

*Exercise 10.3.2
(a)
$20 * (5/20)^t > s$
t = 1 → maximum s = 5 // t = 2 → maximum s = 1 (inconsistent)
(t, s) = (1, 5)

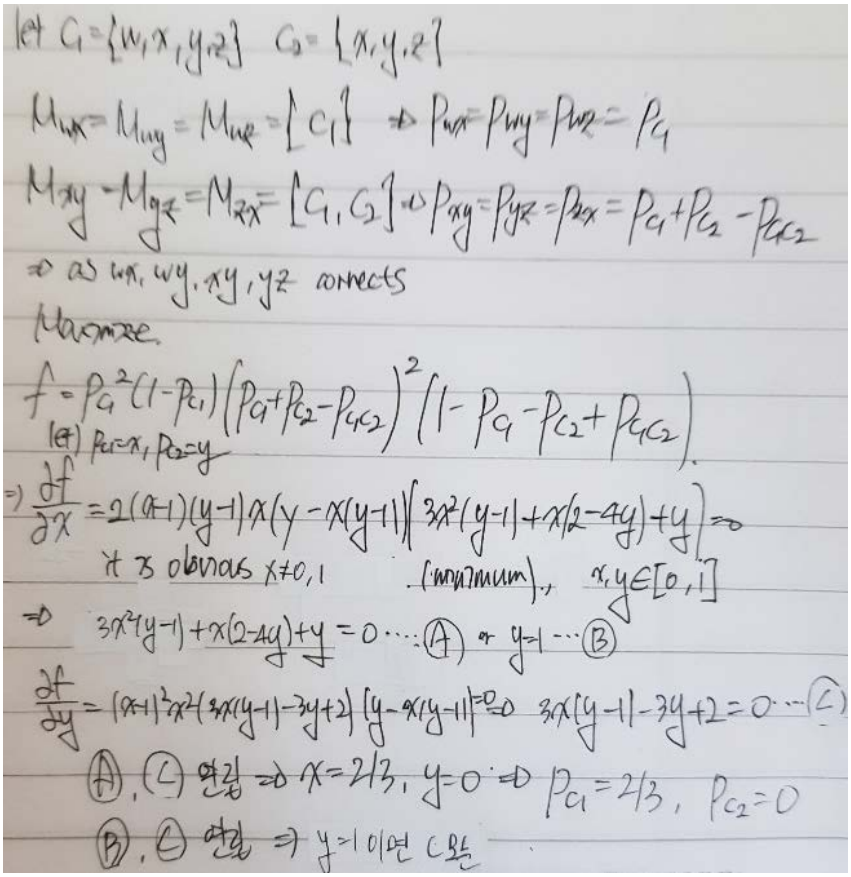
(b)
$200 * (150/200)^t > s$
Possible pairs of (t, s)
(1, 150), (2, 112), (3, 84), (4, 63), (5, 47), (6, 35) …(10, 11)

*Exercise 10.5.2

(a)



(b)



Answer to Problem 3

*Exercise 12.3.2
(a) ([2, 7, 2], +1), ([3, 3, 2], -1)
(b)
As **w** = [1, 1, 1] and b = -10 all points are outside the margin.

(c)
w = [-0.072, 0.670, 0.667] and b = -4.21 is assumed to be proper **w** and b

```
training
```

```
[(array([1, 4]), 1),
 (array([2, 2]), 1),
 (array([3, 4]), 1),
 (array([1, 1]), -1),
 (array([2, 1]), -1),
 (array([3, 1]), -1)]
```

```
def calc(_w, _b, _train, etha = 0.01, c = 10):
    new_w = np.append(_w, _b)
    grad_w = new_w.copy()
    grad_b = 0
    logic = 1
    for vector in _train:
        x = vector[0]
        x = np.append(x, 1)
        y = vector[1]
        val = y * np.dot(new_w, x)
        out_of_margin = (val >= 1)
        if not out_of_margin:
            grad_w -= c*y*x
            logic = 0
            ##print("ERROR")

    if(logic==1):
        nextw = new_w - etha*grad_w
    else:
        nextw = new_w - etha*grad_w
    nextb = nextw[-1]
    nextw = nextw[:-1]
    ##nextw /= np.sqrt(np.dot(nextw, nextw))
    return nextw, nextb, logic, grad_w
```

```
w = np.array([1., 1., 1.]) #last element is b
b = -10
logic = 0

count = 0
while(True):
    w, b, logic, grad = calc(w, b, training)
    count +=1
    if(np.dot(grad, grad)<19):
        if logic==1:
            break
```

```
grad
```

```
array([-0.07217596,  0.64984479,  0.66742955, -4.25145248])
```

```
w
```

```
array([-0.0714542 ,  0.64334635,  0.66075526])
```

```
b
```

```
-4.208937951914241
```

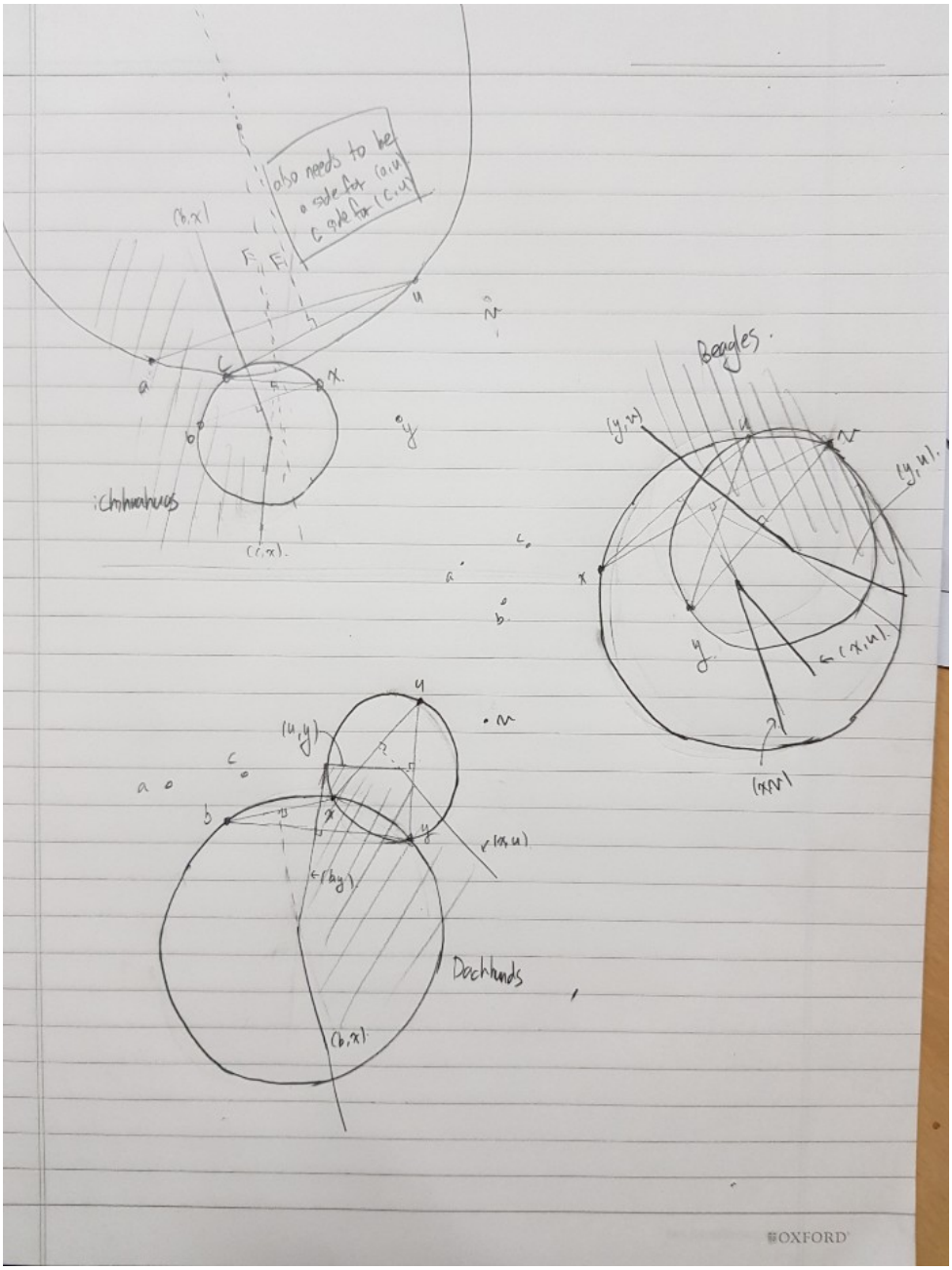
```
logic
```

```
1
```

```
for i in range(6):
    print(training[i][1] * (np.dot(w, training[i][0]) + b))
```

```
1.4538611181632488
1.4730885827151958
1.9542990701957326
1.0114336906807408
1.171750994475465
1.1177057085832804
```

*Exercise 12.4.1
(a)
Let query point a, b, c are for Chihuahuas, x, y are for Dachshunds and u, v are for Beagles.
*(s, t) means line segment which has same distance from both s and t.
*colored region on figure below is 2-nearest neighbored region.



(b)

For 2-NN classifier, boundaries always consist of straight-line segment.

It compares nearest neighbors, and nearer point in L2 distance between two points are determined by line segment.

Still, if it is weighted and use different distance measure, then boundaries might not be line segment.

*Exercise 12.5.3

While fraction of first class is ‘x’, then fraction of second class would be ‘1-x’.

(a) GINI

$f(x) = 1 – x^2 – (1-x)^2$

→ $f(x) = 2x – 2x^2$

$f'(x) = 2 – 4x$

$f''(x) = -4$

$f'' < 0$ → convex

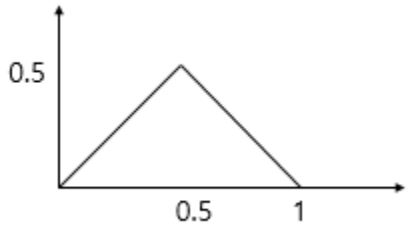
.

(b) Entropy

:$f(x) = x*\log_2(1/x) + (1-x)*\log_2(1/(1-x))$

$f'(x) = (\log(1/x) - \log(1/(1-x))) / \log(2)$
$f''(x) = 1 / (x-1)x$
$f'' < 0$ for $0 < x < 1$. → f is convex.

(c) Accuracy measure
Graph of $f(x) = 1 - \max(x, 1-x)$ is



This is consist of straight line, so it is not convex. (by Hint)