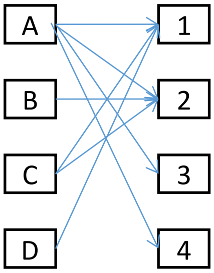
COMP 6000I Search Engines and Applications
Fall 2019 Homework 3

Due: Dec 6, 2019 11:59pm

Submit your answers in a zip file via Canvas.

Tomas Sousa Pereira
20667036

1. **[50]** In the following bipartite graph, the left nodes (A-F) are authors and the right nodes (1-6) are papers. A link from i to j means author i is the author paper j.



(a) [15] Compute by hand the authority and hub weights of each node in the graph. Let's note in advance that Authority weights of authors and Hub weights of papers are zero. The following steps are taken:

- All initial hub and authority weights are 1/6.
- L1 normalization is applied to the weights for each type of the nodes in an iteration, i.e., summation of each type of weights is one.

i. State below the Hub and Authority formulas for each of the nodes.

A: $x_A = y_1 + y_2 + y_3 + y_4$ 1: $y_1 = x_A + x_C + x_D$

B: $x_B = y_2$ 2: $y_2 = x_A + x_B + x_C$

C: $x_C = y_1 + y_2$ 3: $y_3 = x_A$

D: $x_D = y_1$ 4: $y_4 = x_A$

Asynchronous starting with Auth.

When computing Auth at iteration k we use Hub at iteration k-1. When computing Hub at iteration k we use Auth at iteration k.

ii. Fill in the values (with a little intermediate computation) in the table below.

Authors

Iteration	**0**	**1**		**2**	
		Before Normalization	**After Normalization**	**Before Normalization**	**After Normalization**
Hub(A)	1/4	1	2/5	1	25/59
Hub(B)	1/4	3/8	3/20	17/50	17/118
Hub(C)	1/4	3/4	3/10	17/25	17/59
Hub(D)	1/4	3/8	3/20	17/50	17/118

Papers:

Iteration	**0**	**1**		**2**	
		Before Normalization	**After Normalization**	**Before Normalization**	**After Normalization**
Aut(1)	1/4	3/4	3/8	17/20	17/50
Aut(2)	1/4	3/4	3/8	17/20	17/50
Aut(3)	1/4	1/4	1/8	2/5	4/25
Aut(4)	1/4	1/4	1/8	2/5	4/25

(b) [15] Give a plausible real-world interpretation of the Hub weights of authors and Authority weight of papers. That is, what does it mean when an author has high/low hub weight and a paper has high/low authority weight?

The papers have more weight if they have more authors pointing to them, i.e. the papers have higher Authority because they have more Authors linking to themselves.

The authors have more weight if they are pointing to more papers, i.e. they have higher Hub weight because they have linked to more papers.

(c) [5] What is the main difference between this graph and a typical web graph.

In the web all the nodes are the same, each node can point to all of the nodes including to himself. Here the papers can't point to the authors.

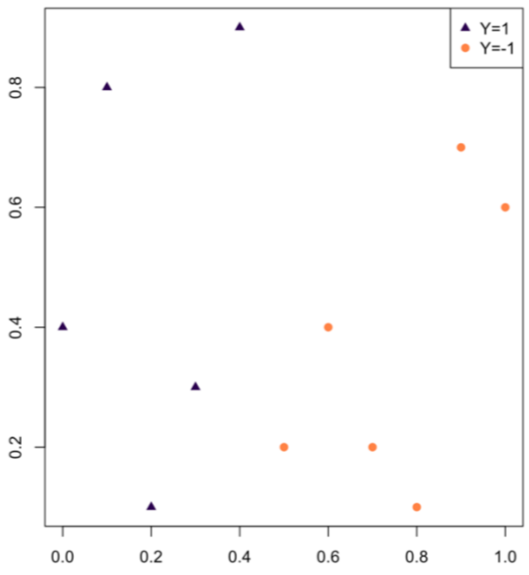
(d) [15] If we apply PageRank to the graph, ignoring the fact that there are two types of nodes, can you give a meaningful interpretation of PageRank of the nodes? Yes or no or indecisive, please Justify.

No. Nowadays is not very likely that one node don't point to no other node. Almost every every node (in page rank graphs) points to another node. In this case we cant observe that, therefore we cant retrieve any meaningful interpretation.

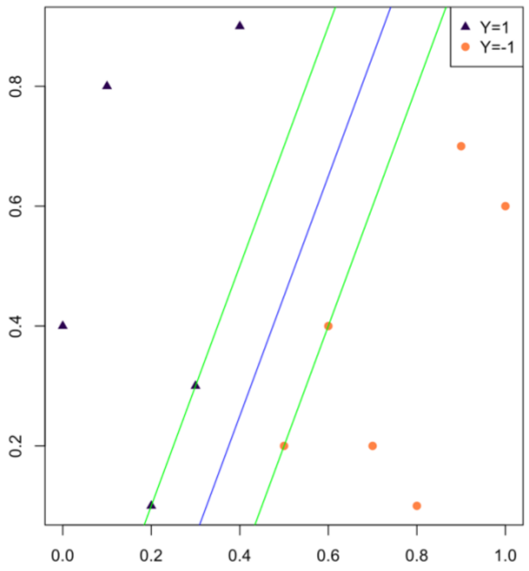
2. **[50]** After click log mining, the search engine has discovered the relevance of 11 documents, shown in the table below, where numbers in the first column are page IDs, X1 and X2 are features, y=1 indicates relevance, y=-1 non-relevance.

```
    X1  X2  y
1   0.0 0.4  1
2   0.1 0.8  1
3   0.2 0.1  1
4   0.3 0.3  1
5   0.4 0.9  1
6   0.5 0.2 -1
7   0.6 0.4 -1
8   0.7 0.2 -1
9   0.8 0.1 -1
10  0.9 0.7 -1
11  1.0 0.6 -1
```

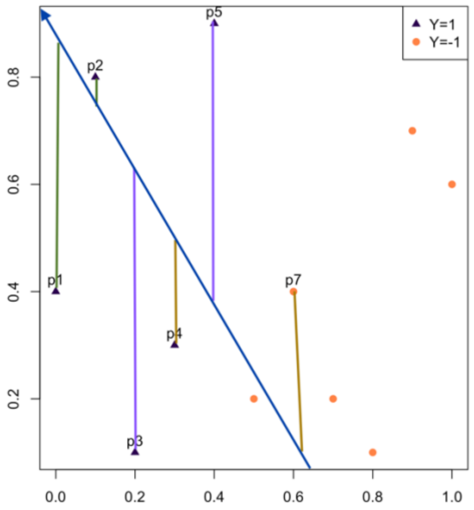
(a) [5] Plot the data points on a two-dimensional graph.



(b) [20] By hand, draw the decision function and identify the support vectors on the graph.



(c) [25] With the same page set, suppose preference mining reveals that p1>p2, p5>p3, p4>p7 (p*i* are pages). Draw a decision function, clearly indicating the direction of the vector (vector pointing to direction of high preference). If any preference is not satisfied by your decision function, please explain why.



Unfortunately, we can’t satisfy all of the preferences. This happens because we cant find a vector α which the order satisfy the preference order. Therefore, we should satisfies as many preferences as possible we choose to satisfy p1>p2 and p4>p7. We also tried to optimize trying to find largest distance between the two closest projects.