

Name	
ASU ID Number	

CSE 472: Social Media Mining

Homework II - Network Models and Data Mining

Prof. Huan Liu
Due at 2023 Sept 22, 11:59 PM

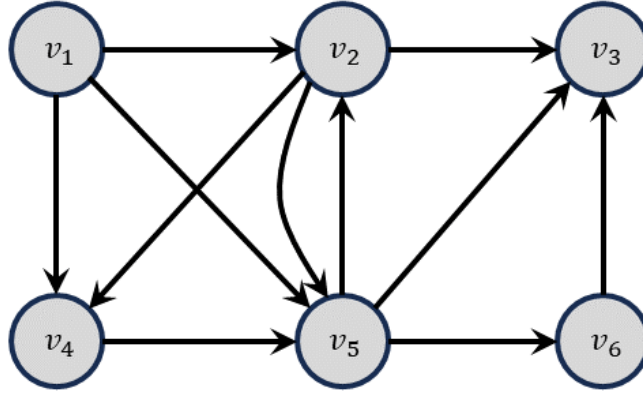
This is an *individual* homework assignment. Please submit a digital copy of this homework to **Grade-scope**. This is a fillable PDF and you are able to type into answer boxes provided for each question.

1. [Network Models]

- a. Why are random graphs incapable of modeling real-world graphs?

- b. We can make a simple random graph model of a network with clustering or transitivity as follows. We take the n vertices and go through each distinct trio of three vertices, of which there are $\binom{n}{3}$, and with independent probability P we connect the members of the trio together using three edges to form a triangle, where $P = \frac{c}{\binom{n-1}{2}}$ with c constant. Show that the mean degree of a vertex in this network is $2c$.

- c. In a followee/follower network, each node represents a user, and a directed-edge exists between two nodes v_i and v_j ($v_i \rightarrow v_j$) if user v_i follows user v_j or vice-versa. When a new user (new node) joins to the network, that user follows other users with probability *proportional to other users' in-degree*, and is followed with probability *proportional to other users' out-degree*, hence, **following the preferential attachment model**. At $t=0$, we have 6 users of $V_0 = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, Please calculate the probability that the new node follows other nodes and also calculate the probability that the new node is followed by others. At $t=1$, v_7 joins and follows a ($= 4$) users with the highest in-degree probabilities. Additionally, v_7 is followed by users b ($= 3$) with the highest out-degree probabilities. Please update each node's probabilities (P_{in} and P_{out}); and for $t=2$, v_8 joins and follows a ($= 2$) users with the highest in-degree probabilities and is followed by b ($= 4$) users with the highest out-degree probabilities.



Algorithm 1: Preferential Attachment Algorithm

Input : Graph $G(V_0, E_0)$, where $|V_0| = n_0$ and $d_v \geq 1 \forall v \in V_0$, a number of edges going out from v , b number of edges coming into v , $a + b \leq n_0$, time to run algorithm t

Output: A scale-free network

```

1 //Initial graph with  $n_0$  nodes with degrees at least 1
2  $G(V, E) = G(V_0, E_0)$ 
3 for 1 to  $t$ :
4    $V = V \cup \{v_i\}$ 
5   While  $d_i^{out} \neq a$ :
6     Connect  $v_i$  to the node  $v_j \in V$ ,  $i \neq j$  (i.e.,  $E = E \cup \{e(v_i, v_j)\}$ ) with the highest
    in-degree probability  $P_{in}(v_j) = \frac{d_j^{in}}{\sum_k d_k^{in}}$ 
7   While  $d_i^{in} \neq b$ :
8     Connect  $v_j \in V$  to the node  $v_i$ ,  $i \neq j$  (i.e.,  $E = E \cup \{e(v_j, v_i)\}$ ) with the highest
    out-degree probability  $P_{out}(v_j) = \frac{d_j^{out}}{\sum_k d_k^{out}}$ 
9 return  $G(V, E)$ 

```

Probability proportional to other users when the new user follows others:

t	x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
$t = 0$	$P_{in}(v_x)$							NA	NA
$t = 1$	$P_{in}(v_x)$								NA
$t = 2$	$P_{in}(v_x)$								

Probability proportional to other users when the new user is followed by others:

t	x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
$t = 0$	$P_{out}(v_x)$							NA	NA
$t = 1$	$P_{out}(v_x)$								NA
$t = 2$	$P_{out}(v_x)$								

2. [Data Mining]

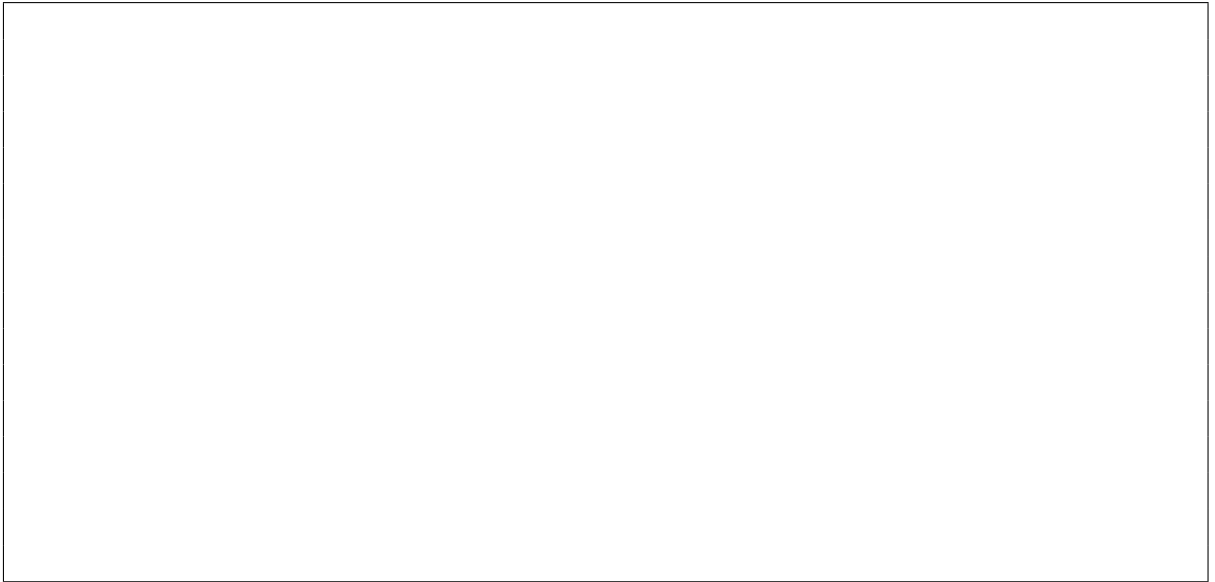
Consider the given dataset from an employee database. For a given entry row, the column “*count*” represents the number of data tuples (*department*, *salary*, *status*) with the values given in the row. For example, there are 15 instances with values of (department = sales, salary = high, status = senior). Let “*status*” be the target class label, answer the following questions. [It is necessary to set the default base value for all logarithms to 2.]

Department	Salary	Status	Count
Sales	High	Senior	15
Sales	Low	Junior	20
Systems	Medium	Junior	10
Marketing	Low	Junior	10
Marketing	Medium	Senior	10
Marketing	High	Senior	5
Human Resources	Medium	Junior	7
Human Resources	Very High	Senior	3

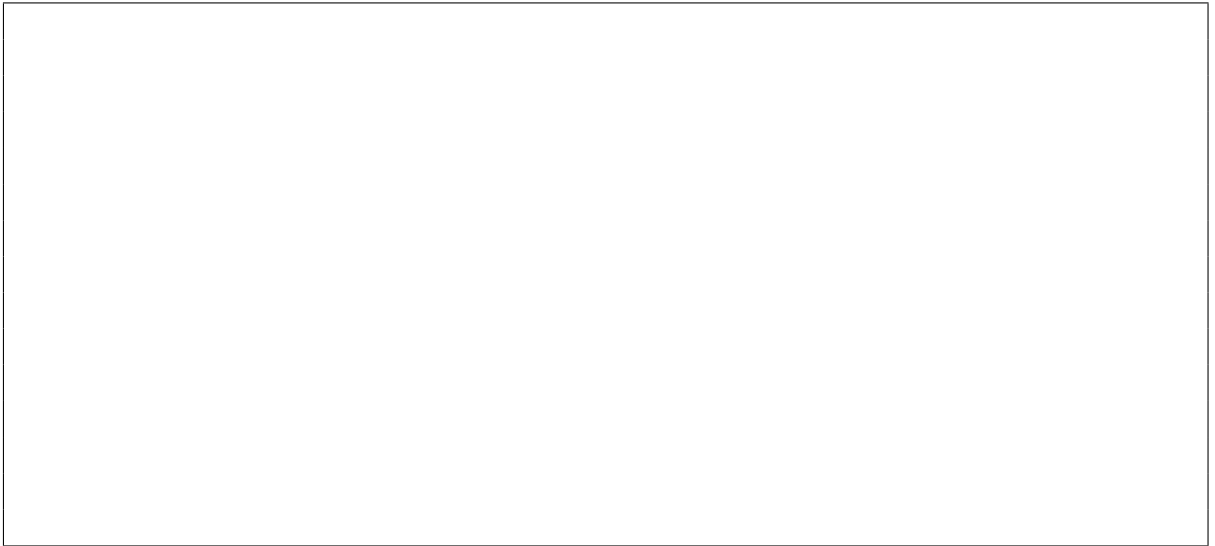
- (a) What is the value for the $H(Status)$? Here $H(x)$ defines the entropy of x .

- (b) Based on the Information Gain values, which feature is most probable to be the root node of the decision tree? Show all your work.

- (c) Draw the final decision tree. An example of how to draw the tree on the text box:



- (d) Given a data instance having the values “*Human Resources*” and “*Medium*” for the attributes “*department*” and “*salary*”, respectively, what would a Naive Bayesian classifier predict for the instance’s “*Status*”? Detail all your calculations.



Good Luck