Name	Solution
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 **Gradescope**. This is a fillable PDF and you are able to type into answer boxes provided for each question.

1. [**Network Models**]

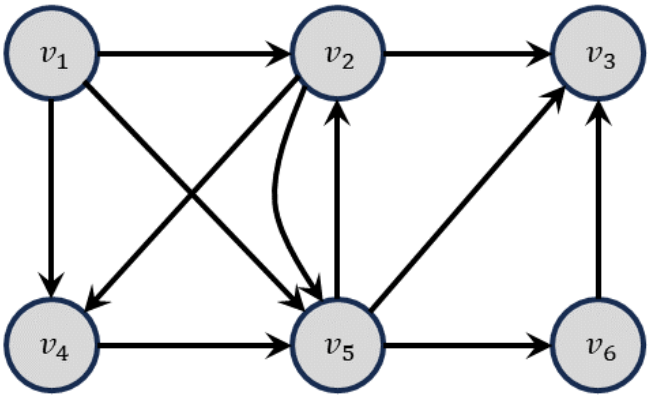
   a. Why are random graphs incapable of modeling real-world graphs?

   Real-world graphs have low average path length, high clustering coefficient and power-law degree distribution while random graphs have low average path length and small clustering coefficient. Moreover, their degree distribution is not power-law at all.

   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 $\mathsf{P} = \frac{c}{\binom{n-1}{2}}$ with c constant. Show that the mean degree of a vertex in this network is 2c.

   For each vertex there are (n-1)*(n-2)/2 pairs of others with which it could form a triangle, and each triangle is present with the probability c/[(n-1)*(n-1)/2] for an avg, number of triangles per vertex. Each triangle contributes two edges to the degree so average degree 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 vise-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

//Initial graph with n_0 nodes with degrees at least 1
$G(V, E) = G(V_0, E_0)$
for 1 to t:
    $V = V \cup \{v_i\}$
    While $d_i^{out} \neq a$:
        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 $\mathsf{P}_{in}(v_j) = \frac{d_j^{in}}{\sum_k d_k^{in}}$
    While $d_i^{in} \neq b$:
        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 $\mathsf{P}_{out}(v_j) = \frac{d_j^{out}}{\sum_k d_k^{out}}$
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	$\mathsf{P}_{in}(v_x)$	0	** 2/11 ≈ 0.18	** 3/11≈ 0.27	** 2/11 ≈ 0.18	** 3/11≈ 0.27	1/11≈ 0.09	NA	NA
t = 1	$\mathsf{P}_{in}(v_x)$	0	3/18 ≈ 0.16	** 4/18 ≈ 0.22	3/18 ≈ 0.16	** 4/18 ≈ 0.22	1/18 ≈ 0.05	3/18 ≈ 0.16	NA
t = 2	$\mathsf{P}_{in}(v_x)$	0	3/24 = 0.125	5/24 ≈ 0.2083	3/24 = 0.125	5/24 ≈ 0.2083	1/24 ≈ 0.0416	3/24 = 0.125	4/24 ≈ 0.1666

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	$\mathsf{P}_{out}(v_x)$	** 3/11≈ 0.27	** 3/11≈ 0.27	0	1/11≈ 0.09	** 3/11≈ 0.27	1/11≈ 0.09	NA	NA
t = 1	$\mathsf{P}_{out}(v_x)$	** 4/18 ≈ 0.22	** 4/18 ≈ 0.22	0	1/18 ≈ 0.05	** 4/18 ≈ 0.22	1/18 ≈ 0.05	** 4/18 ≈ 0.22	NA
t = 2	$\mathsf{P}_{out}(v_x)$	5/24 ≈ 0.2083	5/24 ≈ 0.2083	0	1/24 ≈ 0.0416	5/24 ≈ 0.20	1/24 ≈ 0.0416	5/24 ≈ 0.20	2/24 ≈ 0.8833

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.

N(Status=Senior) = 15 + 10 + 5 + 3 = 33
N(Status=Junior) = 20 + 10 + 10 + 7 = 47

H(Status) = -33/80 log2(33/80) - 47/80 log2(47/80) = 0.9777

(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.

Department as the root node feature:
Sales: 15 senior, 20 junior, 35 total | Systems: 0 senior, 10 junior, 10 total
Marketing: 15 senior, 10 junior, 25 total | Human Resources: 3 senior, 7 junior, 10 total

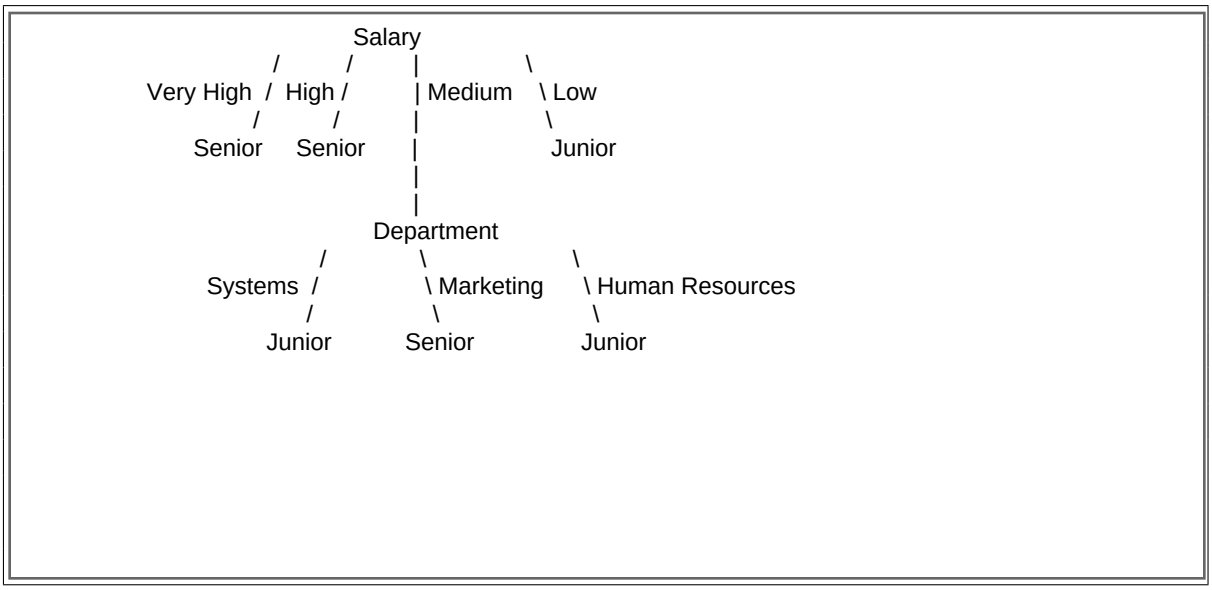
H(Sales) = -15/35log2(15/35) - 20/35log2(20/35) = 0.9852
H(System) = -10/10log2(10/10) = 0
H(Marketing) = -15/25log2(15/25) - 10/25log2(10/25) = 0.9709
H(Human Resources) = -3/10log2(3/10) - 7/10log2(7/10) = 0.8812
IG(A=department, S=status) = 0.9777-(35/80)*0.9852-(10/80)*0-(25/80)*0.9709-(10/80)*0.8812= 0.1331

Salary as the root node feature:
Very High: 3 senior, 0 junior, 3 total | High: 20 senior, 0 junior, 20 total
Medium: 10 senior, 17 junior, 27 total | Low: 0 senior, 30 junior, 30 total

H(Very High) = 0
H(High) = 0
H(Medium) = -10/27log2(10/27) - 17/27log2(17/27) = 0.6602
H(Low) = 0
IG(A=Salary, S=status) = 0.9777 - (27/80) * 0.6602 = 0.7548 > 0.1331 = IG(A=department, S=status)

=> Salary is the most probable feature

(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.

*HR: Human Resources

P(Senior | HR, Medium) = P(HR, Medium | Senior) * P(Senior) / P(HR, Medium)
= P(HR | Senior) * P(Medium | Senior) * P(Senior) / P(HR, Medium)
= 3/33 * 10/33 * 33/80 / P(HR, Medium)
= 1/88 / P(HR, Medium) = 0.011 / P(HR, Medium)

P(Junior | HR, Medium) = P(HR, Medium | Junior) * P(Junior) / P(HR, Medium)
= P(HR | Junior) * P(Medium | Junior) * P(Junior) / P(HR, Medium)
= 7/47 * 17/47 * 47/80 / P(HR, Medium)
= 119/3760 / P(HR, Medium) = 0.031 / P(HR, Medium)

=> P(Senior | HR, Medium) < P(Junior | HR, Medium)

=> Status should be Junior

Good Luck