Name	Saurabh Zinjad
ASU ID Number	1229549687

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?

   Random graphs are a type of mathematical model that assumes that links in real-world networks are established entirely at random. However, this assumption does not correspond to the fundamental mechanisms that underlie the development of relationships or connections in actual networks. In real-world networks, friendships frequently develop as a result of common interests, acquaintances, or other non-random causes.

   Additionally, random graphs fall short of capturing the universal traits seen in real-world networks, such as degree distribution, clustering coefficient, and average path length. Degree distribution describes the distribution of node degrees within a network. In contrast to the power-law degree distribution often seen in real-world networks, random graphs show a Poisson degree distribution.

   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.

   E(D) (the expected degree of a vertex), The expected degree of a single vertex is calculated as the sum of the probabilities of it being part of each possible triangle:

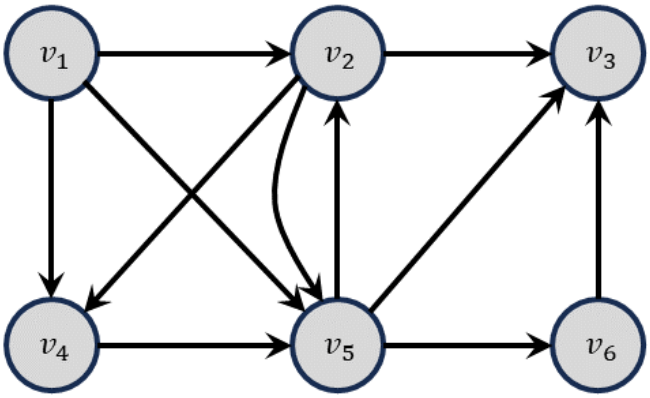
   E(D)=Sum(all possible triangles) * P

   There are ((n-1) choose 2) possible triangles since there are n−1 other vertices to form triangles with the selected vertex. Therefore:

   E(D)=(((n-1)C2) * P

   Now, substitute the expression for P:

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	3/11	2/11	3/11	1/11	NA	NA
t = 1	$\mathsf{P}_{in}(v_x)$	0	3/18	4/18	3/18	4/18	1/18	3/18	NA
t = 2	$\mathsf{P}_{in}(v_x)$	0	3/24	5/24	3/24	5/24	1/24	3/24	4/24

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	3/11	0	1/11	3/11	1/11	NA	NA
t = 1	$\mathsf{P}_{out}(v_x)$	4/18	4/18	0	1/18	4/18	1/18	4/18	NA
t = 2	$\mathsf{P}_{out}(v_x)$	5/24	5/24	0	1/24	5/24	1/24	5/24	2/24

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.

Answer: H(Status) = 0.978

total_count = 15+20+10+10+10+5+7+3
p_junior = (20+10+10+7) / total_count = 47/80 = 0.5875
p_senior = (15+10+5+3) / total_count = 33/80 = 0.4125

Using entropy formula = H(x) = -1*[(summation from i=1 to n) pi *log2(pi)]
H(status) = - p_junior * log2(p_junior) - p_senior*log2(p_senior)
H(status) = -(0.5875)*(-0.7673) - (0.4125) * (-1.278)

(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.

"SALARY" is most probable root node of the decision tree. Cause Information gain of "Salary" is greater than "Department"

IG(Department, Staus) = H(Status) - (35/80)*H(Sales) - (10/80)*H(System) - (25/80)*H(Marketing) - (10/80)*H(Human Resources)

H(Sales) = -[(15/35)*log2(15/35)]-[(20/35)*log2(20/35)] = 0.985
H(System) = -[(0/10)*log2(0/10)]-[(10/10)*log2(10/10)] = 0
H(Marketing) = -[(10/25)*log2(10/25)]-[(15/25)*log2(15/25)] = 0.9709
H(Human Resources) = -[(7/10)*log2(7/10)]-[(3/10)*log2(3/10)] = 0.8812

IG(Department, Staus) = 0.978 - (35/80)*(0.985) - (10/80)*(0) - (25/80)*(0.9709) - (10/80)*(0.8812)

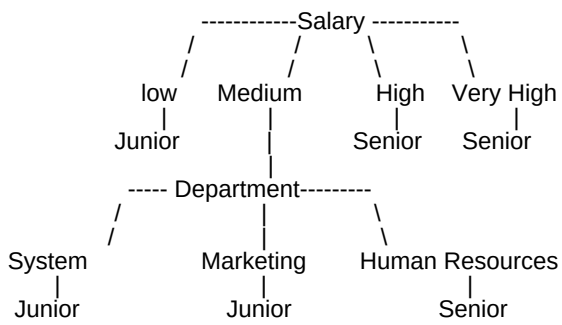
==========> IG(Department, Staus) = 0.1335 <==========

IG(Salary, Staus) = H(Status) - (20/80)*H(High) - (30/80)*H(Low) - (27/80)*H(Medium) - (3/80)*H(Very High)

H(High) = -[(0/20)*log2(0/20)]-[(20/20)*log2(20/20)] = 0
H(Low) = -[(30/30)*log2(30/30)]-[(0/30)*log2(0/30)] = 0
H(Medium) = -[(17/27)*log2(17/27)]-[(10/27)*log2(10/27)] = 0.95
H(Very High) = -[(0/3)*log2(0/3)]-[(3/3)*log2(3/3)] = 0

IG(Department, Staus) = 0.978 - (27/80)*(0.95) = 0.657

(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.

Naive bayesian classifier predict "Junior" for the instance's "Status".

We want to calculate P(J|HR,M), using Bayes and Naiive Bayes -

P(J|HR,M) = P(HR|J) * P(M|J) * P(J) / ((P(HR|J) * P(M|J) * P(J)) + P(HR|S) * P(M|S) * P(S))
P(J|HR,M) = (7/42 * 17/42 * 42/80) / ((7/42 * 17/42 * 42/80) + (3/38 * 10/38 * 38/80))
P(J|HR,M) = 0.782

If we calcuate P(S|HR,M) in same way it is 0.0027

Since Junior "Probability" is greater than "Senior".

Good Luck