Name	PRAVEENAA KULANDHAIVEL
ASU ID Number	1225814683

CSE 472: Social Media Mining

Homework II - Network Models and Data Mining

Prof. Huan Liu
Due at 2022 Sept 21, 11:59 PM

This is an ***individual*** homework assignment. Please submit a digital copy of this homework to **Gradescope**. For your solutions, even when not explicitly asked you are supposed to concisely justify your answers.

1. **[Network Models]**

   a. Assuming that we are interested in a dense random graph, what should we choose as the value of p? How does the value of p affect the sparseness? Where p defines the probability of forming edges.

   Expected number of edges connected to a node, $c = (n-1)p$
   higher value of $p \Rightarrow$ more edges $\Rightarrow$ dense graph
   highest value of $p = 1$, there will be $(n-1)$ edges connected.

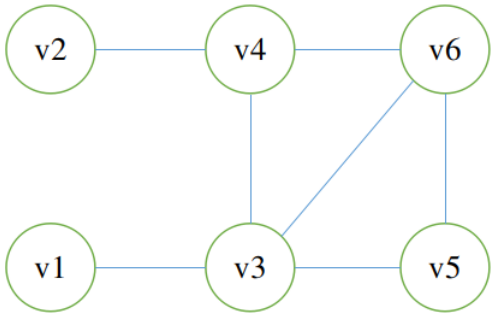
   b. We can make a simple random graph model of a network with clustering or transitivity as follows. We take n vertices and go through each distinct trio of three vertices, of which there are $\binom{n}{3}$, and with independent probability p we connect the members of the trio together using three edges 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.

   Total number of pairs forming triangles - $\binom{n-1}{2}$

   Probability for each triangle $\Rightarrow \frac{c}{\binom{n-1}{2}}$

   $\Rightarrow$ Therefore, each triangle contributes 2 edges for degree making the average 2c.

c. In a citation network, each node represents a paper, and an edge exists between two nodes if one paper cites another. When a new paper is published (i.e., a new node added to the graph), that paper cites previous papers with probability *proportional to their degrees*, hence, following the **preferential attachment model**. Suppose at each time-step a new paper is submitted to the following citation network. At t=0, we have 6 papers of $V_0 = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, please calculate the probability of the new node connecting to each node. At t=1, v_7 joins and attaches to n (= 4) papers with the highest probabilities, please update each node's probability; and do the same for t=2, when v_8 joins. Leave your answers as common fractions (e.g., $\frac{1}{6}, \frac{4}{5}$).



t	x	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
t=0	$P(v_x)$	1/14	1/14	4/14	3/14	2/14	3/14	NA	NA
t=1	$P(v_x)$	1/22	1/22	5/22	4/22	3/22	4/22	4/22	NA
t=2	$P(v_x)$	1/30	1/30	6/30	5/30	3/30	5/30	5/30	4/30

Note: *NA* in the table represents the situation when the new node cannot connect to the other nodes in the network.

Algorithm 1: Preferential Attachment Algorithm

Input : Graph $G(V_0, E_0)$, where $|V_0| = n_0$ and $d_v \geq 1\ \forall v \in V_0$, number of expected connections $n \leq n_0$, time to run the 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 ∪ {v_i}        //add new node v_i
    while d_i ≠ n
        Connect v_i to a random node v_j ∈ V, i ≠ j (i.e., E = E ∪ {e(v_i, v_j)}) with probability P(v_j) = d_j / Σ_k d_k
return G(V, E)
```

2. **[Data Mining]** Consider the given dataset from an employee database. For a given entry row, *count* column represents the number of data tuples having the values for *department*, *salary*, and *status* given in that row. For example, there are 15 instances with values of (department = sales, salary = high, status=senior). Let *status* be the class label attribute, answer the following questions. Please use 2 as the default value for base of all logs.

Department	Salary	Status	Count
Sales	High	Senior	15
Sales	Low	Junior	20
Systems	Medium	Junior	10
Systems	High	Senior	10
Marketing	High	Junior	10
Marketing	High	Senior	5

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

$H(Status) = -\frac{30}{70}\log_2\frac{30}{70} - \frac{40}{70}\log_2\frac{40}{70} \Rightarrow -0.42\times1.2 - 0.57\times-0.8$
$= 0.504 + 0.456$
$= 0.96$

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

Salary: High → $30/40, 10/40$; Medium → $0, 10/10$; Low → $0, 20/20$

$I(Salary=High) = -\left[\frac{30}{40}\log_2\frac{30}{40} + \frac{10}{40}\log_2\frac{10}{40}\right] = (-0.3 + -0.5) = 0.8$
$I(Sal=Med) = -\left[0 + \log_2\frac{10}{10}\right] = 0$
$I(Sal=Low) = -\left[0 + \log_2\frac{20}{20}\right] = 0$

Dept: Sales → $15/35, 20/35$; Systems → $10/20, 10/20$; Marketing → $5/15, 10/15$

$I(Dept=Sales) = -\left[\frac{15}{35}\log_2\frac{15}{35} + \frac{20}{35}\log_2\frac{20}{35}\right] = -(-0.520 + -0.46) = 0.98$
$I(Dept=Systems) = -\left[\frac{10}{20}\log_2\frac{10}{20} + \frac{10}{20}\log_2\frac{10}{20}\right] = 1$
$I(Dept=Marketing) = -\left[\frac{5}{15}\log_2\frac{5}{15} + \frac{10}{15}\log_2\frac{10}{15}\right] = (0.52 + 0.36) = 0.9$

$I(Status, Sal) = \frac{40}{70}\times0.8 + \frac{10}{70}\times0 + \frac{20}{70}\times0 = 0.45$

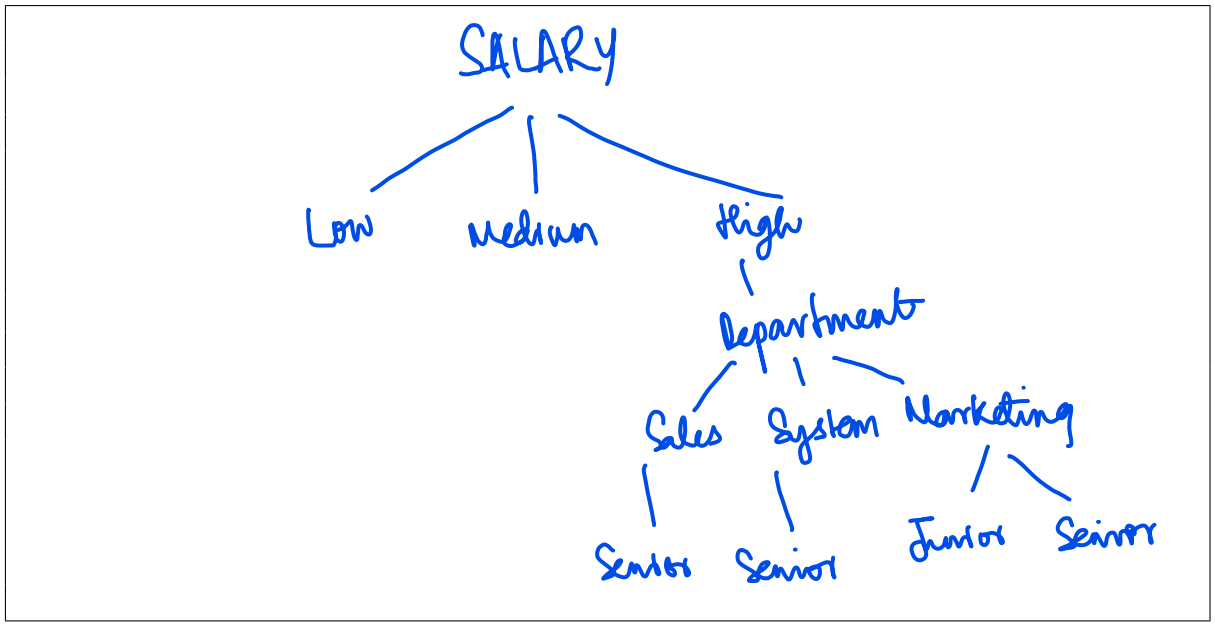
$I(Status, Dept) = \frac{35}{70}\times0.98 + \frac{20}{70}\times1 + \frac{15}{70}\times0.9 = 0.49 + 0.28 + 0.19 = 0.96$

$I.G(Status, Salary) = 0.96 - 0.45 = 0.51$
$I.G(Status, Dept) = 0.96 - 0.96 = 0$

Salary has higher Information Gain. Therefore, will be chosen as root node.

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

```
    A
   / \
  B   C
```



d. Given a data instance having the values *"Sales"* and *"High"* for the attributes *department* and *salary*, respectively, what would a Naive Bayesian classification of the class attribute *Status* for the instance be? Detail all your calculations.

$$P(\text{Junior} \mid \text{Sales \& High}) = \frac{P(\text{Sales, High} \mid \text{Junior})\,P(\text{Junior})}{P(\text{Sales, High})}$$

$$= \frac{(20/40) \times (10/40) \times (40/40)}{P(\text{Sales, High})} = 0.07 * \left(\frac{1}{P(\text{Sales, High})}\right)$$

$$P(\text{Senior} \mid \text{Sales \& High}) = \frac{P(\text{Sales, High} \mid \text{Senior})\,P(\text{Senior})}{P(\text{Sales, High})}$$

$$= \frac{(15/30) \times (30/30) \times (30/40)}{P(\text{Sales, High})} = 0.21 * \left(\frac{1}{P(\text{Sales, High})}\right)$$

Therefore, Value for Senior > Value of Junior. Status would be Senior