Assessment 3 - Computer Science

Make sure that you always use notations consistent with lecture notes.

Marks: 15 marks, weighted 15% of your final mark.

Submission format: Adobe PDF (preferred) or MS Word Doc.

Submit to: Moodle.

Late submission penalty: No late submission allowed.

The deadline for assignment 3 is: Mon 5th, August 10:00 pm.

Q1. Binary Search Trees (BST) (3 marks total)

(1) (1 mark) Starting from an empty binary search tree, what is the final binary search tree constructed by inserting the following elements in the specified order: {20, 40, 13, 4, 89, 21, 65, 15, 28, 46, 25, 36, 8,14,18}.

(2) (1 mark) Draw a BST with the same elements from (q1.1) but has a maximum height of 3 instead.

(3) (1 mark) Using the tree you built from (q.1.1), perform removing node {21} and draw the resulting BST.

Q2. Decision Trees (4 marks total)

Outlook	Temperature	Wind	Play Tennis
Sunny	Hot	Weak	No
Sunny	Hot	Strong	No
Overcast	Hot	Weak	Yes
Rain	Mild	Weak	Yes
Rain	Cool	Weak	Yes
Rain	Cool	Strong	No
Overcast	Cool	Weak	Yes
Sunny	Mild	Strong	Yes
Sunny	Mild	Strong	Yes
Overcast	Mild	Strong	No

(1) (3 marks) Computing the conditional entropy and the information gain for each predictor using the ID3 algorithm, give the steps.

(2) (1 mark) Build a decision tree using the computed information gain for the predictors.

Q3. Bloom Filter (4 marks total)

UNSW Estate Management needs to check if a license plate might have Parking Permits quickly. It needs to check a large number of license plates every day. To boost efficiency, they employ a Bloom filter. They have a Bloom filter of size *m* = 10 and *k* = 3. The hash functions that they employ are publicly listed below: $h_1,\ h_2,\ and\ h_3$.

$$h_1(plate\_num) = \left(\sum_{n=1}^{l}(c_n - 1)\right) mod\ m$$

$$h_2(plate\_num) = \left(\sum_{n=1}^{l}(c_n + h_1(plate\_num))\right) mod\ m$$

$$h_3(plate\_num) = (h_1(plate\_num) + h_2(plate\_num))\ mod\ m$$

Notes:

i. ***plate_num*** denotes a single plate number entry.

ii. l denotes the length of the plate number.

iii. c_n denotes the n[th] character of a ***plate_num***, e.g., for "*VWX901*", $c_1 = 'V'$, $c_2 = 'W'$, $c_3 = 'X'$, $c_4 = 9'$, $c_5 = '0'$, $and\ c_6 = '1'$.

iv. If c_n is not a number, consult the following table below.

A	B	C	D	E	F	G	H	I	J	K	L	M
↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕
0	1	2	3	4	5	6	7	8	9	10	11	12
N	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕	↕
13	14	15	16	17	18	19	20	21	22	23	24	25

Tasks:

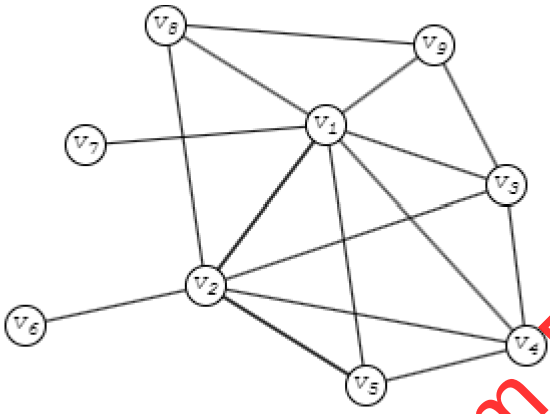
(1) (2 marks) Given an empty bloom filter, initialize the bloom filter by inserting the following stream of **4** plate numbers with Parking Permit from S_1 into the Bloom filter. S_1 = {"NSW171", "CMD256", "XUV971", "KBM899"}. Using the bloom filter to check plate numbers in S_2 = {"MNO121", "QRS247"}, give the steps.

(2) (1 mark) If the bloom filter size m=100 and there are 15 plates to initialize the bloom filter, what is the optimal number of hash functions? Compute the new probability of false positive.

(3) (1 mark) What can we do to decrease the false positive rate?

Q4. K-cores (4 marks total)

The k-core of a graph is the *maximal* subgraph such that every vertex in the k-core has a degree of at least k. We provide the following graph G with 9 vertices.



(1) (1 mark) Draw the 3-core of G.

(2) (1 mark) Add one edge to the original graph G to make a 4-core. Justify your answer.

(3) (2 marks) Another subgraph metric k-truss, is similar to k-core, the definition of k-truss is as follows.

Definition: *A k-truss is a subgraph where every edge is part of at least k−2 triangles within the subgraph.*

Note: In a graph G=(V,E), let V be the set of vertices and E be the set of edges. A triangle is a set of three vertices {v1,v2,v3}⊆V such that (v1,v2)∈E, (v2,v3)∈E, and (v1,v3)∈E.

Draw the 4 -truss of G, justify your answer.