BT2101 Decision Making Methods and Tools
SEMESTER I 2020-2021
Assignment 2
Due: Friday, 23 October 2020, 11.59 pm

Instructions:

- Work with the same group members as you have for Assignment 1.
- Upload a softcopy as pdf file into Luminus folder Assignment2-Submission. Deadline, 11.59 pm, Friday, 23 October.
- Ensure that your answer include the names of all group members.

---

1. (10 points) Consider the following ABC Bank's employee database:

Employee No	Education Level	Job Type	Year Born (19xx)	Gender	Years Prior	Bonus above median
1	3	1	84	Male	1	no
2	1	5	72	Female	15	yes
3	1	1	75	Female	12	no
4	2	2	70	Female	17	yes
5	3	1	82	Male	5	no
6	3	1	86	Female	0	yes
7	3	1	83	Female	3	no
8	3	1	88	Male	2	yes
9	1	3	65	Male	24	yes
10	4	4	63	Male	32	yes

The variables in the dataset are:

- *Education Level:* a categorical variable with categories 1 (A level), 2 (polytechnic), 3 (bachelor's degree), 4 (post-graduate degree).
- *Job Type:* a categorical variable indicating the type of job the employee holds: 1 (management), 2 (sales), 3 (administration), 4 (office), 5 (miscellaneous support)
- *Year Born:* year employee was born, a continuous variable.
- *Gender:* Male or Female.
- *Years Prior:* number of years of work experience at another bank prior to working at ABC Bank, a continuous variable.
- *Bonus above median:* class label with value is "yes" if the employee receives a year-end bonus that is above the median bonus of the previous year, "no" otherwise.

We are building a binary classification tree to predict the class label *Bonus above median* using the values of the other variables in the data.

(a) (2 points) How heterogeneous are the samples in the dataset? Compute using the Gini index.

(b) (2 points) Using the values of the variable *Year Born*, what is the maximum reduction in the diversity that can be achieved? Measure the reduction in terms of Gini index.

(c) (2 points) If we consider splitting the data according to the values of the variable *Job Type*, how many possible splits are there?

(d) (2 points) If we consider splitting the data according to the values of the variable *Education Level*, how many possible splits are there? List all these possible splits.

(e) (2 points) Suppose the information from Employee No. 11 is now available, but with its value for the variable *Year Born* missing. Suggest 2 possible ways to handle the missing value so that this new data sample can be included for building the classification tree.

2. (10 points) We are interested in finding out who among the clients in a small bank has been given a 'GOOD CREDIT' rating; these clients are considered to be a safe credit risk. The following information from 12 clients are available:

AGE	PROF-EXP	INCOME	UNIVERSITY	GOOD-CREDIT
25	1	49	0	NO
29	5	45	1	NO
34	9	180	1	YES
35	8	125	0	YES
35	9	100	1	YES
35	10	81	0	NO
37	13	29	1	NO
45	19	34	1	NO
50	24	22	1	NO
53	27	72	1	YES
58	15	21	1	YES
65	39	105	0	YES

AGE, years of professional experience (PROF-EXP), annual INCOME (thousands of $) are continuous attributes, while UNIVERSITY = 0 if the client does not have a university degree, 1 otherwise.

A binary decision tree to differentiate between two groups of clients (GOOD-CREDIT YES vs GOOD-CREDIT NO) is proposed.

(a) (2 points) What is the entropy of the training samples with respect to the classification?

(b) (4 points) Suppose INCOME is to be the first attribute used for splitting the root node of the decision tree. What is the best cut-off point for splitting that will maximize the information gain?

(c) (4 points) Build a complete decision tree that classifies all the training data samples correctly.

3. (10 points) Goldking Food is considering the introduction of a new line of desserts. In order to produce the new line, the company is considering either a major renovation or a minor renovation of its current production facility. The following table shows the expected returns of the various alternatives that it can choose:

Alternative	Market condition	
	Favorable market	Unfavorable market
Major renovation	$2,500,000	-$800,000
Minor renovation	$1,000,000	-$200,000
Do nothing	$0	$0

Before making the final decision, Goldking Food can conduct a marketing research survey at a cost of $50,000. The effectiveness of similar research surveys in the past in predicting the actual nature of the market is shown in the table below.

	Actual State of Nature	
Survey results	Favorable market (FM)	Unfavorable market (UM)
Positive (PS)	0.80	0.40
Negative (NS)	0.20	0.60

(for example, when the market condition is favorable, the probability of a positive outcome in the survey is 80%)

Assume that without any survey information, the probability of an unfavorable market is twice the probability of a favorable market.

(a) (2 points) What is the decision that minimizes the expected opportunity loss (regret)?

(b) (1 point) Determine the Expected Value with Original Information.

(c) (1 point) Determine the Expected Value of Perfect Information.

(d) (6 points) Determine the Expected Value of Sample Information and the best course of action.