**CS412, Spring 2021**
**Assignment 1**
**Due: Mar 3rd, 2021, 23:59, Blackboard**

**Instructions:**

- The main submission of Assignment 1 must be uploaded on Blackboard. Your submission will be graded via Blackboard platform.
- Your PDF file should be named as **“ID_name(Chinese) _Assigment1.pdf”**. For example, if your name is 张三, name your file as “3xxx_张三_Assignment1.pdf”.
- Type out your solutions on a separate blank file (using Word, Latex, markdown, etc.), and do not forget to convert the file to PDF extension before submitting. You should not type out on the original pdf file.
- Every student has 3 late days in total for the assignments.
- You are to complete this assignment individually. We require you to:
  - Not explicitly tell each other the answers
  - Not to copy answers
  - Not to allow your answers to be copied

**Question 1. (4 points)**

Discuss whether each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender.

(b) Dividing the customers of a company according to their profitability.

(c) Computing the total sales of a company.

(d) Sorting a student database based on student identification numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

(f) Predicting the future stock price of a company using historical records.

(g) Monitoring the heart rate of a patient for abnormalities.

(h) Monitoring seismic waves for earthquake activities.

**Question 2. (4 points)**

Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

**Question 3. (2 points)**

For each of the following data sets, explain whether or not data privacy is an important issue.

a) Census data collected from 1900–1950.
b) IP addresses and visit times of Web users who visit your Website.
c) Images from Earth-orbiting satellites.
d) Names and addresses of people from the telephone book.

**Question 4. (4 points)**

Classify the following attributes as binary, discrete, or continuous. Further, classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some of the cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

*Example: Age in years.*
*Answer: Discrete, quantitative, ratio*

(a) Brightness in lumens as measured by a light meter.

(b) Brightness as measured by people's judgments.

(c) Percentage of ones in an m-by-n binary matrix (only 0/1 entries).

(d) Increase in profit of the current year over the profit obtained in the year 2000.

**Question 5. (6 points)**

a. For the following vectors, $\boldsymbol{x}$, and $\boldsymbol{y}$, calculate the indicated similarity or distance measures.

(i) $\boldsymbol{x}$ = (0, 1, 0, 1, 0), $\boldsymbol{y}$ = (1, 0, 1, 0, 1) cosine, correlation, Euclidean

(ii) $\boldsymbol{x}$ = (1, 1, 1, 1, 1), $\boldsymbol{y}$ = (1, 1, 1, 1, 10) Cosine, Euclidean, Correlation

Now, let us further explore the cosine and correlation measures.

b. If two objects have a cosine measure of 1, are they identical? Explain briefly.

c. Under what conditions (if any?), is the cosine measure between two vectors the same as their correlation? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

**Question 6. (6 points)**

(a) Suppose you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain briefly. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

(b) The collection of books in a library can be represented by a vector whose length is equal to the number of distinct books available at any library and whose elements indicate the number of copies of that book the library owns. You want to compare the similarity of two libraries based on the collection of their books.

(i) Which similarity measure is more suited for this task between cosine and correlation. Briefly justify your answer.

(ii) What is one strength and one weakness of Euclidean distance for this task? Briefly justify your answer.

**Question 7. (6 points)**

For the following questions, answer whether the given measure is a metric. Provide either a brief explanation, or short proof if it is, or a counterexample if you think it is not.

a. The proximity measure between two integers $x$ and $y$: $|x^2 - y^2|$

b. The proximity measure between two vectors $(x_1, y_1)$ and $(x_2, y_2)$: $(x_1 \times x_2)^2 + (y_1 - y_2)^2$

c. Hamming distance between two binary strings of length $n$.

**Question 8. (4 points)**

Answer the following questions with True/False. Give a brief explanation for your answers.

a) It is not a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute has outliers.

b) The correlation of the vectors (1, 1, 1, 1) and (2, 2, 2, 2) is 1.

c) Cosine Similarity is better than Correlation because it ignores 0-0 matches in non-binary vector data.

d) For binary vectors, mutual information is more like correlation than the Jaccard measure

**Question 9. (4 points)**

Let X be an unfair 6-sided die with probability distribution defined by P(X=1) = 1/2, P(X=2) = 1/4, P(X=3) = 0, P(X=4) = 0, P(X=5) = 1/8, and P(X=6) = 1/8. What is the entropy of the unfair 6-sided die?

**Question 10. (8 points)**

Consider the following joint distribution over random variables $X$ and $Y$:

| | | | | | |
|---|---|---|---|---|---|
| $y_1$ | 0.01 | 0.02 | 0.03 | 0.1 | 0.1 |
| $y_2$ | 0.05 | 0.1 | 0.05 | 0.07 | 0.2 |
| $y_3$ | 0.1 | 0.05 | 0.03 | 0.05 | 0.04 |
| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |

Remember to compute these in bits which means using base 2 for logarithms.

Hint: you can use python or excel to compute the following quantities, but you should give the necessary formula and process.

In the case of $P(x_i) = 0$ for some i, the value of the corresponding summand $0 \times \log_2 0 = 0$.

(a) What is the entropy $H(X)$, $H(Y)$?

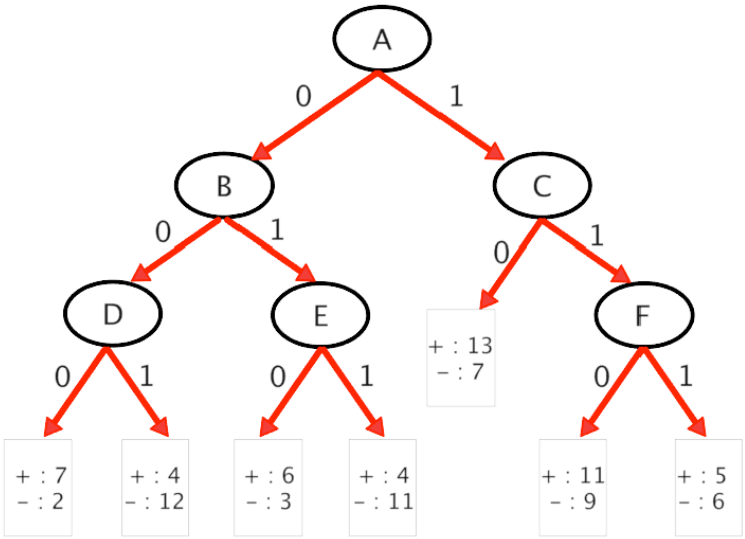
(b) What is the conditional entropy $H(X|Y)$ and $H(Y|X)$?

(c) What is the joint entropy $H(X, Y)$?

(d) What is the mutual information $I(X; Y)$?

**Question 12. (10 points)**

Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.




a) If each leaf node is labeled according to the majority class of the training instances that reach the node, compute the training error for the tree.

b) Estimate the generalization error for the tree using the pessimistic error rate approach, assuming the cost of each leaf is $\Omega = 0.5$.

c) Suppose the nodes labeled as E and F in the tree are replaced by their corresponding leaf nodes. Estimate the generalization error of the pruned tree using the pessimistic error rate approach. Compare with your answer from part (b) to determine whether the original tree should be pruned.

d) Using the validation set below, determine whether nodes E and F should be pruned by considering the validation error rate of the two trees.

| A | B | C | D | E | F | Class |
|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 0 | - |
| 0 | 1 | 1 | 1 | 1 | 1 | - |
| 0 | 0 | 0 | 1 | 0 | 1 | - |
| 1 | 0 | 1 | 0 | 1 | 1 | - |
| 0 | 1 | 1 | 1 | 0 | 0 | - |
| 1 | 0 | 1 | 1 | 1 | 1 | - |
| 0 | 1 | 0 | 1 | 0 | 0 | + |
| 1 | 0 | 1 | 1 | 0 | 1 | + |
| 0 | 1 | 1 | 0 | 0 | 0 | - |
| 0 | 1 | 0 | 1 | 0 | 1 | - |

**Question 12. (12 points)**

Consider the decision tree shown in Figure 12.1, and the corresponding training and test sets in Tables 12.1 and 12.2, respectively.

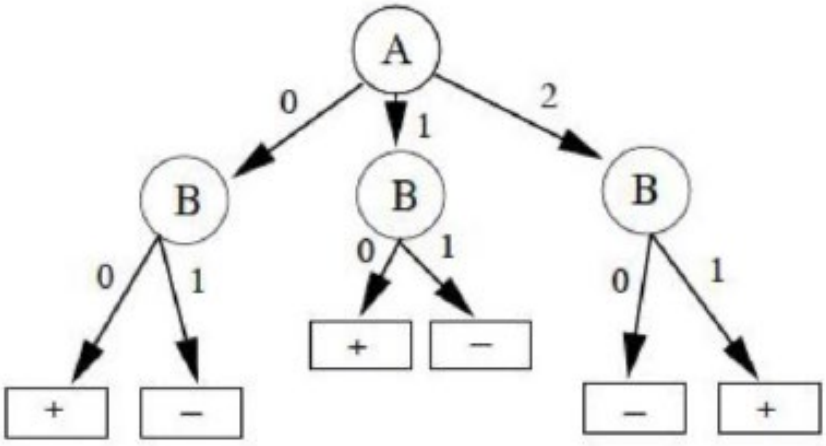



**Figure 12.1**: Decision tree

**Table 12.1**: Training set

| A | B | Number of + instances | Number of - instances |
|---|---|---|---|
| 0 | 0 | 5 | 3 |
| 0 | 1 | 3 | 4 |
| 1 | 0 | 22 | 7 |
| 1 | 1 | 7 | 32 |
| 2 | 0 | 2 | 5 |
| 2 | 1 | 6 | 4 |

**Table 12.2**: Test set

| A | B | Number of + instances | Number of - instances |
|---|---|---|---|
| 0 | 0 | 4 | 1 |
| 0 | 1 | 3 | 1 |
| 1 | 0 | 6 | 3 |
| 1 | 1 | 3 | 15 |
| 2 | 0 | 5 | 2 |
| 2 | 1 | 2 | 5 |

a) Estimate the generalization error rate of the tree using both the optimistic approach and the pessimistic approach. Use $\Omega = 2$ as the cost of adding a leaf node while calculating the pessimistic estimate.

b) Compute the error rate of the tree on the test set shown in table 12.2.

c) Figure 12.2 shows a pruned version of the original decision tree. Estimate the generalization error rate of this tree using both the optimistic approach and the pessimistic approach, as in Part (a). Also compute the error rate of this tree on the test set shown in table 12.2.

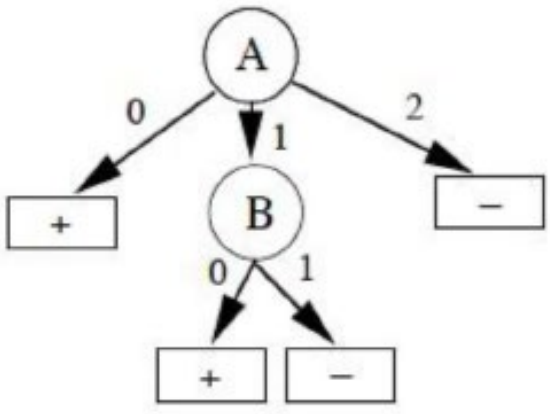



**Figure 12.2:** Pruned tree

**Question 13. (12 points)**

Consider the following set of training examples.

| $X$ | $Y$ | $Z$ | No. of Class C1 Examples | No. of Class C2 Examples |
|---|---|---|---|---|
| 0 | 0 | 0 | 5 | 40 |
| 0 | 0 | 1 | 0 | 15 |
| 0 | 1 | 0 | 10 | 5 |
| 0 | 1 | 1 | 45 | 0 |
| 1 | 0 | 0 | 10 | 5 |
| 1 | 0 | 1 | 25 | 0 |
| 1 | 1 | 0 | 5 | 20 |
| 1 | 1 | 1 | 0 | 15 |

(a) Compute a two-level decision tree using the greedy approach described in this chapter. Use the classification error rate as the criterion for splitting. What is the overall error rate of the induced tree?

(b) Repeat part (a) using X as the first splitting attribute and then choose the best remaining attribute for splitting at each of the two successor nodes. What is the error rate of the induced tree?

(c) Compare the results of parts (a) and (b). Comment on the suitability of the greedy heuristic used for splitting attribute selection.

**Question 14. (10 points)**

The following table summarizes a data set with three attributes A, B, C and two class labels +, −. Build a two-level decision tree.

| A | B | C | Number of Instances | |
|---|---|---|---|---|
| | | | + | − |
| T | T | T | 5 | 0 |
| F | T | T | 0 | 20 |
| T | F | T | 20 | 0 |
| F | F | T | 0 | 5 |
| T | T | F | 0 | 0 |
| F | T | F | 25 | 0 |
| T | F | F | 0 | 0 |
| F | F | F | 0 | 25 |

(a) According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

(b) Repeat for the two children of the root node.

(c) How many instances are misclassified by the resulting decision tree?

(d) Repeat parts (a), (b), and (c) using C as the splitting attribute.

(e) Use the results in parts (c) and (d) to conclude about the greedy nature of the decision tree induction algorithm.

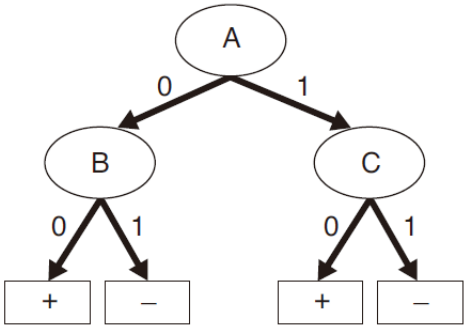
**Question 15. (6 points)**

Consider the decision tree shown in Figure 15.1.

(a) Compute the generalization error rate of the tree using the optimistic approach.

(b) Compute the generalization error rate of the tree using the pessimistic approach. (For simplicity, use the strategy of adding a factor of 0.5 to each leaf node.)

(c) Compute the generalization error rate of the tree using the validation set shown above. This approach is known as reduced error pruning.




Training:

| Instance | A | B | C | Class |
|---|---|---|---|---|
| 1 | 0 | 0 | 0 | + |
| 2 | 0 | 0 | 1 | + |
| 3 | 0 | 1 | 0 | + |
| 4 | 0 | 1 | 1 | – |
| 5 | 1 | 0 | 0 | + |
| 6 | 1 | 0 | 0 | + |
| 7 | 1 | 1 | 0 | – |
| 8 | 1 | 0 | 1 | + |
| 9 | 1 | 1 | 0 | – |
| 10 | 1 | 1 | 0 | – |

Validation:

| Instance | A | B | C | Class |
|---|---|---|---|---|
| 11 | 0 | 0 | 0 | + |
| 12 | 0 | 1 | 1 | + |
| 13 | 1 | 1 | 0 | + |
| 14 | 1 | 0 | 1 | – |
| 15 | 1 | 0 | 0 | + |

**Figure 15.1.** Decision tree and data sets