CS 412: Fall'23
Introduction To Data Mining

Assignment 2

(Due October, 13 11:59 pm)

- The homework is due on October 13 at 11:59 p.m. We will be using Gradescope for homework assignments. If you are still having issues with joining Gradescope, you may use the code shared on Canvas Announcements dated September 5. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. Unfortunately, We will NOT accept late submissions without a reasonable justification.

- Please use Slack or Canvas first if you have questions about the homework. You can also join office hours and/or send us emails. If you are sending us emails with questions on the problems, please start the subject with "CS 412 Fall'23: " and send the email to *all of us* (Arindam, Ruby, Hyunsik, Rohan, Kowshika, Sayar) for faster response.

- Please write down your solutions entirely by yourself and make sure the answers are clear. The homework should be submitted in PDF format; there is no need to submit source code about your computing. You are expected to typeset the solutions, i.e., **handwritten solutions will not be graded**. We encourage you to use Latex.

- For each question, you will NOT get full credit if you only give a final result. Please show the necessary steps and details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean.
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $S = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

- Please make sure to mark each question clearly. You may use as many pages as needed. Please do not change the order of the questions and answers.

- When submitting Assignment 2, Gradescope will ask you to select pages for each problem; please do this precisely!

1. (30 Points) Suppose a group of 20 sales price records has been sorted as follows:

$$5, 7, 7, 9, 10, 11, 13, 15, 20, 29, 35, 47, 50, 55, 60, 65, 72, 80, 87, 92$$

Answer the following questions:

(a) (6 points) Partition the data into four bins using the equal-frequency method. Apply the bin's mean to approximate each bin.

(b) (6 points) Partition the data into four bins with an equal-width method. Apply the bin's mean to approximate each bin.

(c) (6 points) Compare the effect of the equal-frequency binning 1a and equal-width binning 1b in terms of the quality of approximation based on the average variance of the bins.

(d) Normalize the data using the following methods:

(i) (6 points) Min-max, target interval [500-1000].

(ii) (6 points) Decimal scaling.

2. (25 Points) Ms. Jones, the fifth-grade teacher at Heartland Elementary School, is concerned about the number of students who failed the final exam in her class last year. To make proper adjustments to her teaching plans, she decided to train a decision tree on historical data she collected last year. Table 2 summarizes the Final Exam results for 13 students along with the following features: hobby, favorite color, practice time in hours, and submitted HW.

Hobby	**Favorite Color**	**Practice time in Hours**	**Submitted HW**	**Final Exam**
Painting	Red	5	Yes	Pass
Painting	Green	9	Yes	Pass
Painting	Blue	6	No	Pass
Music	Yellow	6	Yes	Pass
Music	Red	7	No	Pass
Swimming	Red	8	Yes	Pass
Swimming	Purple	5	Yes	Pass
Painting	Purple	9	No	Pass
Music	Red	8	No	Fail
Music	Red	8	No	Fail
Swimming	Blue	6	No	Fail
Swimming	Yellow	7	No	Fail
Music	Yellow	5	Yes	Fail

Table 1: Summary of data collected about 13 students' Final grade, hobby, favorite color, practice time in hours, and submitted HW.

Answer the following questions:

(a) (15 points) Compute the Information Gain for all of the features. Show all intermediate computations.

(b) (5 points) Which feature would the decision-tree building algorithm (using Information Gain) choose for the root?

(c) (5 points) Draw the full decision tree that would be learned for this dataset (with no pruning and using Information Gain).

Note for 2c

- *Hand drawing is not acceptable. You may use any drawing tool or write a script to generate the tree (no need to submit any script). You may create the sketch as an image and embed it in the Latex file.*
- *Using decision tree code, e.g., library to solve 2c is acceptable (given it produces the correct result). No need to submit the script.*

3. (15 Points) A Pharma startup at Illinois Research Park created a new affordable COVID-19 Antigen test to detect COVID-19 named `I-Test`. The first round of evaluating the correctness of the `I-Test` results showed 5% false positive, i.e., the `I-Test` displayed a positive result where, in fact, the person did not have COVID-19 (based on a PRC validation test). Moreover, the `I-Test` always showed a correct positive result when the test subject did have COVID-19 (also, the result was validated by a PRC test).

   Given a pool of test takers where one in thousand `I-Test` taker has COVID-19, suppose a person at random took the `I-Test`. If the test result is positive; what is the probability that the test taker DOES have COVID-19? Explain your answer.

   Show all solution steps including intermediate computations

4. (30 Points) The following dataset summarizes the outcome of a hypothetical COVID-19 predictor given three symptoms: fever, cough, and headache.

Fever	Cough	Headache	COVID-19 Outcome
No	No	Yes	Negative
No	Yes	No	Negative
Yes	Yes	No	Negative
No	No	Yes	Positive
Yes	Yes	Yes	Positive
Yes	No	No	Positive
Yes	Yes	No	Positive

Answer the following questions:

(a) (20 Points) Fit a Naïve Bayes model in the data, i.e., create the likelihood tables for all the features (Fever, Cough, and Headache).

(b) (10 Points) Using the likelihood from 4a classify the following example as Positive or Negative (Fever = No, Cough = No, Headache = Yes). Show all the intermediate calculations.