CS 412: Fall'22
Introduction To Data Mining

Assignment 5

(Due Monday, December 05, 11:59 pm)

- The homework is due at 11:59 pm on the due date. We will be using Gradescope for the homework assignments. You should join Gradescope using the entry code shared on Aug 3. Please do NOT email a copy of your solution. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- Please use Slack or Canvas first if you have questions about the homework. You can also come to our (zoom) office hours and/or send us e-mails. If you are sending us emails with questions on the homework, please start subject with "CS 412 Fall'22: " and send the email to *all of us* (Arindam, Mukesh, Chandni, Mayank, Hang, and Shiliang) for faster response.
- Please write down your solutions entirely by yourself and make sure the solutions are clear. The homework should be submitted in pdf format and there is no need to submit source code about your computing. You are expected to typeset the solutions. If your solution has any handwritten components, e.g., equations, tables, etc., please make sure they are legible—otherwise you may not get credit.
- The Assignment has an extra credit question worth 36 points. So, if you get everything correct, you can get 136 (out of 100) in this assignment.
- For each question, you will NOT get full credit if you only give out 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 $S = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

1. (32 points) This question considers decision tree learning for classification:

   (a) (5 points) Define Gain Ratio as a splitting criteria for constructing decision trees. Clearly describe all quantities (including SplitInfo) in the definition using suitable mathematical notation.

   (b) (5 points) Define Gini Reduction in Impurity as a splitting criteria for constructing decision trees. For Gini Reduction in Impurity, please assume you can do k-way split (not just 2-way splits) similar to Information Gain. Clearly describe all quantities in these definitions using suitable mathematical notation.

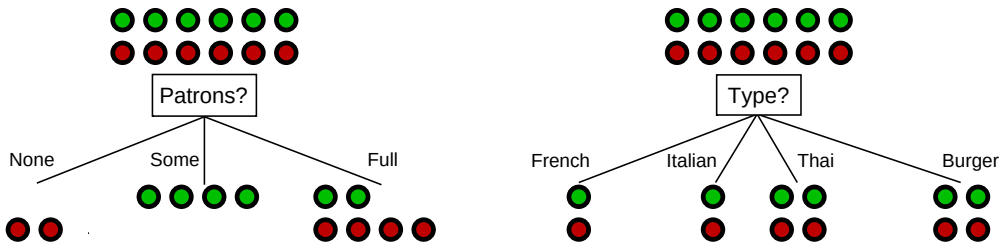

Figure 1: Decision tree splits with 'Patrons?' vs 'Type?'.

   (c) (10 points) Consider the two class classification problem with two possible choices for the root of a decision tree for the restaurant dataset, as shown in Figure 1. Compute the Gain Ratio for the attributes 'Patrons?' and 'Type?'. Please show all intermediate steps. Based on the Gain Ratio, which attribute will you split on at the root? Briefly explain your answer.

   (d) (12 points) Consider a two class classification problem with two possible choices for the root of a decision tree for the restaurant dataset, as shown in Figure 1. Consider a new splitting criterion Gini Reduction Ratio = Gini Reduction in Impurity / SplitInfo, similar to Gain ratio but based on Gini Redcution in Impurity rather than Information Gain. Compute the Gini Reduction Ratio (with k-way split) for the attributes 'Patrons?' and 'Type?'. Please show all intermediate steps. Based on the Gini Reduction Ratio, which attribute will you split on at the root? Briefly explain your answer.

2. (20 points) Suppose there are three kinds of candy bags:

   - $\frac{1}{4}$ are type h_1: 100% cherry candies,
   - $\frac{1}{2}$ are type h_2: 50% cherry candies and 50% lime candies,
   - $\frac{1}{4}$ are type h_3: 100% lime candies.

   Any candy bag of any type has exactly 100 candies at the beginning.

   We have one candy bag, but we do not know which type it is. We want to compute the posterior probabilities of the different types of bags as we draw candies out of the bag we have. The candies are **drawn without replacement**.

   (a) (10 points) We draw a candy (Candy_1) from the bag, and it turns out to be "lime." What are the posterior probabilities $p(h_1|\text{Candy}_1 = \text{lime}), p(h_2|\text{Candy}_1 = \text{lime}), p(h_3|\text{Candy}_1 = \text{lime})$ of each type of bag? Clearly explain your answer, e.g., how you are using Bayes rule, and show your calculations.

   (b) (10 points) We draw another candy (Candy_2) from the bag, and it turns out to be "lime." What are the posterior probabilities $p(h_1|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{lime}), p(h_2|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{lime}), p(h_3|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{lime})$ of each type of bag? Clearly explain your answer, e.g., how you are using Bayes rule, and show your calculations.

3. (25 points) This question considers training a naive Bayes classifier for 2-class classification using the dataset in Table 1. Each row refers to an apple instance with three categorical features (size, color, and shape) and one class label (whether the apple is good or not).

RID	Size	Color	Shape	Class: good apple
1	Small	Green	Irregular	No
2	Large	Red	Irregular	Yes
3	Large	Red	Circle	Yes
4	Large	Green	Circle	No
5	Large	Green	Irregular	No
6	Small	Red	Circle	Yes
7	Large	Green	Irregular	No
8	Small	Red	Irregular	Yes
9	Small	Green	Circle	No
10	Large	Red	Circle	Yes

Table 1: Apple classification dataset.

(a) (6 points) How many independent parameters[1] are required for training the naive Bayes classifier from this data set? Please explain your answer and enumerate all of them.

(b) (14 points) Estimate the values of these parameters based on the observations in Table 1. Please show the details of the computation for one of the conditional probability parameters to illustrate your understanding.

(c) (5 points) Given a new apple with features $x = (Large; Green; Circle)$, what is the estimated class posterior probabilities given x, i.e., $P(y = Yes \mid x)$ and $P(y = No \mid x)$? Please show details of your computation. Based on the class posterior probabilities, which class will naive Bayes predict for x? Briefly explain your answer.

[1]For a random variable X with two possible values, a and b, there is only one independent parameter say $P(X = a)$ since we have $P(X = b) = 1 - P(X = a)$.

4. (23 points) This question considers Random Forests (RFs).

   (a) (6 points) Briefly describe the three key parameters in RFs Forest-RI: d, the tree depth; m, the number of attributes randomly selected as candidates for splits; and T, the total number of trees.

   (b) (6 points) In the context of classification, clearly describe how RFs Forest-RI (random input selection) are trained and how prediction is done on a test point. Your answer can assume the use of the CART methodology without describing the methodology.

   (c) (6 points) RFs are built by bootstrap sampling, i.e., given an original set of samples of size n, the bootstrapped sample is obtained by sampling with replacement n times. Assuming n is large, what is the expected number of unique samples from the original set of n samples in the bootstrapped sample?

   (d) (5 points) Professor Very Random Forest claims to have a brilliant idea to make RFs Forest-RI more powerful: since RFs prefers trees which are diverse, i.e., not strongly correlated, Professor Forest proposes setting $m = 1$ for Forest-RI, where m is the number of random features used in each node of each decision tree. Professor Forest claims that this will improve accuracy while reducing variance. Do you agree with Professor Forest's claims? Clearly explain your answer.

Extra Credit.

1. (36 points) Professor Stewart Gilligan Griffin has developed two models for spam detection respectively based on (M1) a neural network and (M2) a time machine. The models were evaluated using $n = 10000 = a+b+c+d$ emails and based on the following confusion matrix:

		Prediction	
		Spam	Not Spam
Truth	Spam	a	b
	Not Spam	c	d

The two models had these specific values for the confusion matrix:

(M1) $a = 2588, b = 412, c = 46, d = 6954$,

(M2) $a = 2999, b = 1, c = 1, d = 6999$.

Clearly define the following quantities in terms of a, b, c, d in the confusion matrix and compute their numerical values (rounded to 3 places after the decimal) for M1 and M2.

(a) (6 points) Sensitivity.

(b) (6 points) Specificity.

(c) (6 points) Accuracy.

(d) (6 points) Precision.

(e) (6 points) Recall.

(f) (6 points) F1 score.