CS 412: Spring'21
Introduction To Data Mining

Assignment 5

(Due Tuesday, May 4, 11:59 pm)

- Feel free to talk to other students of the class while doing the homework. We are more concerned that you learn how to solve the problems than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Slack first if you have questions about the homework. You can also use CampusWire, send us e-mails, and/or come to our office hours. If you are sending us emails with questions on the homework, please cc all of us (Arindam, Carl, Jialong, and Qi) for faster response.
- The homework is due at 11:59 pm on the due date. We will be using GradeScope for collecting the homework assignments. Please submit your answers via GradeScope (`https://www.gradescope.com/`). Please do NOT email a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- The homework should be submitted in pdf format.
- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary details, calculation steps, and explanations as appropriate.

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

   (a) (8 points) Define Gain ratio and Gini impurity as a splitting criteria for constructing decision trees. Clearly describe all quantities in these definitions using suitable mathematical notation.

   (4 points) For dataset D,

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

   where n is the number of classes and p_j is the relative frequency of class j in D. If D is split on attribute A into D_1 and D_2, then

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

   Reduction in impurity:

$$\Delta gini(A) = gini(D) - gini_A(D)$$

   (4 points) For dataset D, The expected information for classifying a tuple in a dataset D is

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

   where m is the number of distinct values taken on by the attribute we are splitting on. When splitting on an attribute A,

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$

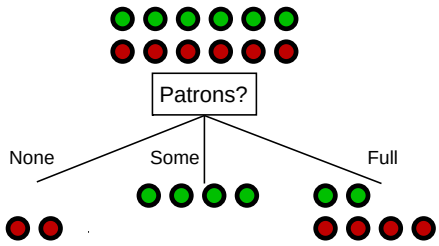
$$Gain(A) = Info(D) - Info_A(D)$$

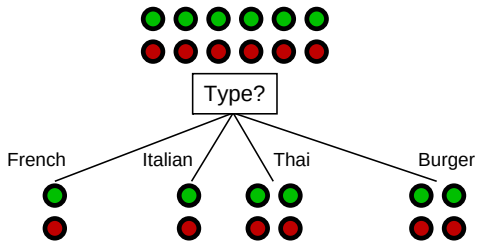
$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \frac{|D_j|}{|D|}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

   (b) (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 1b. Compute the Gini impurity for the attributes 'Patrons?' and 'Type?'. Based on the Gini impurity, which attribute will you split on at the root? Briefly explain your answer.

$$gini(D) = 1 - \left(\frac{6}{12}\right)^2 - \left(\frac{6}{12}\right)^2 = 1 - \frac{1}{4} - \frac{1}{4} = \frac{1}{2}$$





(5 points)

$$gini_{Patrons?}(D) = \frac{2}{12}gini(None) + \frac{4}{12}gini(Some) + \frac{6}{12}gini(Full)$$

$$= \frac{2}{12}\left(1 - \left(\frac{0}{1}\right)^2 - \left(\frac{1}{1}\right)^2\right)$$

$$+\frac{4}{12}\left(1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2\right)$$

$$+\frac{6}{12}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right)$$

$$= \frac{6}{12}\left(1 - \frac{1}{9} - \frac{4}{9}\right) = \frac{6}{12} \times \frac{4}{9} = \frac{2}{9}$$

(5 points)

$$gini_{Type?}(D) = \frac{2}{12}gini(French) + \frac{2}{12}gini(Italian) + \frac{4}{12}gini(Thai) + \frac{4}{12}gini(Burger)$$

$$= \frac{2}{12}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)$$

$$+\frac{2}{12}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right)$$

$$+\frac{4}{12}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$+\frac{4}{12}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$

$$= \left(\frac{2}{12} + \frac{2}{12} + \frac{4}{12} + \frac{4}{12}\right)\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) = \left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) = \frac{1}{2}$$

(2 points) Based on Gini Impurity, I will split based on 'Patrons?' This is because it has the smallest Gini impurity. Additionally, if we look at 'Type?', we can see that each split

has the same distribution as the dataset, so we aren't actually gaining any information by splitting.

$$\Delta gini(Type?) = gini(D) - gini_{Type?}(D) = \frac{1}{2} - \frac{1}{2} = 0$$

(c) (12 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 1b. Compute the Gain ratio for the attributes 'Patrons?' and 'Type?'. Based on the Gain ratio, which attribute will you split on at the root? Briefly explain your answer.

$$Info(D) = -\frac{6}{12}\log_2\frac{6}{12} - \frac{6}{12}\log_2\frac{6}{12} = 1$$

(5 points)

$$Info(None) = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$$Info(Some) = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$$

$$Info(Full) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = \frac{\log 3}{\log 8} + \frac{2\log 3}{3\log 2} - \frac{2}{3} = 0.918$$

$$= \frac{\log_2 3}{\log_2 8} + \frac{2}{3}\log_2 3 - \frac{2}{3} = \log_2 3 - \frac{2}{3}$$

$$Info_{Patrons?} = \frac{2}{12}Info(None) + \frac{4}{12}Info(Some) + \frac{6}{12}Info(Full)$$

$$= 0 + 0 + \frac{1}{2}\left(\log_2 3 - \frac{2}{3}\right)$$

$$Gain(Patrons?) = 1 - \frac{1}{2}\left(\log_2 3 - \frac{2}{3}\right) = \frac{4}{3} - \frac{\log_2 3}{2} = 0.541$$

$$SplitInfo_{Patrons?}(D) = -\frac{2}{12}\log_2\frac{2}{12} - \frac{4}{12}\log_2\frac{4}{12} - \frac{6}{12}\log_2\frac{6}{12} = \frac{2}{3} + \frac{\log_2 3}{2} = 1.459$$

$$GainRatio(Patrons?) = \frac{\frac{4}{3} - \frac{\log_2 3}{2}}{\frac{2}{3} + \frac{\log_2 3}{2}} = \frac{12}{4 + 3\log_2 3} - 1 = 0.371$$

(5 points)

$$Info(French) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Info(Italian) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$Info(Thai) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Info(Burger) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Info_{Type?} = \frac{2}{12}Info(French)+\frac{2}{12}Info(Italian)+\frac{4}{12}Info(Thai)+\frac{4}{12}Info(Burger) = 1$$

$$Gain(Type?) = 1 - 1 = 0$$

$$SplitInfo_{Type?}(D) = -\frac{2}{12}\log_2\frac{2}{12} - \frac{2}{12}\log_2\frac{2}{12} - \frac{4}{12}\log_2\frac{4}{12} - \frac{4}{12}\log_2\frac{4}{12} = \frac{2}{3} + \frac{\log_2 3}{2}$$

$$= \frac{\log_2 54}{\log_2 8} = \frac{1+3\log_2 3}{3} = 1.918$$

$$GainRatio(Type?) = \frac{0}{\frac{1+3\log_2 3}{3}} = 0$$

(2 points) Attribute 'Patrons?' has a larger gain ratio, so we will split on it. In particular, note that nothing is gained for 'Type?' since all the splits have the same distribution as D.

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

   - $\frac{1}{4}$ 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.

   We have one bag of candies, but we don't 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 with replacement.

   (a) (12 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.

   (4 points) $p(h_1|\text{Candy}_1 = \text{lime}) = \frac{p(\text{Candy}_1=\text{lime}|h_1)}{p(\text{Candy}_1=\text{lime})p(h_1)} = \frac{0*0.25}{0.5} = 0$

   (4 points) $p(h_2|\text{Candy}_1 = \text{lime}) = \frac{p(\text{Candy}_1=\text{lime}|h_2)}{p(\text{Candy}_1=\text{lime})p(h_2)} = \frac{0.5*0.5}{0.5} = 0.5$

   (4 points) $p(h_3|\text{Candy}_1 = \text{lime}) = \frac{p(\text{Candy}_1=\text{lime}|h_3)}{p(\text{Candy}_1=\text{lime})p(h_3)} = \frac{1*0.25}{0.5} = 0.5$

   (b) (8 points) We draw another candy (Candy_2) from the bag, and it turns out to be cherry. What are the posterior probabilities $p(h_1|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry}), p(h_2|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry}), p(h_3|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry})$ of each type of bag? Clearly explain your answer, e.g., how you are using Bayes rule, and show your calculations.

   (4 points) $p(h_2|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry}) = \frac{p(h_2,\text{Candy}_1=\text{lime},\text{Candy}_2=\text{cherry})}{p(\text{Candy}_1=\text{lime},\text{Candy}_2=\text{cherry})} = 1$

   (2 points) $p(h_1|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry}) = 0$

   (2 points) $p(h_3|\text{Candy}_1 = \text{lime}, \text{Candy}_2 = \text{cherry}) = 0$

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	No
9	Small	Green	Circle	No
10	Large	Red	Circle	Yes

Table 1: Apple classification dataset.

(a) (10 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.
(7 points) There are 7 independent parameters required for training the naive Bayes classifier (1 point for each parameter) : $p(Small|Yes)$, $p(Small|No)$, $p(Green|Yes)$, $p(Green|No)$, $p(Irregular|Yes)$, $p(Irregular|No)$, $p(Yes)$. (This is not the only answer.)
(3 points) Explain the answer. We can obtain all other parameters from the above parameters.

(b) (10 points) Estimate the values of these parameters based on the observations in Table 1.
(9 points) (1.5 point for each parameter)
$p(Small|Yes) = 0.25$
$p(Small|No) = 0.5$
$p(Green|Yes) = 0$
$p(Green|No) = 0.83$
$p(Irregular|Yes) = 0.25$
$p(Irregular|No) = 0.67$
(1 point) $p(Yes) = 0.4$

(c) (5 points) Given a new apple with features $x = (Small; Red; Circle)$, 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 point) $P(y = Yes \mid x) \propto P(Small \mid Yes) * P(Red \mid Yes) * P(Circle \mid Yes) * P(Yes) = 0.075$
(1 point) $P(y = No \mid x) \propto P(Small \mid No) * P(Red \mid No) * P(Circle \mid No) * P(No) = 0.01667$
(1 point) Correct answer: $P(y = Yes \mid x) = 0.075/(0.075 + 0.01667) = 0.8182$
(1 point) Correct answer: $P(y = No \mid x) = 0.01667/(0.075 + 0.01667) = 0.1818$

[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 $P(X = b) = 1 - P(X = a)$.

(1 point) Since $0.8182 > 0.1818$, a new apple with feature x is classified as a good apple.

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

   (a) (8 points) In the context of classification, clearly describe how RFs are trained and how prediction is done on a test point. Clearly describe the two key parameters in RFs: d, the tree depth, and m, the number of features (attributes) randomly selected for split.

   (2 points) How RFS are trained: The individual decision trees are generated using a random selection of attributes at each node to determine the split. More formally, each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest.
   (2 points) How prediction is done on a test point: During classification, each tree votes and the most popular class is returned as the prediction result.
   (2 points) The tree depth d: d is the maximum length of the tree. More information will be obtained with a larger d.
   (2 points) The number of features randomly selected for split m: m is the number of attributes to be used to determine the split at each node and m is much smaller than the number of available attributes.

   (b) (8 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?

   (3 points) Correct answer: the expected number of unique samples from the original set of n samples in the bootstrapped sample is 0.632*n.
   (5 points) Explanation: Since it is replacement, it is possible that some of the original data tuples will occur more than once in this sample. Each unique number has the $(1/n)$ probability of being selected, so the probability of not being chosen is $(1 - 1/n)$. Thus, the probability that a number not being selected at the end is $(1–1/n)^n$. When n is large, it approach $e^{-1} = 0.368$. Therefore, the remaining 0.632 * n unique number will form the training set.

   (c) (7 points) Professor Very Random Forest claims to have a brilliant idea to make RFs more powerful: since RFs prefers trees which are diverse, i.e., not strongly correlated, Professor Forest proposes setting $m = 1$, 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.

   (2 points) No, I don't.
   (5 points) If m = 1, splitting attribute is completely random chosen, and the node is not picking a good splitting attribute. This will decrease the quality of RFs and the accuracy will not be improved.

Extra Credit.

1. (35 points) Professor Stewart Gilligan Griffin has developed a Neural Spam Detector for detecting email spam using neural networks. The Neural Spam Detector was tested using 10,000 emails and the following confusion matrix was obtained: Please clearly define and

		Prediction		
		Spam	Not Spam	Total
Truth	Spam	2588	412	3000
	Not Spam	46	6954	7000
	Total	2634	7366	10000

compute the following quantities, and show details of your definition and calculations using $a = 2588, b = 412, c = 46, d = 6954$:

(a) (5 points) Sensitivity.
(2 points) Definition: Sensitivity is the true positive recognition rate. Sensitivity = TP/P = a / (a+b).
(3 points) Calculation: Sensitivity = a / (a+b) = 2588 / (2588+412) = 0.863.

(b) (5 points) Specificity.
(2 points) Definition: Specificity is the true negative recognition. Specificity = TN/N = d/(c+d).
(3 points) Calculation: Specificity = d / (c+d) = 6954/6954+46 = 0.993.

(c) (5 points) Accuracy.
(2 points) Definition: Percentage of test set tuples that are correctly classified. Accuracy = (TP+TN)/ALL = (a+d)/(a+b+c+d).
(3 points) Calculation: Accuracy = (TP+TN)/ALL = (a+d)/(a+b+c+d) = (2588+6954)/(2588+412+46+6954) = 0.9542.

(d) (5 points) Precision.
(2 points) Definition: Exactness: what % of tuples that the classifier labeled as positive are actually positive. Precision = TP/(TP+FP) = a/(a+c).
(3 points) Calculation: Precision = a/(a+c) = 2588/(2588+46) = 0.983.

(e) (5 points) Recall.
(2 points) Definition: Completeness: what % of positive tuples did the classifier label as positive. Recall = TP/(TP+FN) = a/(a+b).
(3 points) Calculation: Recall = a/(a+b) = 2588/(2588+412) = 0.863.

(f) (5 points) F1 score.
(2 points) Definition: harmonic mean of precision and recall, F1 = (2P*R)/(P+R) = (2 * Precision * Recall)/(Precision + Recall).
(3 points) Calculation: F1 score = (2 * Precision * Recall)/(Precision + Recall) = (2 * 0.983 * 0.863)/(0.983 + 0.863) = 0.919.

(5 points) If an email is predicted as spam, it is immediately deleted without notifying the end user. Professor Griffin claims that the ideal spam detector should have high recall at the

cost of having low precision. Do you agree with Professor Griffin? Explain your answer.
(2 points) Correct answer: No, I don't.
(3 points) Explanation: High recall means spam emails are predicted correctly. However, low precision means that many useful emails are also predicted as spam and deleted by mistake, which may cause huge lost to the customers. Therefore, having high recall at the cost of low precision is not acceptable.