

# CS 412: Fall'21

## Introduction To Data Mining

### Assignment 5

(Due Monday, December 6, 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.
- The homework is due at 11:59 pm on the due date. We will be using Gradescope for collecting the homework assignments. You should have been added to the Gradescope page for our class – if not, please email us. 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!
- Please use Slack first if you have questions about the homework. You can also send us e-mails, and/or come to our office (zoom) hours. If you are sending us emails with questions on the homework, please start subject with “CS 412 Fall'21: ” and send the email to *all of us* (Arindam, Yikun, Dongqi, Zhe, and Hang) for faster response.
- The homework should be submitted in pdf format and there is no need to submit source code with details of your computation.
- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary 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  $\mathcal{X} = \{x_1, \dots, x_n\}$ , 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  $\chi^2$  statistic?

**A:** For a categorical variable taking  $k$  possible values, if the expected values are  $e_i, i = 1, \dots, k$  and the observed values are  $o_i, i = 1, \dots, k$ , then the  $\chi^2$  statistic can be computed as:  $\chi^2 = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$ . 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  $\chi^2 = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$ .

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. For **Gini 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.

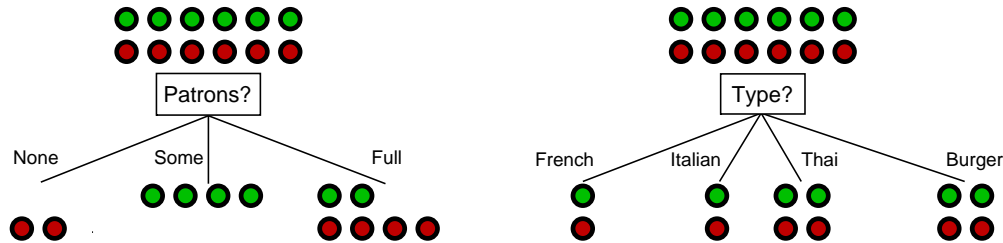


Figure 1: Decision tree splits with 'Patrons?' vs 'Type?'.

- (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 1. Compute the **Gini impurity** (with  $k$ -way split) for the attributes 'Patrons?' and 'Type?'. Please show all intermediate steps. Based on the **Gini impurity**, which attribute will you split on at the root? Briefly explain your answer.
- (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 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.

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 ( $\text{Candy}_1$ ) 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) (8 points) We draw another candy ( $\text{Candy}_2$ ) 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.

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<sup>1</sup> are required for training the naive Bayes classifier from this data set? Please explain your answer and enumerate all of them.
- (b) (10 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  $x = (Small; Red; Circle)$ , what is the estimated class posterior probabilities given  $x$ , i.e.,  $P(y = Yes | x)$  and  $P(y = No | x)$ ? Please show details of your computation. Based on the class posterior probabilities, which class will naive Bayes predict for  $x$ ? Briefly explain your answer.

---

<sup>1</sup>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) (5 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.
  - (b) (3 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.
  - (c) (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?
  - (d) (7 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. (14 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  $a + b + c + d$  emails and the following confusion matrix was obtained:

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

Please clearly define the following quantities using  $a, b, c, d$  from the above table, and compute their numerical values (rounded to 3 places after the decimal) when  $a = 2588, b = 412, c = 46, d = 6954$ :

- (a) (2 points) Sensitivity.
- (b) (2 points) Specificity.
- (c) (2 points) Accuracy.
- (d) (2 points) Precision.
- (e) (2 points) Recall.
- (f) (2 points) F1 score.

(2 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.