IS603 Midterm Exam, due 7:00 pm ET 4/5/2022

Total points: 20

Name: Rohan Salvi

ID number: VH10935

**MANDATORY ACADEMIC INTEGRITY AGREEMENT:**

**I will not discuss the exam with anyone except Dr. Quader until after it finishes at 7:00 pm ET on 4/5/2022. All of my work is my own and is not copied from other sources (even if paraphrased). I have acknowledged and cited all sources. I have not committed any act of cheating, fabrication, or plagiarism, and I have not helped others to do so. I have not used software to attempt to hide plagiarism. I will delete my copy of the exam afterward.**

**I understand that there is a zero-tolerance policy on plagiarism, and any violations will result in a 0 grade on the exam. I understand that academic integrity violations may incur additional consequences such as a failure grade in the entire course and further disciplinary action.**

Signature, certifying that I agree to the above statement: Rohan Salvi

Today’s date: 4//5/2022

**Your exam submission will not be accepted without a signature for the above.** You may attach a signed copy of this sheet or copy the complete statement to another document and sign it there. This exam is open book. You may consult other sources, including the textbook, the course slides, homework assignments and the internet, but you must follow the policies in the academic integrity agreement. For definitions of academic integrity terms including cheating, fabrication, plagiarism, and dishonesty, see [https://www.umbc.edu/policies/pdfs/UMBC%20III%201.10.03%20Undergraduate%](https://www.umbc.edu/policies/pdfs/UMBC%20III%201.10.03%20Undergraduate%20Student%20Academic%20Conduct%20Policy.pdf) [20Student%20Academic%20Conduct%20Policy.pdf](https://www.umbc.edu/policies/pdfs/UMBC%20III%201.10.03%20Undergraduate%20Student%20Academic%20Conduct%20Policy.pdf) .

Again, **you may not discuss the exam with anyone.**

1.

a. It is a supervised learning task, since the data is labeled, we can call it into supervised learning task. Supervised learning needs supervision to train the model, which is similar to as a student learns things in the presence of a teacher.

b.

I would say the more the dataset the better result for prediction. Another issue that can be addressed that If an individual or legal entity would like to use personal data, one needs informing and explicitly expressed consent of what personal data moves to whom, when, and for what purpose from the owner of the data. Also the data should be kept confidential to electronics store.

c.

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy. Choose attribute with the largest information gain as the decision node, divide the dataset by its branches and repeat the same process on every branch.

d.

p(no) = 5/14

p(yes) = 9/14

entropy (parent)= *−p*1 log2 *p*1 *− p*2 log2 *p*2

=-7/14 \* log2(7/14) - 7/14 \* log2(7/14)

=0.94

e.

To calculate entropy for student = yes

p(yes)= 6/7 and p(no)=1/7

entropy(yes)= -p1log2p1- p2log2p2

= -(6/7)log2(6/7)-(1/7)log2(1/7)

= 0. 59167

f.

p(no)= 4/7

p(yes)=3/7

entropy(no)= -p1log2p1-p2log2p2

=-(4/7)log2(4/7)- (3/7)log2(3/7)

= 0.985

g.

Information Gain(IG)= entropy(parent) − [p(c1) × entropy(c1) + p(c2) × entropy(c2) + . . .]

Entropy(parent)= 0.94

p(c1)=7/14 entropy(c1)=0. 59167

p(c1)=7/14 entropy(c1)=0. 59167

p(c2)=7/14 entropy(c2)= 0.985

IG=0.94 - [(7/14)\*0.59+(7/14)\*0.98]

IG=0.155

2.

1. The supervised learning category includes decision tree classifiers. They can be used to address problems involving regression and classification. I would use the Confusion Matrix approach to estimate the classification accuracy for new consumers.
2. An N x N matrix is used to evaluate the performance of a classification model, where N is the number of target classes. The matrix compares the actual goal values to the machine learning model's predictions. This provides us with a comprehensive picture of how well our classification model is working and the types of errors it makes.
3. Important predictive analytics such as recall, specificity, accuracy, and precision are visualized using confusion matrices. Confusion matrices are important because they allow you to compare values such as True Positives, False Positives, True Negatives, and False Negatives in a straightforward manner.

d) The curve I choose to visualize the performance of the model is Profit curve. The relationship between a firm's economic profit and the quantity of output sold is visually represented by a curve.

e) Profit curves assist you in estimating a model's commercial impact. There is an asymmetry between the reward of successful predictions and the penalty (or cost) of wrong predictions in many classification issues. The average profit chart aids in the evaluation of a model based on your supplied costs or benefits, allowing you to observe how earnings fluctuate as different inputs are changed.

f) Profit curves are important for determining profitability. For example, profit yields are displayed along the x-axis in descending order for a population dimension such as Customers or Products. The most profitable customer is on the far left, while the least profitable customer is on the far right.

3.

a) The benefit (cost) for this ad is $5.

b) The benefit (cost) for this ad is $0.

c) The benefit (cost) for this ad is $195.

d) The benefit (cost) for this ad is $0.

e)

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Yes | 195 | -5 |
| No | 0 | 0 |