Lastname: Rivas S.S. #: 766671137

Firstname: Abelardo

ISC4933/5935 Data Mining

Mid Term Examination

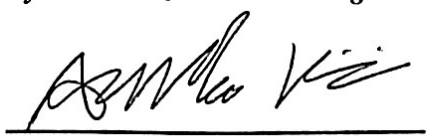
Fall 2020

Instructions for this test:

1. Put your name and SS at the top of each page, especially this page.
2. Read each problem carefully and be sure to provide the answer requested.
3. Use the provided pages only. Do *not* use any extra paper.
4. SHOW ALL WORK.
5. Do not cheat.

Upon completion of the test please sign the following statement:

I have neither given nor received aid from any unauthorized source during this exam.



Problem 1: Both UGs and Grads

(a) Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity: (i) Number of courses registered by a student in a given semester, (ii) Speed of a car (in miles per hour), (iii) Decibel as a measure of sound intensity (see `http://en.wikipedia.org/wiki/Decibel`) and (iv) Saffir-Simpson Hurricane Scale (see `http://www.nhc.noaa.gov/aboutsshs.shtml`).

(b) For the following vectors $\mathbf{x} = (-7, 8, -10, 5)$ and $\mathbf{y} = (-2, 7, -11, 0)$ determine the correlation, cosine and Euclidean measures!

a) i) # of courses. discrete
quantitative, ratio, ^ because it is a count

ii) Speed of a car (mph).
quantitative, ratio, continuous.

iii) Decibel
quantitative, interval, continuous. Differences matter, but not ratio.

iv) S-S Hurricane Scale
qualitative, ordinal, discrete. Has order.

b) $X = (-7, 8, -10, 5)$ $Y = (-2, 7, -11, 0)$

$\bar{X} = -1$ $\bar{Y} = -1.5$

$s_X = \sqrt{\frac{1}{3}\sum_{k=1}^{4}(X_k - \bar{X})^2} = 8.83$

$s_Y = \sqrt{\frac{1}{3}\sum_{k=1}^{4}(Y_k - \bar{Y})^2} = 8.77$

$cov(X,Y) = \frac{1}{3}\sum_{k=1}^{4}(X_k - \bar{X})(Y_k - \bar{Y}) = 58$

$corr(X,Y) = \frac{58}{8.83 \times 8.77} = .748$

$cos(X,Y) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{152}{(15.42)\cdot(13.19)} = .747$

$d = \sqrt{(-7+2)^2 + (8-7)^2 + (-10+11)^2 + (5+0)^2} = 7.21$

$X \cdot Y = (-7 \cdot -2) + (8 \cdot 7) + (-10 \cdot -11) + (5 \cdot 0) = 152$

$\|X\| = (-7 \cdot -7) + (8 \cdot 8) + (-10 \cdot -10) + (5 \cdot 5) = \sqrt{238} = 15.42$

$\|Y\| = (-2 \cdot -2) + (7 \cdot 7) + (-11 \cdot -11) + (0 \cdot 0) = \sqrt{174} = 13.19$

Problem 2: Both UGs and Grads

Consider the training dataset given below. M, N, L are the attributes and Y is the class variable.

M	N	L	Y
0	1	0	Yes
1	0	1	Yes
0	0	0	No
1	0	1	No
0	1	1	No
1	1	0	Yes

(a) Can you draw a decision tree having 100% accuracy on this training set? If your answer is yes, draw the decision tree. If your answer is no, explain why?

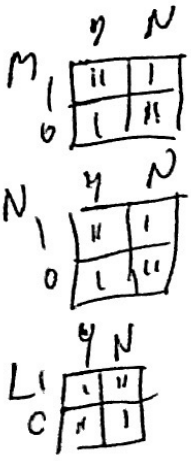
(b) Which attribute among M, N and L has the highest information gain? Explain your answer.

(c) You are given a collection of datasets that are linearly separable. Is it always possible to construct a decision tree having 100 % accuracy on such datasets? True or False. Explain your answer.

a) No because of the entries with M=1 N=0 and L=1 one both classified as Yes and No. There is no split possible with 100% accuracy.

b) $I_{parent} = -(3/6 \log_2(3/6) + 3/6 \log_2(3/6)) = 1$

$Entropy(M) = -3/6(2/3 \log_2(2/3) + 1/3 \log_2(1/3)) +$
$-3/6(1/3 \log_2(1/3) + 2/3 \log_2(2/3)) = .9182$



By symmetry all attributes have the same entropy and therefore all the same information gain.

c) False True by counter example, positives

+	−
−	+

No way to split + and − and negatives, in a linearly separable example, you can draw lines or planes that separate the classes. The lines/planes are the splits the tree makes.

Problem 3: Grads and UG optional

Suppose you are given a census data, where every data object corresponds to a household and the following continuous attributes are used to characterize each household: total household income, number of house- hold residents, property value, number of bedrooms, and number of vehicles owned. Suppose we are interested in clustering the households based on these attributes.

(a) Explain why cosine is not a good measure for clustering the data.

(b) Explain why correlation is not a good measure for clustering the data.

(c) Explain what preprocessing steps and corresponding proximity measure you should use to do the clustering.

Problem 4: Grads

Consider the following [illegible] measure

$$d(x, y) = 1 - s(x, y) \tag{1}$$

[illegible] the [illegible] similarity between two data objects, x and y. Does the distance [illegible] the [illegible] symmetry, and triangle inequality properties? For each [illegible] Assume x and y are non-negative vectors (e.g., term [illegible]

Problem 5: UGs and Grads

Consider an attribute X of a data set that takes the values $\{x_1, \cdots, x_9\}$ (sorted in increasing order of magnitude). We apply two methods (equal interval width and equal frequency) to discretize the attribute into 3 bins. The bins obtained are shown below:

Equal Width: $\{x_1, x_2, x_3\}, \{x_4, x_5, x_6, x_7, x_8\}, \{x_9\}$

Equal Frequency: $\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}, \{x_7, x_8, x_9\}$

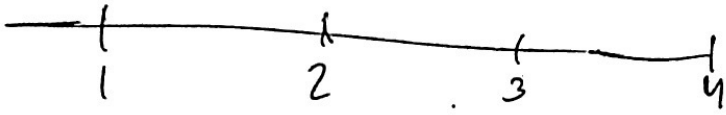
Explain what will be the effect of applying the following transformations on each discretization method, i.e., whether the elements assigned to each bin can change if you discretize the attribute after applying the transformation function below. Note that $\bar{X}$ denotes the average value and σ_x denotes standard deviation of attribute X.

(a) $X \rightarrow X - \bar{X}$ (i.e., if the attribute values are centered).

(b) $X \rightarrow \frac{X-\bar{X}}{\sigma_x}$ (i.e., if the attribute values are standardized).

(c) $X \rightarrow \exp\left[\frac{X-\bar{X}}{\sigma_x}\right]$ (i.e., if the values are standardized and exponentiated).

X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9 equal width



Example

$X_1 = 1.1$ $X_2 = 1.2$ $X_3 = 1.5$ $X_4 = 2.1$

$X_5 = 2.3$ $X_6 = 2.5$ $X_7 = 2.75$ $X_8 = 2.9$ $X_9 = 3.5$

a) Centering the values will only shift the values by the mean to the left. Meaning both methods will have the same ~~amount~~ number of elements in each bin.

~~b) Standardizing the values~~

b) Standardizing the values ~~centers~~ centers them around the mean and puts them in terms of how far away from the mean they are. This can squish or stretch the values, meaning equal width bins can have a different number of elements in each bin but, equal frequency will stay the same.

c) Applying the exponential to the standardized values will change the number of elements in each bin only if standardizing the values also changed it. (only for equal width).

Problem 4: Grads

Consider the following distance measure:

$$d(x, y) = 1 - c(x, y) \tag{1}$$

where $c(x; y)$ is the cosine similarity between two data objects, x and y. Does the distance measure satisfy the positivity, symmetry, and triangle inequality properties? For each property, show your steps clearly. Assume x and y are non-negative vectors (e.g., term vectors for a pair of documents).

Problem 6: UGs and Grads

(a) Compute the GINI-gain for the following decision tree split. Assume the parent node is $(12, 4, 6)$ and the children nodes are $(3, 3, 0), (9, 1, 0), (0, 0, 6)$.

(b) Assume there are 3 different classes and 50% of the examples belong to class 1. and 25% of the examples belong to class 2 and class 3, respectively. Compute the entropy of this class distribution, giving the exact number not only the equation.

(c) Why is the decision tree learning algorithms a greedy algorithm?

(d) Why is pruning important when using decision trees? What is the difference between pre-pruning and post pruning?

a) Gini parent: $1 - \left(\left(\frac{12}{22}\right)^2 + \left(\frac{4}{22}\right)^2 + \left(\frac{6}{22}\right)^2\right) = .595$

Gini A $= 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2 + \left(\frac{0}{6}\right)^2\right) = .5$

Gini B $= 1 - \left(\left(\frac{9}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{0}{10}\right)^2\right) = .18$

Gini C $= 1 - (0 + 0 + 1) = 0$

$\Delta = .595 - \left(\frac{6}{22}.5 + \frac{10}{22}.18 + \frac{6}{22}.0\right) = .376$

b) Entropy $= \left(\frac{-1}{2}\log_2(1/2) + \frac{-1}{4}\log_2(1/4) + \frac{-1}{4}\log_2(1/4)\right) =$ 1.5

c: Because it chooses the attribute with the highest gain as the choice for the child node.

d) To handle overfitting, aka minimizing the generalization error. Prepruning changes the stopping criterion to a threshold of the gain ratio (stop when the gain ratios are under this value). Post pruning allows for the full formed tree to form and removes internal nodes from the bottom-up until the gain is no longer improved.

Problem 7: Grads

(a) Construct a multilayer artificial neural network with one hidden layer and 8 hidden layer neurons.

(b) The necessary number of neurons in a hidden layer is always problem-oriented. Discuss the advantages to remove or to add iteratively neurons from the hidden layer.

(c) Neural networks are known to be function approximators. Describe why this is the case.

(d) Describe overfitting, evaluation of the performance of a classifier and methods for comparing the classifiers (Chapter 4). Name for each strategies.