Data Mining, CSCI 347, Fall 2019
Exam 2, Nov. 1

1. The data mining method for classifying a new instance by determining its nearest neighbor and matching the prediction of that neighbor is most likely to be called: (4 pts.)
    a. Clustering
    b. Covering algorithms
    c. Mining association rules
    d. Statistical modeling
    e. Instance based learning

2. Which of the following is most appropriate for numeric values? (4 pts.)
    a. Tables
    b. Linear models
    c. Trees
    d. Rules
    e. Clusters

3. Input into data mining algorithms can best be described by: (4 pts.)
    a. A wide variety of real world data collected in a wide variety of ways
    b. Structured data consisting of attribute values and relationships
    c. Individual, independent records, which provide values for the same set of attributes
    d. Data normalized to reduce repetition
    e. Data normalized in order to facilitate comparison of attributes

4. Choose the term which best describes a measurement where values are ordered, measured in fixed and equal units, and zero is defined: (4 pts.)
    a. Interval quantity
    b. Ratio quantity
    c. Normalized quantity
    d. Nominal quantity
    e. Ordinal quantity

5. Choose the term that best describes a measurement where values are ordered but no distance between values is defined: (4 pts.)
    a. Interval quantity
    b. Ratio quantity
    c. Normalized quantity
    d. Nominal quantity
    e. Ordinal quantity

Short Answer

6. Learning via linear regression finds weights to minimize the squares of the differences between the actual and predicated values in the formula:

$x = w_0 + w_1 * a_1 + w_2 * a_2 + \ldots + w_n * a_n$

where there are k attributes $a_1, \ldots, a_k$

That is, minimizing the squares of the differences:

$$\sum_{j=1}^{m} \left(x^{(j)} - \sum_{i=0}^{n} w_i * a_i^{(j)} \right)^2$$

Tell what is meant by the following: (8 pts.)

m

n

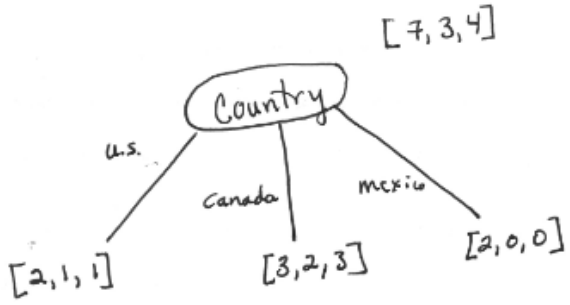
$x^{(j)}$

$a_i^{(j)}$

7. Describe what is meant by a "balanced" training set? (5 pts.)

8. Decision trees are created by selecting a root, splitting on that root and recursively repeating the process on the resulting subsets of instances. Measurements of information grain, gain ratio or gini ratio (which we didn't study) can be used to select the root. Give the formula for the gain ratio of a split. Describe each element in your formula. (5 pts.)

9. Calculate the gain ratio of the following split, showing all work. (You don't have to perform the logarithms and arithmetic.) (7 pts.)



10. State all of the association rules that can be created from the 2-item sets

attr1 = x, attr2 = y.

(Note, all 2-items are to be used in each rule.) (10 pts).

11. Consider the following dataset.

Relation: TestMaterialStrength

No.	1: Measurement1	2: Measurement2	3: Weight	4: Strength	5: **Class**
	Numeric	Numeric	Numeric	Numeric	Nominal
1	1.0	5.0	27.0	15.0	0
2	6.5	-2.5	27.0	5.0	0
3	-0.5	4.5	32.0	9.2	1
4	0.0	0.0	0.0	0.0	1
5	0.5	2.0	23.0	4.2	1

What class would be predicted for the testing instance:

Measurement1 = 1.5
Measurement2 = 0.5
Weight = 29
Strength = 5.0

a. Using instance-based learning and the Euclidean distance. Show your work. (5 pts.)

b. Using instance-based learning and the Manhattan distance. Show your work. (5 pts.)

Essay Questions

For the essay questions you may use your notes, the web and books. Do not discuss the problems with anyone aside from me. These questions do not have a single best answer. They will be graded on the logic, clarity and completeness of your answers. Ethical decision-making needs to be rationally addressed, using logical reasoning based on facts and commonly held values.

12. In one of two sentences, give the significant characteristics of each of the following learning algorithms. Include clustering, if your response to the second part of this question includes clustering. (5pts.)

    Statistical learning (Naive Bayes)

    Decision trees

    Classification Rules

    Association Rules

    Linear regression

    Instance based learning

    Clustering

    Presentations were made on the following special topics:
    Stream learning, natural language processing, image mining, web mining (content, usage and structure), recommender systems, psycholinguistic data mining, mining in adversarial situations.

    If web mining is seen as 3 separate topics, this gives 9 types of mining. For 4 of these topics, discuss the type of data mining algorithms discussed in class that are most likely to be helpful. In your discussion refer to characteristics that you have listed for the algorithm. (20 pts.)

13. In the book "Weapons of Math Destruction" Cathy O'Neil talks about mathematical models which are opaque, affect large numbers of people, and are potentially damaging, begin used to:
    - Determine credit worthiness used to grant/deny loans
    - Screen applications and sort resumes, so human resource personnel don't need to review all applications/resumes
    - Help policing, sentencing and setting parole, determining the likelihood of recidivism
    - Evaluate workers, teacher quality
    - Rand colleges
    - Target voters

    In the Data Science Code of Professional Conduct, Rule 8 labeled "Data Science Evidence, Quality of Data and Quality of Evidence" states:
    (h) A data scientist shall use reasonable diligence when designing, creating and implementing machine learning systems to avoid harm. …

    Discuss the obligations of a data scientist regarding dangerous algorithms, and what, if anything ought to be done. (10 pts.)