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 - m is the number of instances in the dataset

n - n is the number of attributes in the dataset

$x^{(j)}$ - the actual value of the jth instance

$a_i^{(j)}$ - the value of the ith attribute in the jth instance.

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

A training set is balanced if it has equal numbers of each of the outcomes (i.e. there are the same number of instances for each class value)

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.)

Gain ratio =
(info. before splitting – weighted info. of each branch after splitting)
/ intrinsic info

(info. before splitting – weighted info. of each branch after splitting) is called the "information gain"

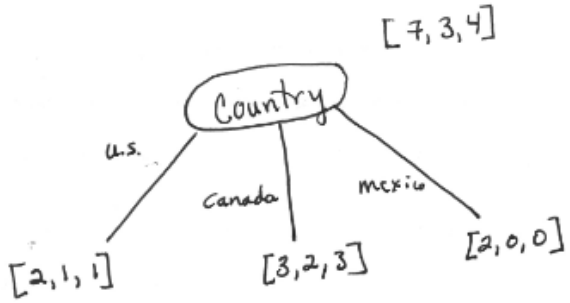
The information of a split, say for 3 class values is:

$$\text{Info}([a,b,c]) = \text{entropy}(a/(a+b+c), b/(a+b+c), c/(a+b+c)) \text{ bits}$$
$$= - a/(a+b+c) \log_2(a/(a+b+c)) +$$
$$- b/(a+b+c) \log_2(b/(a+b+c)) +$$
$$- c/(a+b+c) \log_2(c/(a+b+c))$$

The information before and after the splitting are split by the class values. The intrinsic information is simply the branching of the split, ignoring class values.

The weights are the proportion of the records.

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



Information before the split:

info([7, 3, 4]) = -7/14 $\log_2$(7/14) – 3/14 $\log_2$(3/14) – 4/14 $\log_2$(4/14)

= 0.5 + 0.476 + 0 .516

= 1.492

Weighted information after the split:

4/14 * info([2, 1, 1])

= 4/14 (-2/4 $\log_2$(2/4) – 1/4 $\log_2$(1/4) – 1/4 $\log_2$(1/4))

= 4/14 (0.5 + 0.5 + 0.5)

= 0.428 bits

8/14 * info([3, 2, 3])

= 8/14 (-3/8 $\log_2$(3/8) – 2/8 $\log_2$(2/8) – 3/8 $\log_2$(3/8))

= 8/14 (0.531 + 0.5 + 0.531)

= 0.892 bits

2/14 * info([2, 0, 0])

= 0 bits

Intrinsic value is info([4, 8, 2])

= -4/14 $\log_2$(4/14) – 8/14 $\log_2$(8/14) – 2/14 $\log_2$(2/14))

= 0.516 + 0.461 + 0.401

= 1.378

Overall gain ratio = $\frac{1.492 - (0.428 + 0.892 + 0)}{1.378}$

= $\frac{1.492 - 1.32}{1.378}$

= 0.125

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).

IF true THEN attr1 = x AND attr2 = y

IF attr1 = x THEN attr2 = y

IF attr2 = y THEN attr1 = x

In general, for an n-item set, 2^n-1 rules can be generated.

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.)

Distance from test instance to instance 1:

sqrt $((1.5–1.0)^2 + (0.5\text{-}5.0)^2 + (29\text{-}27)^2 + (5.0\text{-}15.0)^2)$
sqrt $((0.5)^2 + (-4.5)^2 + (2)^2 + (-10.0)^2)$
sqrt (124.5)
11.2

Distance from test instance to instance 2:

sqrt $((1.5–6.5)^2 + (0.5\text{- }-2.5)^2 + (29\text{-}27)^2 + (5.0\text{-}5.0)^2)$
sqrt $((-5.0)^2 + (3.0)^2 + (2)^2 + (0)^2)$
sqrt (38)
6.2

Distance from test instance to instance 3:

sqrt $((1.5\text{– }-0.5)^2 + (0.5\text{-}4.5)^2 + (29\text{-}32)^2 + (5.0\text{-}9.2)^2)$
sqrt $((2.0)^2 + (-4.0)^2 + (3)^2 + (-4.2)^2)$
sqrt (46.6)
6.8

Distance from test instance to instance 4:

sqrt $((1.5–0.0)^2 + (0.5\text{-}0.0)^2 + (29\text{-}0)^2 + (5.0\text{-}0.0)^2)$
sqrt $((1.5)^2 + (0.5)^2 + (29)^2 + (5.0)^2)$
sqrt (868.5)
29.5

Distance from test instance to instance 5:

sqrt $((1.5–0.5)^2 + (0.5\text{-}2.0)^2 + (29\text{-}23)^2 + (5.0\text{-}4.2)^2)$
sqrt $((1.0)^2 + (-1.5)^2 + (6)^2 + (0.8)^2)$
sqrt (39.89)
6.3

Test instance is closest to instance 2, which has a class value of 0, so 0 is predicted.

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

Distance from test instance to instance 1:

|1.5–1.0| + |0.5-5.0| + |29-27| + |5.0-15|
0.5 + 4.5 + 2 + 10.0
17

Distance from test instance to instance 2:

|.5–6.5|+ |0.5- -2.5| + |29-27| + |5.0-5.0|
|-5.0| + |3.0| + |2| + |0|
10

Distance from test instance to instance 3:

|1.5– -0.5| + |0.5-4.5| + |29-32| + |5.0-9.2|
|2.0|+ |-4.0| + |3| + |-4.2|
13.2

Distance from test instance to instance 4:

|1.5–0.0| + |0.5-0.0| + |29-0| + |5.0-0.0|
|1.5| + |0.5| + |29| + |5.0|
36

Distance from test instance to instance 5:

|1.5–0.5| + |0.5-2.0| + |29-23| + |5.0-4.2|
|1.0| + |-1.5| + |6| + |0.8|
9.3

Test instance is closest to instance 5, which has a class value of 1, so 1 is predicted.

Essay Questions

2019, exam 2
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)
    Assumes that all attributes are independent and contribute equally. Fast, highly scalable supervised learning.

    Decision trees
    Recursive method that finds the attribute that best predicts, and then recursively repeats this process. Supervised learning, easy for humans to understand and easy to overfit. Can use different methods to determine the attribute that best predict. Builds one tree to predict class values.

    Classification Rules
    Supervised learning technique, looking for "nuggets" of information expressed as rules (if… then…). Uses a few attributes that predict a class given a desired coverage and accuracy. Supervised learning, easy for humans to understand (can be easier than a large tree).

    Association Rules
    Just looking for relations amongst the attributes, unsupervised. Still "nuggets" of information expressed as if… then… Also called a "market basket approach".

    Linear regression
    Use for numeric attributes, finding a formula (weights) to be predict. Can be predicting a numeric value (regression), or a Boolean value (logistic regression)

    Instance based learning
    Predicting a class based on the class of the instance(s) most similar. Every attribute, contributes equally. Supervised learning but no structure is learned.

    Clustering
    Grouping similar instances together, unsupervised.

    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.)

Stream learning – Learning involves a continual flow of massive amounts of data so it requires cost-effective techniques easily adapted to new instances. Linear regression, where the information is numeric, as numerical methods work faster. Instance-based learning is easily updated with new instances, statistical methods are fast.

Natural language processing – linear modeling (support vector machines) and Bayesian classification. Bayesian classifiers can treat each word as independent and does not need to consider the relations between words.

Psycholinguistic data mining – A subfield of natural language processing. Linear modeling and Bayesian classifiers may be used.

Content web mining – A subfield of natural language processing. Linear modeling and Bayesian classifiers may be used.

Usage web mining – Determining how users are using a site, clicking on advertisements, how long spent on a page, return rate, can use each of the methods. Clustering usage first could reduce the number of sites that need to be analyzed.

Structural web mining – Page rank algorithms are numerical methods.

Image mining – Linear methods, as images are easiest to work with when represented numerically. Decision trees to classify features. Clustering and association rules to group pixels into features.

Recommender systems – Suggest “relevant” items to users based on customer ratings (either directly or by purchasing) using collaborative filtering or content based systems. Memory-based collaborative filtering employ distance-measurement approaches, like nearest neighbor. Content-based systems use more classical learning techniques. Of the techniques covered in class, instance-based learning, clustering, or association rules.

Mining in adversarial situations – Initially a simple classifier, for example spam versus not spam, could catch adverse data. However, the adversarial situation

contains data carefully crafted to evade the filtering process. Thus, Bayesian statistical approaches were successful, but then weren't. Many approaches are used.

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
    - Rank 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.)

    Comments from students:
    Many people don't have a full grasp of what a data scientist does or can do, which means that data scientists need to be the referees to their own game.

    Governments are behind in laws concerning data, and thus it is necessary for data scientists to create a code of ethics.

    Data scientists are obligated to be constantly aware of the possible repercussions of their algorithms. They need to pay attention, not just to perfecting the algorithm itself, but also to its place in the larger picture.

    The algorithms we design assume an ideal system, when in reality, the real world to which the algorithm applies is far from perfect. It is our responsibility as purveyors of knowledge through these algorithms to be aware of these imperfections (such as systemic racism or gender bias) to the best of our ability, and to try to adjust the algorithms to compensate for them, if we can't correct the systems themselves.

    We must avoid creating, designing and implementing algorithms that cause harm, that is, algorithms that are opaque, affect large numbers of people and are damaging.

Identity is the first moral pillar that data scientists should actively uphold. Transparency is another and privacy another.

We could post the Data Science Code of Professional Conduct on the wall in the museum lab.

Suggestions:

- Make data scientists a licensed profession
- Post Data Science Association Code of Professional Conduct in the museum lab
- Require training on what makes algorithms dangerous and how to avoid them as part of the bachelor's degree
- Move towards creating better laws for who is responsible for the harm created by dangerous algorithms
- Create a mix of government intervention and industry self-regulation