CS548 homework

Cate Dunham

January 29, 2025

1 Chapter 1:

1.1 Problem 1

Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender. Not data mining. Assuming that gender is a given feature in the dataset, this is not data mining. If a model is predicting gender based off other features and then dividing the data, then this would be data mining.

(b) Dividing the customers of a company according to their profitability. Not data mining. Similar to 1a, this is not data mining unless profitability is also being predicted.

(c) Computing the total sales of a company. Not data mining. Again, this would be data mining if we were predicting sales but computing them could be done in Excel.

(d) Sorting a student database based on student identification numbers. Not data mining. This is just sorting data.

(e) Predicting the outcomes of tossing a (fair) pair of dice. Not data mining. We can just use probability; we don't need to learn anything about the behavior of the dice.

(f) Predicting the future stock price of a company using historical records. Data mining. This is a predictive task where the target variable is stock price. Unlike the dice above, we will need to find patterns in the data rather than using probability.

(g) Monitoring the heart rate of a patient for abnormalities. Data mining. This is a descriptive task (although we could have a predictive task if trying to predict when abnormalities would occur).

(h) Monitoring seismic waves for earthquake activities. Data mining. This is the same type of task as 1g.

(i) Extracting the frequencies of a sound wave. Not data mining. This is data transformation.

1.2 Problem 2

Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied. Clustering and/or classification can be applied to users to provide the results that will be most relevant to the person performing the search. Imagine a user searches for "good sing along songs". If that user has been classified as a, "parent of young children", they might want to see results for kids music. But if they are classified as a, "millennial graduate student" they might want to see results for 90s throwbacks. Many search engines already seem to do this, but what they often do not capture is that, depending on the situation, users may belong to very different classes. As a millennial graduate student who is also a mom of young children, I would love a search engine that could use available data to accurately predict which class I am currently searching as! For example, a browser that can use the time and date to predict that I probably don't want kids songs if I'm executing this search at 11pm on New Years Eve.

1.3 Problem 3

For each of the following data sets, explain whether or not data privacy is an important issue.

(a) Census data collected from 1900–1950. Not an issue. Census data is publicly available.

(b) IP addresses and visit times of Web users who visit your Website. Important issue. IP addresses can act as identifying information and users may not want outside parties to have access to their website visit history.

(c) Images from Earth-orbiting satellites. I don't think this would be a data privacy issue, but I think it really depends on how detailed the resulting images would be. If, for example, the license plate numbers could be seen in the satellite's data then I could imagine data privacy would become an issue.

(d) Names and addresses of people from the telephone book. Not an issue. This information is publicly available.

(e) Names and email addresses collected from the Web. Not an issue. Assuming that the source for the names and addresses is publicly available.

2 Chapter 2:

2.1 Problem 3

You are approached by the marketing director of a local company, who believes that he has devised a foolproof way to measure customer satisfaction. He explains his scheme as follows: "It's so simple that I can't believe that no one has thought of it before. I just keep track of the number of customer complaints for each product. I read in a data mining book that counts are ratio attributes, and so, my measure of product satisfaction must be a ratio attribute. But when I rated the products based on my new customer satisfaction measure and showed them to my boss, he told me that I had overlooked the obvious, and that my measure was worthless. I think that he was just mad because our best-selling product had the worst satisfaction since it had the most complaints. Could you help me set him straight?"

(a) Who is right, the marketing director or his boss? If you answered, his boss, what would you do to fix the measure of satisfaction? His boss is correct. To fix the measure of satisfaction we would need to account for the popularity of the product by dividing the number of complaints by the total number of purchases of that product.

(b) What can you say about the attribute type of the original product satisfaction attribute? The marketing director was using the original measure as an ordinal attribute, although we saw the issues with sorting the products by number of complaints.

2.2 Problem 4

A few months later, you are again approached by the same marketing director as in Exercise 3. This time, he has devised a better approach to measure the extent to which a customer prefers one product over other, similar products. He explains, "When we develop new products, we typically create several variations and evaluate which one customers prefer. Our standard procedure is to give our test subjects all of the product variations at one time and then ask them to rank the product variations in order of preference. However, our test subjects are very indecisive, especially when there are more than two products. As a result, testing takes forever. I suggested that we perform the comparisons in pairs and then use these comparisons to get the rankings. Thus, if we have three product variations, we have the customers compare variations 1 and 2, then 2 and 3, and finally 3 and 1. Our testing time with my new procedure is a third of what it was for the old procedure, but the employees conducting the tests complain that they cannot come up with a consistent ranking from the results. And my boss wants the latest product evaluations, yesterday. I should also mention that he was the person who came up with the old product evaluation approach. Can you help me?"

(a) Is the marketing director in trouble? Will his approach work for generating an ordinal ranking of the product variations in terms of customer preference? Explain. The marketing director is in trouble. He has probably made the assumption that preference would carry across the pairs, meaning that if a customer prefers 2 to 1 and 3 to 2 that they should also prefer 3 to 1. This may not always be the case, and would explain the difficulties coming up with consistent rankings from the results.

(b) Is there a way to fix the marketing director's approach? More generally, what can you say about trying to create an ordinal measurement scale based on pairwise comparisons? If the marketing director is set on using this approach, he could remove the product 3 vs product 1 comparison to eliminate inconsistencies caused by that third comparison. That would not be my recommended strategy (I'd suggest going back to the old ranking system until a better solution is found) but it would at least help. More generally, the pairwise comparison approach could realistically only work for a small number of items. Since each item needs to be compared to each other item. With just 6 items you would have 15 pairwise comparisons.

(c) For the original product evaluation scheme, the overall rankings of each product variation are found by computing its average over all test subjects. Comment on whether you think that this is a reasonable approach. What other approaches might you take? I think this depends a lot on the sample size. With a sufficiently large sample the average ranking might be fine. With a small sample size the average would be more susceptible to outliers, making it a less ideal choice. In this scenario the median or mode ranking may be more valuable.

2.3 Problem 13

Consider the problem of finding the K nearest neighbors of a data object. A programmer designs Algorithm 2.2 for this task.

Algorithm 2.2 Algorithm for finding K nearest neighbors.

```
1: for i = 1 to number of data objects do
2: Find the distances of the $i^{th}$ object to all other objects.
3: Sort these distances in decreasing order.
(Keep track of which object is associated with each distance.)
4: return the objects associated with the first K distances of the sorted list
5: end for
```

(a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same. I'm confused why step 3 in the algorithm says to sort in decreasing order. If we sort in decreasing order then the first K distances of the sorted list would actually be the furthest away? Assuming this is a mistake, duplicate objects will cause a problem because, with a distance of 0, they will always be each other's nearest neighbors. Since we already know that duplicate objects are similar, this result would not be useful.

(b) How would you fix this problem? A simple solution would be to remove duplicate objects.

2.4 Problem 14

The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of similarity measure from subsection 2.4 would you use to compare or group these elephants? Justify your answer and explain any special circumstances. The first thing I observe about the attributes is that they are on different scales. To correct for this we should standardize the data (to mean of 0 and std of 1) so that large attributes like height and small attributes like tusk length have an equal impact on the distance measurement. Since the attributes are all numerical and continuous, Euclidean distance would be an appropriate choice.

2.5 Problem 15

You are given a set of m objects that is divided into K groups, where the ith group is of size mi. If the goal is to obtain a sample of size n ¡ m, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select $n \times m_i/m$ elements from each group.

(b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

Approach a will result in a sample that has the same number of elements $n \times m_i/m$ for each group whereas with approach b this is not guaranteed.

2.6 Problem 18

This exercise compares and contrasts some similarity and distance measures.

(a) For binary data, the L1 distance corresponds to the Hamming distance; that is, the number of bits that are different between two binary vectors. The Jaccard similarity is a measure of the similarity between two binary vectors. Compute the Hamming distance and the Jaccard similarity between the following two binary vectors:

x = 0101010001
y = 0100011000

Hamming distance – number of bits that are different = 3

Jaccard similarity –
f00 = 5 the number of attributes where x is 0 and y is 0
f01 = 1 the number of attributes where x is 0 and y is 1
f10 = 2 the number of attributes where x is 1 and y is 0
f11 = 2 the number of attributes where x is 1 and y is 1

We can calculate Jaccard Similarity as:

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{2}{1+2+2} = .4$$

(b) Which approach, Jaccard or Hamming distance, is more similar to the Simple Matching Coefficient, and which approach is more similar to the cosine measure? Explain. (Note: The Hamming measure is a distance, while the other three measures are similarities, but don't let this confuse you.)

$$SMC = \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}} = \frac{2+5}{1+2+2+5} = .7$$

Hamming is more similar to SMC. Jaccard does not take into account matches where both bits are 0 but both SMC and Hamming do.

Similarly, both Jaccard and the cosine measure ignore 0-0 matches so Jaccard is more similar to the cosine measure than Hamming is.

(c) Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, Hamming or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain. (Assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise.)

I would use Jaccard for this since we care about how many genes they *share*, not how many genes are different.

(d) If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the Hamming distance, the Jaccard coefficient, or a different measure of similarity or distance? Explain. (Note that two human beings share >99.9% of the same genes.)

I'm not sure if this question is like the last one where the organisms are represented as binary vectors? If it is then Hamming could be appropriate since it would focus on the small genetic differences between the same-species organisms.

If the organisms are represented as continuous data something like Euclidean distance would be appropriate.

2.7 Problem 20

Here, we further explore the cosine and correlation measures.

(a) What is the range of values that are possible for the cosine measure?

Cosine ranges from -1 to 1.

(b) If two objects have a cosine measure of 1, are they identical? Explain.

A cosine measure of 1 it means that the vectors representing two two objects have the same *angle*, but does not tell us anything about their *magnitude*.

(c) What is the relationship of the cosine measure to correlation, if any? (Hint: Look at statistical measures such as mean and standard deviation in cases where cosine and correlation are the same and different.)

We know that a cosine similarity of 0 indicates that the angle between x and y is 90°, which means the vectors are orthogonal and are not correlated. Likewise, a cosine similarity of 1 indicates that the angle between x and y is 0°, which means the vectors are collinear.

(d) Figure 2.20(a) shows the relationship of the cosine measure to Euclidean distance for 100,000 randomly generated points that have been normalized to have an L2 length of 1. What general observation can you make about the relationship between Euclidean distance and cosine similarity when vectors have an L2 norm of 1?

When the vectors have an L2 norm of 1 we can see an inverse relationship between Euclidean distance and cosine similarity - as cosine similarity increases Euclidean distance decreases. This is what we would expect given that cosine similarity measures similarity and Euclidean distance measures dissimilarity.

(e) Figure 2.20(b) shows the relationship of correlation to Euclidean distance for 100,000 randomly generated points that have been standardized to have a mean of 0 and a standard deviation of 1. What general observation can you make about the relationship between Euclidean distance and correlation when the vectors have been standardized to have a mean of 0 and a standard deviation of 1?

Similar to the last question, we see an inverse relationship here. We know that highly correlated vectors are also likely to be close together, and this is what we see on the plot.

(f) Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an L2 length of 1.

Cosine for L2 length of 1 is:

$$\mathbf{x} \cdot \mathbf{y},$$

and Euclidean distance is:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{y}_k)^2} = \sqrt{\sum_{k=1}^{n} \mathbf{x}_k^2 - 2\mathbf{x}_k\mathbf{y}_k + \mathbf{y}_k^2}$$

We can rearrange, use norms and simplify to arrive at:

$$\sqrt{1 - 2\sum_{k=1}^{n} \mathbf{x}_k\mathbf{y}_k + 1} = \sqrt{2 - 2cos(\mathbf{x}, \mathbf{y})} = \sqrt{2(1 - cos(\mathbf{x}, \mathbf{y})}$$

(g) Derive the mathematical relationship between correlation and Euclidean distance when each data point has been standardized by subtracting its mean and dividing by its standard deviation.

With a mean of 0 and standard deviation of 1, correlation between $\mathbf{x}$ and $\mathbf{y}$ becomes:

$$corr(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{n - 1}.$$

With a mean of 0 and variance of 1 per data point we know that:

$$\sum_{k=1}^{n} \mathbf{x}_k^2 = n - 1,$$

$$\sum_{k=1}^{n} \mathbf{y}_k^2 = n - 1.$$

Substituting those into the expanded Euclidean distance:

$$\sqrt{(n-1) - 2(n-1)corr(\mathbf{x}, \mathbf{y}) + (n-1)} = \sqrt{2(n-1)(1 - corr(\mathbf{x}, \mathbf{y})}$$