Homework 1

Automated Learning and Data Analysis
Team 41

Rajshree Jain, rjain27
Srujana Rachakonda , srachak

January 31, 2020

1 Questions

Example of a numbered list:

1. Data Properties

   (a) Classify the following attributes as nominal, ordinal, interval or ratio. Also classify them as binary1, discreet or continuous. If necessary, give a few examples of values that might appear for this attribute to justify your answer. If you make any assumptions in your answer, you must state them explicitly.

   i. Diastolic blood pressure measured in units of millimeters of mercury - This is a Ratio Type and Continuous attribute. This is because, we can do all the operations on the values of this attribute. They have Distinctness, Order. We can also do Addition and Multiplication. Further, the value is continuous because it has real numbers as the values. Eg - 120.8, 130.5 etc.

   ii. Apartment number (101, 203, 411, etc.) - This is an Nominal and Discrete type attribute. This is because we can know that the values can be distinct but we cannot order them or perform addition and multiplication on them. Further, the value is discrete because it is a countably infinite set of values

   iii. Species of birds (sparrows, warblers, ducks, etc.) - This is a Nominal and Discrete attribute. This is because we can determine if the values of the attributes are distinct or not. Eg - sparrow, flamingo, peacock etc. We cannot order them or perform addition or multiplication on them. Also, the values are discrete because they are countably infinite set of values.

   iv. A record of whether or not a CSC student has attended a required seminar (Yes or No) - This is a Nominal attribute with Discrete (Binary) values. We can see that the values can either be yes or no. We can just say that they are distinct. We cannot order, add or multiply them. Further, they are discrete (binary) because they can have only 2 values.

   v. Temperature in Kelvin - This is a Ratio attribute with Continuous values. This is because the values are distinct. We can order them, add and multiply them. The kelvin temperature scale has an absolute 0 as well. Also the values can take real numbers as values which is why we can say that it is continuous.

   vi. Number of marbles in a bag - This is a Ratio attribute with Discrete values. This is ratio because we can detect if the values are distinct, we can order them, add the number of marbles in 2 bags and say that the number of marbles in one bag are twice the number of marbles in the other bag. Further, the values are discrete because they are countably infinite.

   vii. Income ($) - This is a Ratio Attribute with continuous values. We can perform all 4 operations - distinct, order, addition and multiplication on money. Further, the values will have real numbers so its continuous.

   viii. Movie seat number (A1, A2, B1, etc.) - This is a Nominal attribute with Discrete values. This has a similar explanation as Apartment number.

ix. Day of the month - This is an Interval Attribute with Discrete values. This is because the values can be determined as distinct or not, they can be ordered, we can say that a particular date is 5 days ahead of the other, but we cannot multiply them. Further, they have a countably infinite set of values so they are discrete.

x. Project group number (G01, G02, G03, etc.) - This is a Nominal attribute with Discrete values. This will have a similar explanation as the Apartment numbers.

(b) Table 1 is a dataset with 6 attributes describing students. For each of the following statistics/operations, list all of the dataset's attributes where we can apply that operation: mode, median, Pearson correlation, mean, standard deviation, z-score normalization, binary discretization (into a "high" and "low" group). If you make any assumptions in your answer, you must state them explicitly.

Table 1: Students Dataset

Course	**StudentID**	**GroupID**	**# of Teammates**	**Grade**	**Letter**
STAT 501	001	G11	3	92.1	A-
STAT 505	002	G13	3	89.2	B+
STAT 511	005	G02	2	93.6	A
CS 516	007	S03	2	95.0	A
CS 522	202	S03	3	85.3	B
CS 589	203	G02	2	82.4	B-
PSY 501	003	G06	3	78.2	C+
PSY 505	003	S02	3	86.7	B
PSY 516	391	S07	3	93.1	A
PSY 530	226	G08	2	92.6	A

We can see that the attributes in the Table are as follows:

- Course - Nominal - mode
- StudentID - Nominal - mode
- GroupID - Nominal - mode
- # of Teammates - Interval - mode, median, Pearson correlation, mean, standard deviation, z-score normalization (This we have considered as Interval attribute type because there is no definite)
- Grade - Ratio - mode, median, Pearson correlation, mean, standard deviation, z-score normalization
- Letter - Ordinal - Mode, Median, Binary Discretization

(c) Longitude is a measure of how far East/West your on the Globe, ranging from -180 to 180, with 0 going through Greenwich, England. Give an example of a situation where it would make sense to treat Longitude as an Interval attribute. Then given an example of when it would make sense to consider it as a Ratio attribute. Briefly justify each answer.

It depends on how we are trying to use the longitude attribute. A scenario, when the longitude attribute can be used as an Interval attribute is when we are trying to tell the definite position of a point. For example we say that the place is at -10 degree or -20 degree or -30 degree. We can notice that the difference in -10 degree and -20 degree is same as the difference between -20 degree and -30 degree. But in this case there is no absolute 0 value. Hence we can say that in this case the attribute is Interval Type. Further, if we want to know how far is a point from the point 0 degree or we want to tell how much we have travelled, then we will have the values as 10 degree, 20 degree etc. Like we can say that we travelled 10 degrees. In this case absolute 0 value is well defined because there is a possibility that we do not move so in that case the value will be 0. So, in this case we can say that the attribute is Ratio type.

2. Data Transformation and Data Quality
In a blood test, 3 measures (A1, A2, A3) results were collected for 12 patients. Table 2 shows the measures recorded for each patient after test. NA is used to indicate missing data.

(a) Evaluate the following strategies for dealing with missing data (NA) from the medical experiment above. Give an advantage and disadvantage of each strategy, and which you would choose. Briefly justify your answers in terms of the data above.

i. Strategy 1: Remove the patients with any missing values.

- Advantage - If the data-set is large or the number of missing values is less this technique actually gives a legit insight into the analysis. This is because removing the data objects with missing values, would not affect the data-set much.
- Disadvantage - In case the data set is not MCAR (missing completely at random), there is a possibility that we can introduce bias when we remove or delete objects from the data-set. Also, another much obvious disadvantage that if the dataset is small after deletion the data might make less sense or become much less powerful.

ii. Strategy 2: Estimate the value of missing data for an attribute by taking the average value

- Advantage - This allows us to use data in an incomplete data-set fully. If the data-set has a missing of type MCAR (missing completely at random). The sample mean of the data will not be biased. Hence in this case mean substitution might be a valid approach.
- Disadvantage - If the data that you have has got MAR (missing at random) or MCAR (missing completely at random) in that case the mean that will be calculated will be biased. Let's say, for example - We are taking in the values of salaries of people. There is a possibility that people who have a high income might not respond. Hence in that case mean substitution might give us a value of mean that is much less than the actual mean.

(b) Identify a possible outlier in the dataset and justify why it should be considered an outlier. Under what circumstances would it make sense to not consider it an outlier.
A possible outlier in the given data-set where the value of A2 is 11. That is Patient 2 (Patient-2, A1-229, A2-11, A3-44). It is considered as an outlier because, all the other values for the Attribute A2 are either 6, 7 or missing. This is the only patient that has such a high value of A2. Since the value is unusual with respect to other values, it is referred as an Outlier.
It makes sense to not consider the value as an outlier because there are many missing values in the dataset which might have a value of 11, or 12 or higher. We cannot comment on the value of these particular objects. Since we do not know anything about the missing values, we cannot actually say if value of 11 for A2 is an outlier or not.

3. Sampling

(a) State the sampling method used in the following scenarios and give a reason for your answer. Choose from the following options: simple random sample with replacement, simple random sample without replacement, stratified sampling, progressive/adaptive sampling.

i. Data is collected in an experiment until a predictive model reaches 90% accuracy.
-Progressive Sampling
- This approach starts with a small sample, and then increases the sample size until a sample of sufficient size has been obtained. While this technique eliminates the need to determine the correct sample size initially, it requires that there should be a way to evaluate the sample to judge if it is large enough. Further, here in this case we are checking if the accuracy of prediction reaches 90 percent.

ii. To learn the average GPAs of students at NC State University, the population was divided into the following groups: Freshman, Sophomore, Junior, and Senior. 5% of students from each group were selected for the study.
- Stratified Sampling
- Since, equal numbers of objects are drawn from each group even though the groups are of different sizes, it is called Stratified Sampling.

iii. From the following population, 1, 1, 2, 2, 5, a sample 1, 2, 2, 2, 5 was collected.
- Simple random sampling with replacement
- This is Simple random sampling with replacement because, objects are not removed from the population as they are selected for the sample so there is a good chance that they are re picked. We can see in the above samples, that the value 2 comes 3 times which means that the sample has simple random sampling with replacement.

(b) The U.S. Congress is made up of 2 chambers: 1) a Senate of 100 members, with 2 members from each state, and 2) a House of Representatives of 435 members, with members from each state proportional to that state's population. For example, Alaska has 2 Senators and 1 House representative, while Florida has 2 Senators and 27 House representatives. Both the Senate and the House are conducting surveys of their constituents, which they want to reflect the makeup of each chamber. You suggest that they use stratified sampling for this survey, sending surveys to a certain number of people from each state. Each survey will be sent to 1200 participants.

i. Why is stratified sampling appropriate here?
- Stratified Sampling is appropriate here because, when the data population consists of different types of objects, with widely different numbers of objects, simple random sampling can fail to adequately represent those types of objects that are less frequent. This can cause problems when the analysis requires proper representation of all object types. Similarly, in the above case as well there are various types of populations each from different - different states.

ii. For the Senate survey, how many surveys would you recommend sending to people in Alaska?
- There are 100 members in the Senate and 2 are from each state. So there will be 50 states. So the number of people in Alaska to whom Senate Survey should be sent are 1200/50 = 24 (Ans)

iii. For the House survey, how many surveys would you recommend sending to people in Florida?
For the house Survey the number of Surveys that will be send to the people in Florida are as follows.
There are total of 435 people in the House of Representatives.
There are 27 House representatives in Florida
Hence number of surveys that should be send to House of representatives are 27/435 * 1200 = 74.48 (Ans)

iv. What are some advantages of the "Senate" approach and the "House" approach to stratified sampling?
The senate approach will evenly pick people evenly from all the states. Further in the case of House of Representatives, the survey distribution will be in proportions to the state population of each state. When we do stratified sampling with proportion (House approach), the data we get is highly representative of the actual population. So the results of any kind of analysis will be very accurate. When we take senate approach the results will be very un-baised and not depend on the sizes of samples from each category.

4. Dimensionality Reduction
In this problem, you will analyze the PCA results on the BeijingPM2.5 dataset. Figure 1 shows the Eigenvalue Scree plot and the principal components of PCA analysis on the scaled raw dataset. The dataset was then normalized using z-scores, and Figure 2 shows the Eigenvalue Scree plot and the principal components of PCA analysis on dataset after normalization.

(a) In Figure 1, what is the most reasonable number of principal components to retain? Briefly justify your choice.

We would pick the first two principal components since we would always want to extract the components on the steep slope. Since the components on the shallow slope contribute little to the solution. The highest eigenvalues correspond to the maximum variance.

(b) Based on the table in Figure 1, do you think that performing PCA was useful? Why or why not? If not, what properties of the dataset caused PCA to be less useful?

No it was not useful since we did not perform normalization on the data. Each feature will have a different range of values, a higher range would just be considered as higher variance since PCA maximizes variance.

(c) In Figure 2, what is the most reasonable number of principal components to retain for dimensionality reduction? Briefly justify your choice. Hint: There may be more than one reasonable answer.

I would pick the first two principal components since as a rule of thumb we usually pick features with eigenvalues greater than 1 on normalized data.

(d) If you were to use the results in Figure 2 for feature selection, which of the original attributes would you select? Briefly justify your answer.

According to the tabular data, TEMP and PRES show highest covariance for PC1 and pm2.5 shows highest covariance for PC2 so we would pick these 3 attributes.

(e) Explain the difference between PCA1 and PCA2. Which one would you use for analysis and why?

The attributes we would pick depending on the PCs would vary in PCA1 and PCA2, since the data is not normalized in PCA1. We would use PCA2 since it uses z-score normalization and we eliminate issues with respect to varied ranges of feature values.

5. Discretization

(a) Discretize the attribute CHLORIDE by binning it into 5 equal-width intervals (the range of each interval should be the same). Show your work.
When we sort the values of Chloride attribute we get the following - 90, 91, 95, 97, 97, 97, 98, 99, 102, 103, 104, 105, 108, 109, 111.
We have to make 5 equal width intervals, so the width of each interval will be = 111-90/15 = 21/15 = 4.2.
Hence, the intervals and the numbers that lie in those intervals will be as follows :

Ranges	values
[90 , 94.2)	90,91
[94.2 , 98.4)	95, 97, 97, 97, 98
98.4 - 102.6	99, 102
[102.6 , 106.8)	103, 104, 105
[106.8 , 111)	108, 109, 111

(b) Discretize the attribute POTASSIUM by binning it into 5 equal-depth intervals (the number of items in each interval should be the same). Show your work.
When we sort the values of potassium we get the following : 2.7, 3.2, 3.5, 3.7, 3.8, 3.8, 3.9, 4.1, 4.1, 4.1, 4.6, 4.7, 5, 5.6 , 6, 6.5
When we divide these numbers such that there are 5 equal frequency bins we get the following
$[2.7, 3.6) - 2.7, 3.2, 3.5$
$[3.6, 3.9) - 3.7, 3.8, 3.8$
$[3.9, 4.2) - 3.9, 4.1, 4.1$
$[4.2, 5.1) - 4.6, 4.7, 5$
$[5.1, 6.6) - 5.6, 6, 6.5$

(c) Consider the following new approach to discretizing a numeric attribute: Given the mean (x) and the standard deviation () of the attribute values, bin the attribute values into the following intervals: [x + (k 1), x + k), for all integer values k,i.e. k=...4,3,2,1,0,1,2... Assume that the mean of the attribute CHLORIDE above is x = 100 and that the standard deviation = 6. Discretize CHLORIDE using this new approach. Show your work.
k=-1 then 100-12, 100-6 = [88, 94) = 90, 91
k= 0 then . 100-6, 100 = [94,100) = 95, 97, 97, 97, 98, 99
k=1 then 100, 100+6 = [100, 106) = 102, 103, 104, 105
k=2 then 106, 106+6 = [106, 112) = 108, 109, 111

(d) For each of the above discretization approaches, explain its advantages and disadvantages and when you would want to use it.
- Equal Width binning

- Advantages - This approach of binning handles the un skewed well distributed data in a very good way. This approach is easy to apply and understand.
- Disadvantage - It does not work well, when the data is skewed and also this approach will dominate outliers in the presentation.
- Use - When the data is well distributed it is good to use equal width binning. Also, when there is repeated data in the dataset, you should use this approach.

- Equal depth binning

- Advantages - This approach of discretizing is helpful as it divides the range into N intervals, each containing approximately same number of samples. Hence, this approach is helpful in the scaling of data and helps in a good distribution.
- Disadvantage - This approach can become tricky in case of managing Categorical Data. This approach may not be a very good approach for repeating values.
- Use - You can use this approach, when the data is skewed also.

- New discretizing approach

- Advantages - It is just that the approach where equal width intervals are created. So, we can distribute the data well. Also, since it is based on the mean and standard deviation values, all the time the discretization and the intervals will be very specific to the data that is being used. So, it is beneficial because the method will adapt as and when the data changes.
- Disadvantage - This approach is similar to the equal width binning and you might not work well for skewed data.
- Use - It can be used where the data is in form of a bell curve or has a normal distribution. Because in that curve also data is distributed with mean in the center and then having various numbers with a difference of standard deviation.

6. Distance Metrics

(a) For each distance function, describe whether it has each property. If so, give a short explanation of why. If not, give a counter example, including two pairs of items, the distance between them, and how it violates the given property.

i. Euclidean Distance- d(p,q)= $\sqrt{\sum_{i=1}^{d}(p_i - q_i)^2}$

- Positive Definiteness - Yes - This property states that $d(p, q) \geq 0$ for all p and q and d(p, q) = 0 only if p = q. We can see that the values of d(p,q) will always be positive as it will always be a square root of some number. Further, the value of d(p,q) will only be zero when p=q. Let us consider an example where d=1. In that case $\sqrt{(p_1 - q_1)^2} = 0$, so $p_1 = q_1$.
- Symmetry - Yes - This property means that d(p, q) = d(q, p) for all p and q. This will always be true since we $(p_1 - q_1)^2$ square the difference of p and q. So it does not matter that we calculate d(p,q) or d(q,p). Both of them will be the same.
- Triangle inequality -Yes - This property states that $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q, and r. Since, they are all distance values which are sum of squares under a square-root. They will all be positive. Using the Cauchy-Schwarz inequality we can prove that euclidean distance satisfies the triangle inequality.

ii. Hamming distance - Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different.

- Positive Definiteness - Yes - Since, Hamming distance is a count of positions in binary bits which are different in two bianry strings. The count can never be less than zero. It will only be zero when all the bits in the binary strings are the same. That means that the strings are equal. This satisfies our positive definiteness definition.
- Symmetry - Yes - The count of bits different in two strings would be the same regardless of which string is considered first. Therefore, it is symmetric by default.
- Triangle inequality - Yes - So we want to prove that d(x,z) $\leq$ d(x,y) + d(y,z)
  For words of length 1, d(x,y) = 0 or 1 depending on whether x = y or not.
  Now for d(x,z) if x = z then d(x,z) = 0 else it is 1.
  If d(x,z) = 0 then LHS of the equation is 0 and LHS $\leq$ RHS (since RHS is either 1 or 0)
  if LHS = 1 that means x $\neq$ y and either y $\neq$ z or x $\neq$ z
  Now for words of length n, we do the above computation summed over n values and we arrive at d(x,z) $\leq$ d(x,y)
  Therefore, Hamming distance satisfies the triangle inequality property.

iii. Cosine distance between two numeric vectors, defined as one minus the cosine similarity

- Positive Definiteness - No - It does not follow, Positive definiteness because for positive definiteness d(p,q) greater than or equal to 0 for all p and q and d(p,q)=0 when p=q. But in case of cosine distance. If we take the distance between (2,2) and (3,3) the distance is 0 but the points are not equal to each other therefore it does not satisfy Positive definiteness.
- Symmetry - Yes - Since, cosine distance is 1 - cosine similarity and cosine similarity is dot product/ product of magnitudes of the two vectors. Interchanging them is not going to change the final value.
- Triangle inequality - No - Consider the example where a = [1,0] b = [1/sqrt(2), 1/sqrt(2)] c = [0,1] For these values, the inequality does not hold up and hence, it does not satisfy it.

(b) i. What property of distance metrics allows us to skip some d(xi,y) comparisons in the 1-NN algorithm?

We can use the triangle inequality property to skip some of the computations of d(xi, y) in the above mentioned 1-NN algorithm.

ii. What strategy could we use to reduce the number of d(xi,y) comparisons? Give one example with values for y, x1, and x2, that illustrates that strategy. (Hint: it may help to draw it out the positions of x1, x2 and y in a 2D space.)

-Consider 4 points on the graph. y, x_i , x_j , and x_p. We initially compute distances between y and x_i and x_j.
Using triangle inequality we can arrive at the equation:
$D_{ij} + D_{yj} \geq D_{yi}$
This can be modified into $D_{ij} \geq D_{yi} - D_{yj}$ —— 1
Then we have an arbitrary point x_p. For this arbitrary point x_p we compute the distances D_pj. I.e we compute the distance between x_p and one of our known points x_j. This is cheap since both x_p and x_j are x points!
Now, if $D_{pj} > D_{ij}$ then D_{yp} need not be computed as proven below:
Expanding the above equation using 1 we get,
$D_{yp} - D_{yj} > D_{yi} - D_{yj}$
It can be further simplified into
$Dyp > D_{yi}$
Therefore, the distance between the arbitrary point X_p and Y need not be computed since if $D_{pj} > D_{ij}$, the distance between X_p and Y will always be greater than the distance between X_i and Y.

iii. Does this strategy reduce the number of d(xi,y) comparisons in the best case? What about the worst case?

In the best case, the identified xi, xj points would be the two points closest to the y point and every other point identified would be farther than the two identified points and all those comparisons can be skipped.
However, in the worse case scenario, if the two identified points are the farthest points and each new point discovered in the X set is in descending order of distances then the $D_{pj} > D_{ij}$ condition will never be satisfied and we will be forced to compute all the distance values of X from Y. Therefore, in the worst case the number of comparisons will not be reduced.

7. L-infinity is the distance metric that proved to be most useful since the difference between their mean_intra_dist and mean_inter_dist is the largest of all the metrics.