Homework 1

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2019

Instructions

- **Due Date:** January 31, 2019 at 11:45 PM

- **Total Points**: 100
  + 5 bonus points if you submitted HW0 and followed all the instructions on HW1; 0 otherwise

- Make sure you clearly list each team member's names and Unity IDs at the top of your submission.

- Your submission should be a single zip file containing a PDF of your answers, your code, and a README file with running instructions. Please follow the naming convention for your zip file: G(homework group number)_HW(homework number), e.g. G1_HW1.

- If any question is unclear, please post a question on Piazza. If you make any assumptions in your answer, you must state them explicitly (e.g. "Assuming the data in question is normally distributed...").

- If a question asks you to explain or justify your answer, **give a brief explanation** using your own ideas, not a reference to the textbook or an online source.

- If this homework needs to be updated for any reason, **we will post the update on Piazza** and announce it in class, so please stay alert.

- In addition to your group submission, please also *individually* submit your Peer Evaluation form on Moodle, evaluating yours and your teammates' contributions to this homework.

Problems

1. Data Properties (18 points) [**Ruth Okoilu**]

   [Lecture 2] Answer the following questions about attribute types.

   (a) Classify the following attributes as *nominal, ordinal, interval* or *ratio*. Also classify them as *binary*[1], *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. Blood group (A, B, AB, O)
       ii. Ticket number for raffle draws
       iii. Brightness as measured by a light meter
       iv. Grade in terms of Pass or Fail
       v. Time zones (EST, PST, CST)
       vi. Income earned in a month
       vii. Vehicle license plate number
       viii. Distance from the center of campus.
       ix. Dorm room number

---

[1]binary attributes are a special case of discreet attributes

x. Kelvin temperature

(b) Table 1 is an automobile dataset with 6 attributes. For each attribute, list which of the following statistics/operations can be calculated on that attribute: 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: Automobile Dataset

Make	Fuel-type	# of doors	Height	# of Cylinders	Price
alfa-romero	gas	two	48.8	four	$1349
alfa-romero	diesel	four	50.52	six	$16500
alfa-romero	diesel	two	54.3	four	$16500
audi	gas	four	55.7	eight	$13950
audi	diesel	two	70.38	eight	$17450
audi	gas	four	71.74	six	$15250
bmw	diesel	two	55.1	four	$16430
bmw	gas	four	54.3	eight	$116925
bmw	diesel	two	53.33	six	$20970
bmw	diesel	four	53.3	six	$21105

(c) A quiz was conducted to test students' knowledge of data mining. We don't know what questions were on the quiz or how it was graded, but we do know students' scores are on a scale of 0 to 5. Give an example of a situation where it would make sense to treat this as an Ordinal attribute. Then given an example of when it would make sense to consider it as a Ratio attribute. *Briefly* justify each answer.

Solution:

(a)   i. Blood group (A, B, AB, O) Discrete, Nominal
     ii. Ticket number for raffle draws Discrete Nominal/Ordinal
     iii. Brightness as measured by a light meter Continuous Ratio
     iv. Grade in terms of Pass or Fail Binary Ordinal
     v. Time zones (EST, PST, CST) Discrete Nominal
     vi. Income earned in a month Continuous Ratio
     vii. Vehicle license plate number Discrete Nominal
     viii. Distance from the center of campus. Continuous, Interval/Ratio (depends on whether center of campus is regarded as an arbitrary point)
     ix. Dorm room number Discrete Ordinal
     x. Kelvin temperature Continuous Ratio

For questions with multiple options (e.g. Nominal/Ordinal), we accepted either if the justification was appropriate.

(b) Make (Nominal), Fuel Type (Nominal/Binary), No-of-doors (Nominal/Ordinal/Binary), Height (Ratio), No-of-Cylinders (Ordinal), Price (Ratio).

- The following operations can be performed on a nominal attribute: mode.
- The following operations can be performed on an ordinal attribute: median, binary discretization (assuming more than 2 different values), as well as the operations above for nominal attributes above.
- The following operations can be performed on an interval attribute: mean, standard deviation, Pearson's correlation, z-score normalization, as well as the operations above for ordinal attributes above.
- The all of the above operations can be performed on a ratio. attribute.

(c) If higher scores (5) are better than lower scores (3) then the scale is only ordinal. However, if the difference between, say, a 2 and a 3 represents the same difference in knowledge as the difference

between a 4 and a 5 then it is a ratio attribute. This may occur if all questions are practicing the same concept (e.g. all simple multiplication exercises). If all problems test different skills with different difficulties, that attribute may just be ordinal.

2. Data Transformation and Data Quality (12 points) [**Ruth Okoilu**]

   In a medical experiment, 8 patients were subjected to two different treatments. The following results are the systolic blood pressure (SBP) recorded for each patient after treatment, for each treatment, for patients 1-8, respectively. For example, Patient 1 had an SBP of 160 after treatment A and 300 after treatment B, and so on. `NA` is used to indicate missing data.

   Treatment A: 160, 120, 130, `NA`, 120, `NA`, 240, 140
   Treatment B: 300, 100, `NA`, 130, 110, 100, 120, 90

   (a) Transform this data into a tabular, record format. Use attributes ID, Patient, Treatment, and SBP for the columns and observations in rows.

   (b) 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.
       ii. **Strategy 2:** Estimate the value of missing data for an attribute by taking the average value of other participants for that attribute.

   (c) Consider the results of the medical experiments above.

       i. Are the results 240 and 300 noise or outliers? When analyzing the data, what are some strategies to deal with these values? *Briefly* justify both answers.
       ii. The instrument we used to measure SBP will give a reading within +/- 3 of the actual SBP (i.e. if the real SBP is 150, it will give a value between 147-153). Will this inaccuracy create noise or outliers? What are some strategies to deal with this inaccuracy? *Briefly* justify both answers.

   Solution:

   (a) See table 2

Table 2: Medial Experiment in Tabular Form

ID	Patient	Treatment	SBP
1	1	A	160
2	2	A	120
3	3	A	130
4	4	A	NA
5	5	A	120
6	6	A	NA
7	7	A	240
8	8	A	140
9	1	B	300
10	2	B	100
11	3	B	NA
12	4	B	130
13	5	B	110
14	6	B	100
15	7	B	120
16	8	B	90

   (b) Answers may vary. Strategies for handling data:
       Strategy 1: Remove the patients with any missing values.
       Strategy 2: Estimate the value of missing data for an attribute by taking the average value of

other participants for that attribute.
Advantages:
Strategy 1: It is easy to implement, and it faithfully represents the original data (no made-up values)
Strategy 2: You do not lose 3 of the 8 participants in your dataset.
Disadvantages:
Strategy 1: A disadvantage is that if there are many missing observations, the analysis will be deficient.
Strategy 2: Mean substitution method may lead to bias, as it includes fabricated data points. Also, no new information is being added, so the actual information in the sample is not increased.

(c) Consider the results of the medical experiments above.

    i. **Outliers**: Data points which deviate strongly from others in the dataset. They may or may not be correctly measured, and they should be excluded for some analyses that are sensitive to outliers (e.g. mean). This can be demonstrated using some established measure of deviance, such as SD or Variance.
Noise: Data points which have been measured incorrectly due to faulty instruments or data recording. *Random* noise randomly overlaps with true signal in the observed data, so it does not add bias.
For Treatment A, $M_A = 151.7; SD_A = 45.8$, so the z-score of 240 = 1.93.
For Treatment B, $M_B = 135.7; SD_B = 73.7$, so the z-score of 300 = 2.23.
Based on these calculations, both values are around 2 SD above average ($>$ 97th percentile). We can argue that they are therefore outliers.
300 can also be seen as noise if we know that the maximum value for a SBP is 250. Since 300 is above 250, the reading must have been due to faulty equipment. However, this domain-specific knowledge was not assumed for this question.
Note: This question was graded based on each student's assumption and justification provided. For example, if the student assumes maximum value for a SBP is 300 or provides a strategy for detecting outliers such as labelling values that fall beyond a specific number of standard deviations away from the mean as outliers.

    ii. This inaccuracy will create noise because from our definition of outlier and noise above, noise results from inaccuracies in measurements. Dimensionality reduction teachniques such as PCA can be used to reduce noise.

3. Sampling (7 points) [**Ruth Okoilu**]

(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. To determine the average salary of professors at NC State University, the faculty were divided into the following groups: instructors, assistant professors, associate professors, and professors. Twenty faculty members from each group were selected for the study.

    ii. From the following population, {2, 2, 4, 4, 6, 6, 8}, a sample {2,2,2,6,8} was collected.

    iii. Data is collected in an experiment until a predictive model reaches 90% accuracy.

(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 1000 participants.

    i. Why is stratified sampling appropriate here?

    ii. For the Senate survey, how many surveys would you recommend sending to people in Alaska?

    iii. For the House survey, how many surveys would you recommend sending to people in Florida?

    iv. What are some advantages of the "Senate" approach and the "House" approach to stratified sampling?

Solution:

(a)
   i. Stratified Sampling. Stratified sampling divides the data into subgroups based on some attribute (e.g. state). A simple random sample is then selected from each subgroup to get the final sample.

   ii. Simple random sampling with replacement. Here, a population element can be selected more than one time.

   iii. Adaptive or Progressive sampling. This approach starts with a small sample, and then increases the sample size until a sample of sufficient size has been obtained. Progressive sampling is used to learn a predictive model and the accuracy of predictive models increases as the sample size increases.

(b)
   i. A stratified sampling scheme accommodates differing frequencies for the items of interest. For example, since they want the sample to reflect the makeup of each chamber, then a different number of surveys will be sent to each state. Stratified sampling allows this.

   ii. 20 Surveys. Each state gets equal number of Senate surveys

   iii. Approximately 62 surveys i.e. $(27/435) * 1000$. We want the number of suveys sent to Florida to be proportional to its representation in the House, which is $27/425$. Since there are 1000 surveys, we multiply $(27/435) * 1000$.

   iv. The advantage of the "Senate" approach is that each state is equally-well represented, meaning any conclusions from the surveys will apply to people from all states. The advantage of the "House" approach is that the conclusions we gain from the surveys will apply well across the whole of the U.S. Both approaches have the advantage of ensuring that there is some representation from all states, no matter how surveys are mailed out.

4. Dimensionality Reduction (12 points) [**Song Ju**]

In this problem, you will analyze the PCA results on the *iris* dataset. Figure 1 shows the Eigenvalue Scree plot and the principal components of PCA analysis on the 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.
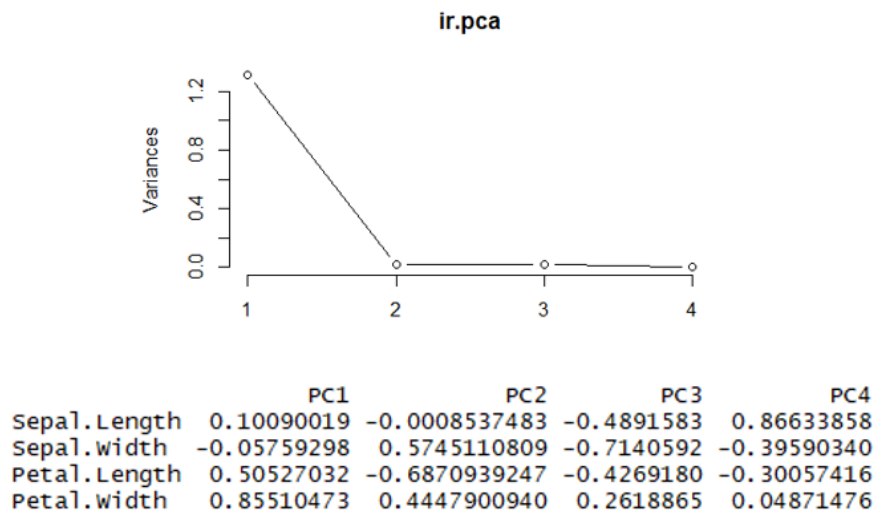

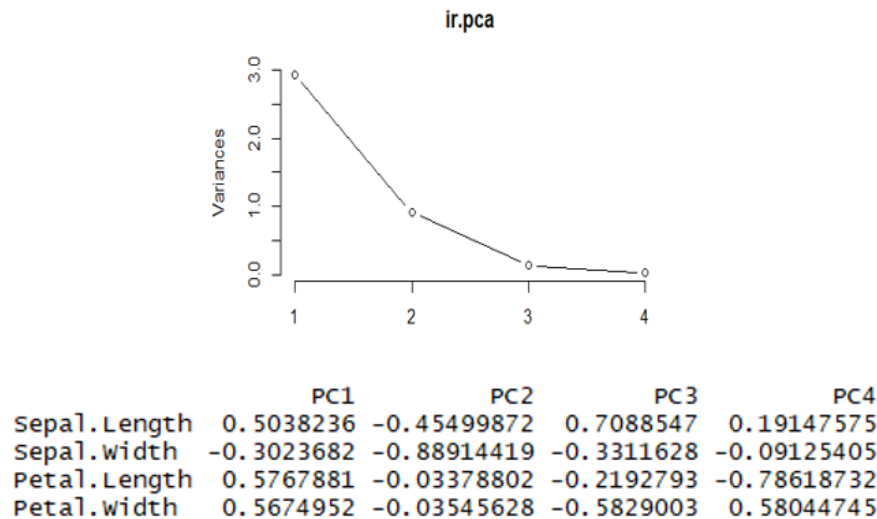|	PC1	PC2	PC3	PC4
Sepal.Length	0.10090019	-0.0008537483	-0.4891583	0.86633858
Sepal.Width	-0.05759298	0.5745110809	-0.7140592	-0.39590340
Petal.Length	0.50527032	-0.6870939247	-0.4269180	-0.30057416
Petal.Width	0.85510473	0.4447900940	0.2618865	0.04871476

Figure 1: PCA1 on Raw Dataset

Figure 2: PCA2 on Normalized Dataset

Please answer the following questions:

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

(b) In Figure 1, according to the first principal component, what are the feature(s) in the original dataset that explain the most variance?

(c) In Figure 2, what is the most reasonable number of principal components to retain? Briefly justify your choice.

(d) In Figure 2, according to the first principal component, what are the feature(s) in the original dataset that explain the most variance?

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

(f) Based on the results of PCA1 and PCA2, which feature(s) would you like to select if you need to do feature selection? Briefly justify your choice.

Solution:

(a) PC1 is the only reasonable principal components to retain, since it's the only one with eigenvalue larger than 1.

(b) `Petal.width`, since it has the largest portion of PC1's variance.

(c) We should retain PC1 and PC2, since they are in the steep slope of the curve and have eigenvalues larger than 1.

(d) `Petal.Length` or `Petal.width` (and arguably `Sepal.Length`) explain most of the variance, as they have the highest eigenvalues in PC1.

(e) The difference is normalization. After normalization, each attribute has a variance of 1. Without normalization, PCA will be biased to only represent the attributes with the largest variance. We would use PCA2 for analysis.

(f) First, we need to use PCA2 to do feature selection. Then, we can select `Petal.Length` and `Petal.Width` based on PC1 and `Sepal.Width` based on PC2. Given that we said both PC1 and PC2 should be retained, we should select features based on these principle components.

5. Discretization (12 points) [**Song Ju**]

Consider the following dataset:

No.	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	71.0	TRUE	yes
12	overcast	73.0	89.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no
15	sunny	95.0	85.0	FALSE	yes
16	rainy	50.0	45.0	YES	no

(a) Discretize the attribute TEMPERATURE by binning it into 4 equal-width intervals (the range of each interval should be the same). Show your work.

(b) Discretize the attribute HUMIDITY by binning it into 4 equal-depth intervals (the number of items in each interval should be the same). Show your work.

(c) Consider the following new approach to discretizing a numeric attribute: Given the mean ($\bar{x}$) and the standard deviation (σ) of the attribute values, bin the attribute values into the following intervals: $[\bar{x} + (k-1)\sigma,\ \bar{x} + k\sigma)$,
for all integer values k, i.e. $k = \ldots - 4, -3, -2, -1, 0, 1, 2 \ldots$
Assume that the mean of the attribute HUMIDITY above is $\bar{x} = 80$ and that the standard deviation $\sigma = 13$. Discretize HUMIDITY using this new approach. Show your work.

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

Solution: (a) [50, 61.25), [61.25, 72.5), [72.5, 83.75), [83.75, 95]

(b) [45, 71), [71, 81), [81, 90), [90,97)

(c) [41, 54), [54, 67), [67, 80), [80, 93), [93, 106)

(d) Equal-width is more sensitive to outliers than equal-depth, since an extreme value would throw off the bins. However, Equal-width is also simpler and more interpretable. The standard deviation approach in (c) would be useful if we wanted extreme values to be in separate bins, since few values will have +/- more than 2 SDs.

6. Distance Metrics (14 points) [**Song Ju**]

(a) A true distance metric has three properties: a) positive definiteness, b) symmetry, c) triangle inequality. Now consider the following distance functions:

   i. Euclidean distance between two numeric vectors

   ii. Manhattan distance between two numeric vectors (L_1)

   iii. A "divergence function" defined between two sets as $d(A, B) = 1 - |A \cap B|/|A|$, representing how many of A's items are not present in B.

   iv. Cosine distance between two numeric vectors, defined as 1 minus the cosine *similarity*:
$d(A, B) = 1 - A \cdot B/(||A||\ ||B||)$

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.

(b) A 1-nearest-neighbor (1-NN) classifier labels a new item y in the test dataset Y by finding the closest item x in the training dataset X, and returning the label of x.

Assume we have a distance function d that is *very expensive* to calculate for any $d(x, y)$ where $x \in X$ and $y \in Y$. However, because we can pre-calculate the distance between any two items in X, $d(x_i, x_j)$ is relatively *cheap* to calculate for any $x_i, x_j \in X$.

To classify a new item y, our 1-NN algorithm will have have to make $|X|$ comparisons between y and some x_i, since it has to compare y to every item $x_i \in X$ to find y's closest neighbor. However, if d is a true distance *metric*, we may be able to reduce the number of comparisons we have to make by *skipping* some of them.

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

   ii. What strategy could we use to reduce the number of $d(x_i, y)$ comparisons? Give one example with values for y, x_1, and x_2, that illustrates that strategy. (Hint: it may help to draw it out the positions of x_1, x_2 and y in a 2D space.)

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

Solution:

(a) For positive definiteness, we have two principles:

$d(x, y) \geq 0$ for any x and y (1)

$d(x, y) = 0$ if and only if x=y (2)

Euclidean distance is positive definite, symmetric and satisfies the triangle inequality.
Manhattan distance is positive definite, symmetric and satisfies the triangle inequality.
Since these two distances represent distances in physical space, they are bound by all three properties of a distance metric.

The divergence function is not positive definite, nor symmetric, and it does not satisfy the triangle inequality. For positive definiteness, it violates principle (2). Suppose $A = \{1, 2\}$ and $B = \{1, 2, 3, 4\}$, then $d(A, B) = 0$ but $A \neq B$. For symmetry, we have $d(A, B) = 0$ but $d(B, A) = 0.5$. For triangle inequality, suppose we have $A = \{1, 2\}, B = \{1, 2, 3, 4\}$ and $C = \{3, 4\}$, then $d(A, B) = 0, d(A, C) = 1, d(B, C) = 0.5$. So we have $d(A, C) > d(A, B) + d(B, C)$ and this violates the triangle inequality.

Cosine distance is not positive definite, it is symmetry and it does not satisfy the triangle equality. For positive definiteness, it violates principle (2). For example, if $A = [1, 1]$ and $B = [2, 2]$, $d(A, B) = 0$, but $A \neq B$. For the triangle inequality, with points $A = [1, 0], B = [2, 2], C = [0, 1]$ we have $d(A, B) = 0.29, d(A, C) = 1, d(B, C) = 0.29$. So we have $d(A, C) > d(A, B) + d(B, C)$ and this violates the triangle inequality.

(b)   i. Triangle inequality

   ii. First, give d_1 as the distance between y and x_1, and d_{12} as the distance between x_1 and x_2. Now we want to estimate the distance d_2 between y and x_2. Based on triangel inequality, we have $d_2 + d_1 \geq d_{12}$. Then we have the lower bound that $d_2 \geq d_{12} - d_1$. If d_{12} is larger than $2 * d_1$, then we know d_2 must be larger than d_1 and we can get rid of x_2.

   iii. The best case is we only need to calculate one distance and get rid of all the other points by triangle inequality. The worst case is we need to calculate the distance between y and every item x_i. See Figure 3.
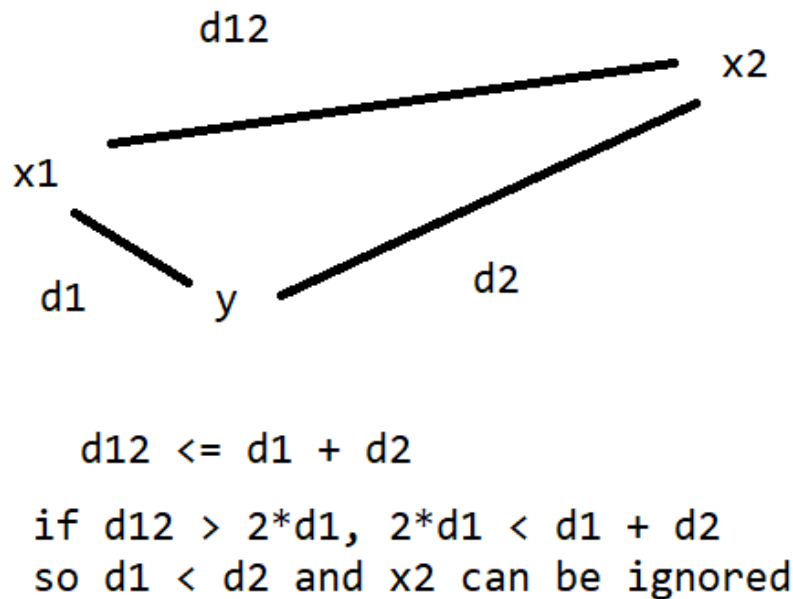
```
d12 <= d1 + d2

if d12 > 2*d1, 2*d1 < d1 + d2
so d1 < d2 and x2 can be ignored
```

Figure 3: 6.b.iii explanation

7. Similarity, Dissimilarity and Normalization (25 points) [**Krishna Gadiraju**]

R Programming Submission Instructions

- Make sure you clearly list each team member's names and Unity IDs at the top of your submission.
- Your code should be named $hw1.R$. Add this file, along with a README to the zip file mentioned in the first page.
- Failure to follow naming conventions or programming related instructions specified below may result in your submission not being graded.
- If the instructions are unclear, please post your questions on piazza.

Programming related instructions

- Carefully read what the function names have been requested by the instructor. **In this homework or the following ones, if your code does not follow the naming format requested by the instructor, you will not receive credit.**
- For each function, both the input and output formats are provided in the $hw1.R$. Function calls are specified in $hw1\_checker.R$. Please ensure that you follow the correct input and output formats. Once again, if you do not follow the format requested, you will not receive credit. It is clearly stated which functions need to be implemented by you in the comments in hw1.R
- You are free to write your own functions to handle sub-tasks, but the TA will only call the functions he has requested. If the requested functions do not run/return the correct values/do not finish running in specified time, you will not receive full credit.
- DO NOT set working directory (setwd function) or clear memory (rm(list=ls(all=T))) in your code. TA(s) will do so in their own auto grader.
- The TA will have an autograder which will first run source(hw1.R), then call each of the functions requested in the homework and compare with the correct solution.

- Your code should be clearly documented.

- To **test you code**, step through the *hw1_checker.R* file. If you update you code, make sure to run `source('./hw1.R')` again to update your function definitions. You can also check the "Source on save" option in R Studio to do this automatically on save.

- You can also check you functions manually by running them in the console with smaller vectors.

- Calculating the distance matrix may take **20-40 seconds**. If you want to test on a smaller dataset, you can use the *hw1_word_frequency_small.csv* file, which contains a small subset of the original data.

Questions

You are given the following dataset(s):

- Hotel reviews for a hotel in New York [1]. You are provided with the list of words in the dataset, as well as frequency count of each word in all the sentences. Each line in *hw1_word_frequency.csv* represents a 200 element vector (i.e., your vocabulary size, or total number of words in your dataset is 200) representing a single sentence from your dataset. Each value in the line represents the frequency count of that particular word over the entire document space. If the word does not exist in the sentence, you will have a zero. The corresponding mapping of the words to the array indices is provided in *hw1_word_index.json* for your reference (but it will not be used in this assignment). In total, there are 155 sentences.

Part 1: Without normalization Using the data provided in *hw1_word_frequency.csv*, you will generate a 155 x 155 element distance matrix containing the pairwise distances for each sentence, using each of the following distance functions. We have already given you the implementation for the function `calculate_matrix` which calls each of the distance methods specified below to compute the distance matrix for every pair of sentences. For euclidean, cosine and manhattan, you are to implement the formulae specified below, and the input type is two vectors of same length i.e., any two sentences from *hw1_word_frequency.csv*. For cheybshev, you will use one of the libraries specified below, and the input is the entire *hw1_word_frequency.csv* as a matrix. The formulae for each distance/similarity measure, as well as whether you need to implement/use library are clearly stated below: Use the `philentropy` R package when using a library for the distance functions, and use the package documentation to find the appropriate function.

(a) **euclidean** (implement): euclidean$(P, Q) = \sqrt{\sum_i (P_i - Q_i)^2}$, where P and Q are vectors of equal length.

(b) **cosine** (implement): cosine$(P, Q) = \frac{\sum_i P_i * Q_i}{||P|| * ||Q||}$, where P and Q are vectors of equal length, and $||P|| = \sqrt{\sum_i P_i^2}$.

(c) **manhattan** (implement): manhattan$(P, Q) = \sum_i |P_i - Q_i|$, where $|x|$ is the absolute value of x, and P and Q are vectors of equal length.

(d) **chebyshev** (use library): chebyshev$(P, Q) = max(|P_i - Q_i|)$, where P and Q are vectors of equal length. Please note that you will not use this formula directly for chebyshev, since you will be using the library.

Part 2: With normalization After calculating the aforementioned distances using the data from *hw1_word_frequency.csv*, you will normalize each row in the data matrix to [0, 1] range using the formula $\frac{row - min(row)}{max(row) - min(row)}$. *You are expected to implement this calculation, not use a library.* This has to be implemented in the function `normalize_data`. Then, recompute the euclidean distance using the method you implemented in the previous section. Do you notice any difference in the two distance matrices? In the function `analyze_normalization`, write some R code to compare the two outcomes. Identify at least 3 properties of the distance matrix that change when using normalized data. In the PDF, list the changes you identified and explain why they occurred. Use both numeric calculations and visualizations to justify your answers. (Hint: use the `hist` function, and the provided `plot_distance_matrix` function.)

*hw*1.*R* has already been provided for you, with the function definitions. Complete all the functions requested for in *hw*1.*R*. Please note that *hw*1_*checker*.*R* is only for you to understand how the TA will run your files. DO NOT submit *hw*1_*checker*.*R*. Also, please note that the TA may be using a dataset different to yours (but with the same dimensions, i.e., 155 sentences, each of length 200), so do not hard code your solutions.

Also, it is recommended you read up on vectorized operations in R. Any submission that takes more than 5 minutes to run on a standard university machine (32 GB RAM, i7 processor) will receive a zero grade. Also, please ensure that all the libraries are correctly loaded using the require method.

Allowed Packages: R Base, utils, plyr, philentropy, data.table, reshape and ggplot2. No other packages are allowed.

Solution: Code is provided in *hw*1_*solution*.*R* The analysis section was designed to make you realize that measures such as mean, euclidean distance are heavily influenced by the large values. Bringing the data into a standard range helps alleviate this problem with euclidean distance.

References

[1] K. Ganesan and C. Zhai, "Opinion-based entity ranking," *Information retrieval*, vol. 15, no. 2, pp. 116–150, 2012.