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 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.

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.

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?

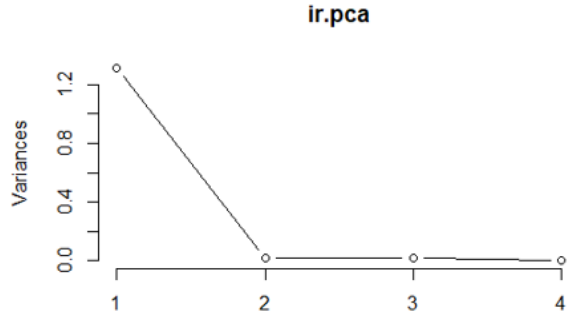
ii. For the Senate survey, how many surveys would you recommend sending to people in Alaska?

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?

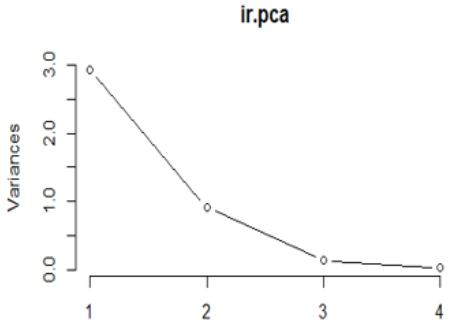
4. Dimensionality Reduction (12 points) [**Song Ju**]

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



	PC1	PC2	PC3	PC4
Sepal.Length	0.5038236	-0.45499872	0.7088547	0.19147575
Sepal.Width	-0.3023682	-0.88914419	-0.3311628	-0.09125405
Petal.Length	0.5767881	-0.03378802	-0.2192793	-0.78618732
Petal.Width	0.5674952	-0.03545628	-0.5829003	0.58044745

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.

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.

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?

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.)

$hw1.R$ has already been provided for you, with the function definitions. Complete all the functions requested for in $hw1.R$. Please note that $hw1\_checker.R$ is only for you to understand how the TA will run your files. DO NOT submit $hw1\_checker.R$. 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.

References

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