Homework 1

Automated Learning and Data Analysis
Dr. Thomas Price

Spring 2020

Instructions

Due Date: January 30, 2020 at 11:45 PM
Total Points: 100 for CSC522; 92 for CSC422. + 5 bonus points if you submitted HW0 and followed all the instructions on HW1
Submission checklist:

- 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. **Name your file**: G(homework group number)_HW(homework number), e.g. G1_HW1.
- 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.
- 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) [**Ge Gao**]

   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. Diastolic blood pressure measured in units of millimeters of mercury
   ii. Apartment number (101, 203, 411, etc.)
   iii. Species of birds (sparrows, warblers, ducks, etc.)
   iv. A record of whether or not a CSC student has attended a required seminar (Yes or No)
   v. Temperature in Kelvin
   vi. Number of marbles in a bag
   vii. Income ($)
   viii. Movie seat number (A1, A2, B1, etc.)
   ix. Day of the month
   x. Project group number (G01, G02, G03, etc.)

   (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.

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

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	96.2	A

(c) Longitude is a measure of how far East/West your are 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.

2. Data Transformation and Data Quality (12 points) [**Ge Gao**]

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.

Table 2: Medical Measures

Patient	**A1**	**A2**	**A3**
1	233	6	48
2	229	11	44
3	226	`NA`	43
4	243	`NA`	41
5	249	6	38
6	`NA`	6	`NA`
7	253	6	39
8	257	7	44
9	251	7	44
10	251	`NA`	43
11	249	7	41
12	253	6	`NA`

(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.

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

(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.

3. Sampling (7 points) [**Ge Gao**]

(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.

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.

iii. From the following population, {1, 1, 2, 2, 5}, a sample {1, 2, 2, 2, 5} was collected.

(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?

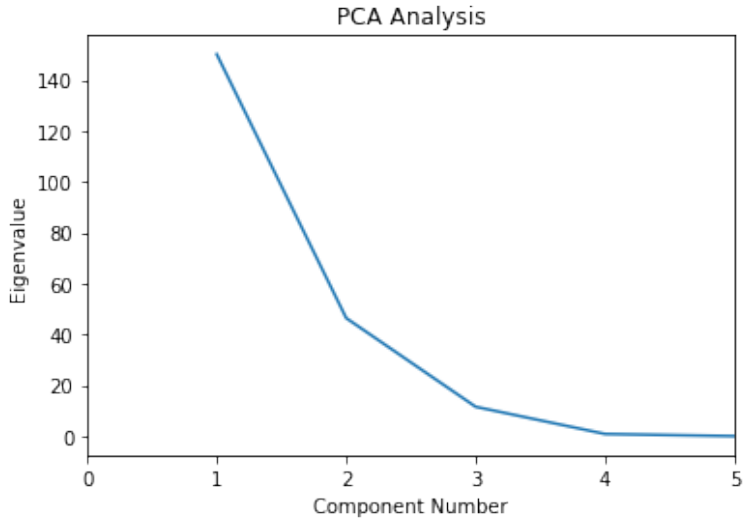
ii. For the Senate survey, how many surveys would you recommend sending to people in Alaska?

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?

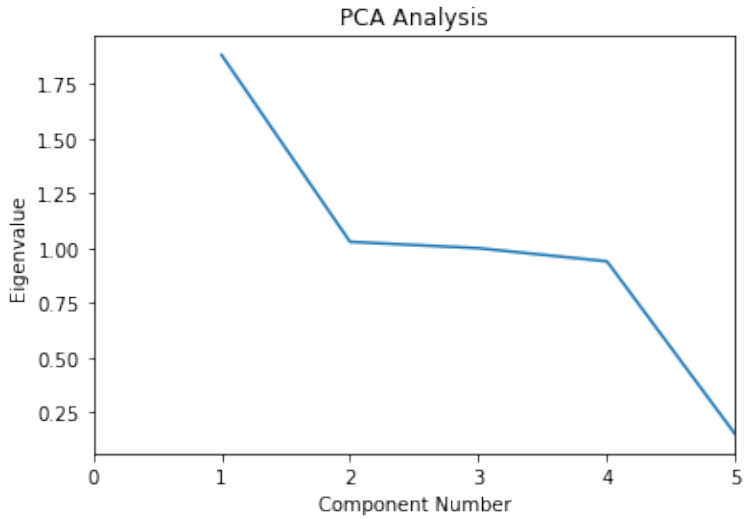
4. Dimensionality Reduction (12 points) [**Ge Gao**]

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.



	PC1	PC2	PC3	PC4	PC5
month	-0.051894	-0.025766	0.998317	0.002566	0.000237
hour	-0.122169	0.992322	0.019257	0.001352	0.000126
pm2.5	0.006882	-0.000472	-0.002225	0.999973	-0.001361
TEMP	-0.991128	-0.12097	-0.054659	0.006641	-0.000731
PRES	0.000688	0.000208	0.000282	-0.001365	-0.999999

Figure 1: PCA1 on Raw Dataset



	PC1	PC2	PC3	PC4	PC5
month	-0.187891	-0.237223	-0.680689	0.66142	0.08722
hour	-0.151187	-0.389684	0.720118	0.546767	0.088111
pm2.5	0.04463	0.863686	0.131547	0.472158	-0.108701
TEMP	-0.697775	-0.003108	0.01064	-0.094765	-0.709935
PRES	0.673022	-0.21426	-0.025957	0.177916	-0.684694

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) 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?

(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.

(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.

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

5. Discretization (12 points) [**Ge Gao**]

Consider the following dataset:

PATIENT	CHLORIDE	POTASSIUM	DATE	NORMAL
1	105	4.1	01/17/2005	yes
2	97	3.8	01/17/2005	yes
3	91	3.2	01/17/2005	no
4	104	4.1	01/18/2005	yes
5	111	5.6	01/18/2005	no
6	108	3.8	01/18/2005	yes
7	95	2.7	01/18/2005	no
8	97	4.6	01/19/2005	yes
9	99	3.9	01/19/2005	yes
10	97	3.5	02/02/2005	yes
11	98	4.7	02/02/2005	yes
12	102	3.7	02/02/2005	yes
13	109	6.0	02/02/2005	no
14	90	6.5	02/04/2005	no
15	103	5.0	02/04/2005	yes

(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.

(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.

(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 CHLORIDE above is $\bar{x} = 100$ and that the standard deviation $\sigma = 6$. Discretize CHLORIDE 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) [**Yang Shi**]

(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. Hamming distance between two numeric vectors
iii. 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) **CSC522: Required / CSC422: Extra Credit (6 points)**
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) [**Yang Shi**]

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 *hw*1.*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 *hw*1.*R*. Function calls are specified in *hw*1_*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 *hw*1.*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 inputs.
- Calculating the distances usually takes no longer than **20 seconds**.

Question

Dataset You are given the following dataset(s):

- Iris dataset [1]. You are provided a subset of the Iris dataset. Each line in *iris.csv* represents a five element vector, representing a single sample from your dataset. Each value if the first four columns are the attributes of the sample, respectively "sepal length", "sepal width", "petal length", "petal width". The fifth column is the class values, specifying which class of iris plant the sample is from. In total, there are 60 sample points.

Part 1: Distance Measurement Before doing analysis, you will need to look through the data file, and write a function named `read_data` to read the dataset in as a dataframe.

1a) Using the data provided in *iris.csv*, you are to implement the distance/similarity measurements defined below. The inputs will be two vectors of the same length:

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

(b) **cosine**: cosine(P, Q) = 1 - $\frac{\sum_i P_i * Q_i}{||P|| * ||Q||}$, where P and Q are vectors of equal length, and $||P|| = \sqrt{\sum_i P_i^2}$.

(c) L_∞: $L_\infty(P, Q) = \max_i |P_i - Q_i|$, where P and Q are vectors of equal length.

1b) Your goal is to investigate how useful each distance function is in telling apart flowers of different species. Ideally, a distance measure should be large for flowers of different species, and relatively smaller for flower of the same species.

To help you with this task, we have provided you with a function: `inter_intra_species_dist`. This function calculates the distance between each flower in the provided iris dataset, using the specified distance function. It then averages the following properties for each flower species:

(a) `mean_intra_dis`: The average distance between flowers of *this* (same) species.

(b) `mean_inter_dis`: The average distance between flowers of this species and *other* (different) species.

(c) `ratio`: The ratio of `mean_intra_dis` / `mean_inter_dis`

In your PDF report, use this function to answer the follow question: Which of the distance metrics that you implemented is most useful for differentiating iris species? Why do you think it is most useful?

Part 2: Principal Component Analysis In this part, you will need to implement a function to calculate the principal components (PCs) of a dataset in the function `principal_component_analysis`. *You are encouraged to leverage the existing function in R, which is* `prcomp`. The input of this function would be an iris dataframe, and you may need to note that the final column is a nominal value, which cannot included in the calculation of PCA. The output of the function is a vector of the weights (the eigenvector) of first principal component. *Hint*: Use `?prcomp` for more information on how to use the function.

After calculating the PC, the next step to do write a function, `principal_component_calculation`, to calculate a PC value given a data object and the component weights.

Part 3: Principal Component Distance

We want to see whether the first PC meaningfully captures the differences between iris species. Implement the `pc1_distance` function, which takes in two data objects (vectors) and a set of PC weights, and returns the distance between those two vectors in the dimension of the first PC, i.e. the absolute difference between their PC values.

Part 4: Comparing Distances

Now we want to compare our PC1 distance to traditional euclidean distance. In your PDF, use the `inter_intra_species_dist` function to answer the following question: Which of the two distance metrics (euclidean, PC1) is most useful for differentiating iris species? Why do you think it is most useful?

Note: *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 for you to understand how to run the code with necessary implementations that are not required. DO NOT submit *hw1_checker.R*. Also, please note that the TA may be using a dataset different to yours, 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, plyr. No other packages are allowed.

References

[1] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.