ALDA Fall 2023
HW1
09/06/2023

HW1 contains 7 questions. Please read and follow the instructions.

- **DUE DATE FOR SUBMISSION: 09/06/2023 11:45 PM**
- **TOTAL NUMBER OF POINTS: 130** (+5 bonus points if you follow all the instructions and 0 otherwise)
- **NO PARTIAL CREDIT** will be given so provide concise answers.
- **You MUST manually add ALL team members in the submission portal when you submit through Gradescope. One submission per group. Team member(s) who are left out will lose 20 points if added after the deadline.**
- Make sure you clearly list **your homework team ID, all team members' names and Unity IDs, for those who have contributed to the homework contribution** at the top of your submission.
- **[GradeScope and NCSU Github]:** Submit a PDF on GradeScope. **You must submit your code, and give the instructors access.** To do so, create a repository on NCSU GitHub for your homework group. Follow this naming convention:

  `engr-ALDA-Fall2023-XXX` where `XXX` is your homework group number.
  For example, if your homework group is H2, then: `engr-ALDA-Fall2023-H2`

  Follow these instructions: Upon signing into your NCSU GitHub account, on the left-hand side, click on the green button that says "New" to begin creating your code repository. Type in your repository name as outlined above. **Do NOT make your code repository public.** Now, "Create repository". Go to your repository's "Settings", and then on the left, select "Collaborators". Confirm account access if necessary, and add your team members by using their username, full name, or email address. **Then, add the instructors: `instructor`.** Create a folder for each homework (there are five) e.g. "HW1", "HW2", etc. **All code MUST be in its corresponding folder before the homework deadline. No credit will be given if the code is not submitted for a programming question.** In your PDF submitted on GradeScope, reference the script/function for each question (e.g. "For the solution to question 2, see matrix.py"). **Include your team's GitHub repository link in the PDF.**
- The materials on this course website are only for use of students enrolled in this course and **MUST NOT** be retained or disseminated to others.
- By uploading your submission, you agree that you have not violated any university policies related to the student code of conduct (`https://policies.ncsu.edu/policy/pol-11-35-01/`), and you are signing the Pack Pledge: **"I have neither given nor received unauthorized aid on this test or assignment"**.

1. (20 points) [**Data Attributes**] [**Graded by Md Mirajul Islam**] Classify the following attributes as:
   1) nominal, ordinal, interval, or ratio; **and** 2) as binary, discrete, or continuous. Some cases may have more than one interpretation, so briefly justify your answer if you think there may be some ambiguity.

   (a) (1 point) Month of a year (E.g. January, February, March, etc.)

   (b) (1 point) Heart rate in beats per minute

   (c) (1 point) Temperature in degrees Fahrenheit

   (d) (1 point) Correct or Incorrect

   (e) (1 point) Distance travelled in meters

   (f) (1 point) Velocity of a car in kilometers per hour

   (g) (1 point) Number of flowers in a bouquet

   (h) (1 point) IQ scores

   (i) (1 point) School Name

   (j) (1 point) Height of a mountain in kilometers

   (k) (1 point) Runs scored in a game of Cricket

   (l) (1 point) Average runs scored by a Batsman in a game of Cricket

   (m) (1 point) Blood type (O+, O-, A+, A-, B+, B-, AB+, AB-).

   (n) (1 point) Open or Closed

   (o) (1 point) Alphabets in a word

   (p) (1 point) Latitude and longitude coordinates

   (q) (1 point) NCSU Unity Id

   (r) (1 point) Size of a soccer field in square feet

   (s) (1 point) Shoe size

   (t) (1 point) MAC address of a device

2. (15 points) [**Matrix Operations**] [**Graded by Rajesh Debnath**] Write code in Python to perform each of the following tasks, please report your output and relevant code in the document file, and also include your code file (ends with extension **.py**) in the .zip file. Use NumPy for the code written for this question.

   (a) (1 point) Generate a 5*5 matrix of all ones A.

   (b) (1 point) Change all elements in the 3^{rd} row to 3.

   (c) (1 point) Sum each row of the matrix in array s.

   (d) (2 points) Concatenate s to the right of A (make s the last column of A). A should now have 6 columns.

   (e) (2 points) Transpose the updated matrix A ($A = A^T$).

   (f) (2 points) Generate matrix B such that its first row contains the standard deviation of each row of A, and the second row contains a uniformly random number in [0, 1]. Use seed=2023 in numpy.random.

   (g) (2 points) Generate a 2*5 matrix C such that $C = BA$.

   (h) (2 points) $X = [1, 2, 3, 4]^T$, $Y = [1, 2, 4, 8]^T$, $Z = [7, 3, 5, 1]^T$. Compute the covariance matrix of X, Y, and Z. Then compute the Pearson correlation coefficients between X and Y.

   (i) (2 points) Verify the equation: $\overline{x^2} = (\bar{x}^2 + \sigma^2(x))$ using $x = [7, 12, 10, 9, 14, 10, 11]^T$ when (python library *math* is allowed):

      i. $\sigma(x)$ is the **population** standard deviation. Show your work.

      ii. $\sigma(x)$ is the **sample** standard deviation. Show your work.

3. (24 points) [**Data Visualization**] [**Graded by Seongsoo Kim**] In this question, please summarize and explore the data in the "Abalone" dataset from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets/Abalone`) which is also provided as "`abalone.zip`". A description of the data set is provided in the `abalone.names` file and the data is provided in the `abalone.data` file. The data file is stored in a csv format.

   Write code in Python to perform the following tasks. Please *report your output and relevant code* in the document file, and also include your code file (ends with extension .py) in the .zip file.

   (a) (4 points) Compute the mean, median, standard deviation, range, 25th percentiles, 50th percentiles, 75th percentiles for the following attributes: *length*, *diameter*, *height*, *whole weight*.

   (b) (3 points) Make a box-and-whisker plot for the attributes *length* and *whole weight* where they are grouped by the *sex* label. Be sure to include a title for each plot of what feature is being described.

   (c) (4 points) Create histogram plot using 16 bins for the two features *shell weight* and *rings*, respectively.

   (d) (4 points) Create a scatter matrix of the data. Include only the following features: *diameter*, *height*, *shell weight*, and *rings*, but use the Sex attribute to change the color of the data points (for convenience, you may use a library for this). For the diagonal of the scatter matrix, plot the kernel density estimation (KDE).

   (e) (5 points) Now, write code to produce a three-dimensional scatter plot using the *diameter*, *shell weight* and *rings* as dimensions, and color the data points according to the *sex* attribute.

   (f) (4 points) The quantile-quantile plot can be used for comparing the distribution of data against the normal distribution. Create a quantile-quantile plot for the two features *length* and *whole weight*, respectively. Give a brief analysis of the two plots.

4. (16 points) [**Short Answer Questions**] [**Graded by Safaa Mohamed**] Please review lecture notes to answer the following questions:

   (a) (10 points) General Short Answer

   i. (3 points) Which distance metric would best describe this: How far away is the agent from their goal in a GridWorld environment? Justify your answer.

   ii. (3 points) What is the definition of covariance? If variables A and B have a covariance of -345 while variables B and C have a covariance of 20. What claims can you draw? Justify your answer.

   iii. (4 points) Provide a scenario in which you might encounter duplicate data. What could have caused the data to be duplicated? How would it be detected? Provide a solution to resolve the duplication, and state the pros/cons.

   (b) (6 points) Noise and Outliers

   i. (2 points) In your own words, explain what is noise. Can noise ever be desirable? If so, give an example when it is desirable. If not, provide an explanation of why not.

   ii. (2 points) In your own words, explain what is an outlier. How could outliers be detected? How do outliers differentiate from noise?

   iii. (2 points) Suppose you are analyzing a dataset of students' exam scores for a particular course. As you explore the data, you notice that there are a few instances where students have exceptionally high or low scores that are significantly different from the rest of the dataset. Upon closer examination, you find that these extreme scores are the result of data entry errors made during the recording process. For example, a student may have mistakenly been assigned a score of 1000 instead of 100. In addition to these extreme values, you also observe that there are some scores that slightly deviate from the expected range, but they are not related to any specific data entry errors or exceptional circumstances. These deviations occur sporadically throughout the dataset. Based on this scenario, please answer the following question: Are the extreme scores (high and low) outliers or noisy data? What about the slightly deviating scores? Justify your answer.

5. (9 points) [**Sampling**] [**Graded by Safaa Mohamed**] Answer the following questions:

   (a) (5 points) A chess dataset contains records for 10,000 unique games, where 6,000 games result in a win for player white and 4,000 games result in a win for player black. Among the 10,000 unique games, 3% witness an en passant move, and 11% have a castling move occur at least once on either the King's side or Queen's side. For simplicity, assume that the two events - en passant or castling - are mutually exclusive. Suppose we are developing a classifier in hopes of predicting the outcome of a chess game as it is being played. However, we are unable to use the entire data set due to computational limitations, and thus can only use a sample of the entire data set. Which sampling method would be appropriate and why? If we are sampling 2,000 games from the provided dataset, how many games should be selected from each group using your choice of sampling methods? Briefly justify your answer

   (b) (4 points) You work for a market research company that specializes in assessing consumer preferences for various products. You have been assigned to conduct a study on the purchasing habits of smartphone users in a particular city. The city has a population of 100,000 smartphone users. Due to time and resource constraints, you are required to collect data from a sample of 1,000 smartphone users. Now, based on this scenario, here's your question: Which sampling method would be appropriate to select a sample of 1,000 smartphone users from the population of 100,000, and why? Justify your answer.

6. (16 points) [**Data Transformation**] [**Graded by Seongsoo Kim**]

   (a) Please identify the appropriate data transformation methods for the following situations. Give a brief justification for your answers:

      i. (4 points) You have proposed an exponential growth model ($y = a(1 + r)^x$) for the population growth of deer in your area. Provide a transformation to make the model linear ($f(y) = mx + b$).

      ii. (4 points) You are creating a model to predict a person's risk of heat stroke. Your model considers two features: outside temperature (max = 105.0°F, min = 48.0°F, mean = 80.1 °F, standard deviation = 10.3 °F), and relative humidity (max = 100.0%, min = 0.0%, mean = 76.0%, standard deviation = 18.0%). 1) For each feature, apply normalization (transformed data has: $x' \in [0, 1]$) and calculate the new mean and new standard deviation of the normalized feature. Compare their means and standard deviations. And 2) for each feature, apply standardization to it and show the range of transformed data and compare their ranges.

   (b) In natural language processing (NLP), there are diverse ways to represent words such as one-hot encoding, bag of words, TF*IDF, and distributed word representations. In **one hot encoding**, a bit vector whose length is the size of the vocabulary of words is created, where only the associated word bit is on (i.e., 1) while all other bits are off (i.e., 0). Here is a toy example: suppose there is a 5-dimensional feature vector to represent a vocabulary of five words: [king, queen, man, woman, power]. In this case, 'king' is encoded into [1,0,0,0,0], 'queen' is encoded into [0,1,0,0,0], etc. Due to the nature of this representation, the feature vector encodes the vocabulary of a sentence where all words are equally distant. On the other hand, in **distributed word vectors**, a real-valued vector whose length is defined by *some common properties of words* is created, then each word can be represented as a linear combination of the defined properties. Using the toy example above, given a 3-dimensional feature vector of [man, woman, power] as the common properties, then words such as 'king', 'queen', 'man', and 'woman' could be encoded into [0.98, 0.1, 0.8], [0, 0.99, 0.85], [0.9, 0, 0.5], and [0, 0.97, 0.5], respectively. In this case, if you subtract a vector of 'man' from a vector of 'king', and add a vector of 'woman', then you will get a vector close to a vector of 'queen'.

      i. (4 points) What is a major disadvantage of one hot encoding as compared to distributed word vectors. Briefly justify your answer.

      ii. (4 points) What is a major disadvantage of distributed word vectors as compared to one hot encoding. Briefly justify your answer.

7. (30 points) [**Distance**] [**Graded by Md Mirajul Islam**] For this exercise, use the provided file "`SeoulBikeData.csv`" (Here), which contains a list of 8760 data instances. There are 14 attributes, including the class attribute; please refer to the included link for documentation. For this exercise, we will only be concerned with a selected few – namely, the *Visibility (10m)* and *Humidity(%)* attributes. Write code in Python to perform the following tasks, please report your output and relevant code in the document file, and also include your code file (ends with .py) in the .zip file.

   (a) (4 points) Data sets sometimes must be cleaned before they can be used. For this question, we have left the data set as it was stored online. Parse and clean the data file into an accessible representation e.g. a Pandas DataFrame. You may consider beginning by examining the provided file for possible problems when reading it.

   Then, generate a scatter plot between the *Visibility (10m)* and *Humidity(%)* of the observations. Label the axes (*Visibility (10m)* should be the x-axis and *Humidity(%)* should be the y-axis). Call this plot "Visibility and Humidity". What general interpretation can you make from this plot?

   (b) (2 points) Define a data point called P such that P = (`mean`(*Visibility (10m)*), `mean`(*Humidity(%)*)). For the remaining parts, please use the transformed attributes.

   (c) (10 points) Compute the distance between P and the rest of the data points using the following distance measures: 1) Euclidean distance, 2) Manhattan block metric, 3) Minkowski metric (for power=7), 4) Chebyshev distance, and 5) Cosine distance. List the closest 6 points for each distance.

   (d) For each distance measure, identify the 20 points from the dataset that are the closest to the point P from (b). (You are allowed to use any package functions to calculate the distances.)

      i. (10 points) Create plots, one for each distance measure. Place P on the plot and mark the 20 closest points. To mark them, you could use different colors or shapes. Make sure the points can be uniquely identified.

      ii. (4 points) Verify if the set of points is the same across all the distance measures. If there is any big difference, briefly explain why it is.