ALDA Fall 2022
HW1
08/24/2022

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

- **DUE DATE FOR SUBMISSION: 9/6/2022 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-Fall2022-XXX` where `XXX` is your homework group number.
  For example, if your homework group is H2, then: `engr-ALDA-Fall2022-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 code is not submitted for a programming question.** In your PDF submitted on GradeScope, reference the script/function for each question (e.g. "For 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) Temperature in kelvin (absolute temperature scale)
   (b) (1 point) Temperature in degrees Celsius
   (c) (1 point) True or False
   (d) (1 point) Placement in a race (E.g. First, Second, Third, etc.)
   (e) (1 point) Speed in miles per hour
   (f) (1 point) Number of Legos in a set
   (g) (1 point) Time of the day to the nearest minute
   (h) (1 point) First Name
   (i) (1 point) Product Satisfaction Rating (Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied)
   (j) (1 point) Ocean depth in meters
   (k) (1 point) Points scored in a game of basketball
   (l) (1 point) Average points scored in a game of basketball over a career
   (m) (1 point) Letter grade in a class (E.g. A+, A, A-, B+, B, etc.)
   (n) (1 point) Open or Closed
   (o) (1 point) Pages in a book
   (p) (1 point) Result of the roll of a dice
   (q) (1 point) Email address
   (r) (1 point) Amount of calories in a food
   (s) (1 point) Size of house in square feet
   (t) (1 point) Cumulative grade point average

2. (15 points) [**Matrix Operations**] [**Graded by Anurata Hridi**] 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.

   (a) (1 point) Generate a 5*5 identity matrix A.

   (b) (1 point) Change all elements in the 5^{th} column of A to 5.

   (c) (1 point) Sum of all elements in the matrix (use ONE "for/while loop").

   (d) (1 point) Transpose the matrix A (A=A^T).

   (e) (2 points) Calculate the sum of the 5^{th} row, the sum of the diagonal and the sum of the 1^{st} column in matrix A, respectively (your answer should be three numbers). Use the transposed matrix from the previous part.

   (f) (1 point) Generate a 5*5 matrix B following standard normal distribution.

   (g) (2 points) From A and B, using matrix operations to get a new 2*5 matrix C such that, the first row of C is equal to the 1^{st} row of B minus the 1^{st} row of A, the second row of C is equal to the sum of the 5^{th} row of A and the 5^{th} row of B.

   (h) (2 points) From C, using ONE matrix operation to get a new matrix D such that,the first column of D is equal to the first column of C, the second column of D is equal to the second column of C times 2, the third column of D is equal to the third column of C times 3, and so on.

   (i) (2 points) $X = [1, 1, 1, 2]^T$, $Y = [0, 3, 6, 9]^T$, $Z = [4, 3, 2, 1]^T$. Compute the covariance matrix of X, Y, and Z. Then compute the Pearson correlation coefficients between X and Y.

   (j) (2 points) Verify the equation: $\bar{x^2} = (\bar{x}^2 + \sigma^2(x))$ using $x = [23, 19, 21, 22, 21, 23, 23, 20]^T$ 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 Anurata Hridi**] In this question, please summarize and explore data in the provided file "seeds\dataset.csv". This data is the "Seeds Data Set" and donated by M. Charytanowicz to the UCI Machine Learning Repository. (`http://archive.ics.uci.edu/ml/datasets/seeds`).

   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: *area*, *perimeter*, *length of kernel*, *width of kernel*.

   (b) (3 points) Make a box-and-whisker plot for the attributes *length of kernel* and *width of kernel* where they are grouped by the *class* 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 *asymmetry coefficient* and *compactness*, respectively.

   (d) (4 points) Create a scatter matrix of the data. Include only the following features: *area*, *compactness*, *length of kernel*, *width of kernel* 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 *length of kernel*, *width of kernel* and *area* as dimensions, and color the data points according to the *class* 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 of kernel groove* and *compactness*, respectively. Give a brief analysis for the two plots.

4. (16 points) [**Short Answer Questions**] [**Graded by Safaa Mohamed**] Please read Chapters 2 in Tan et al., textbook and 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 apart are two binary vectors of the same length? Justify your answer.

      ii. (3 points) What is the definition of covariance? If variables A and B have a covariance of -1 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 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) You are performing analysis of data collected during a recreation of the Millikan oil drop experiment, a famous experiment which was used to determine the charge of an electron. Each trial of the experiment consists of two measurements performed on slightly charged drops of oil. First, the velocity of a drop of oil falling between two plates is measured. Second, an electric field between the two plates is increased until the drop of oil is suspended in place. Using these two measurements you are able to determine the charge on each drop of oil and finding a common divisor of these charges determines the charge of an electron. While doing your data analysis you find that for some measurements the lower plate is improperly grounded changing the strength of the electric field from what was measured. Are these measurements examples of noise in the data or examples of outliers in the data? Briefly justify your answer.

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

   (a) (3 points) A healthcare data set contains information for 119,968 patients. In the data set 13% of patients are left-handed, 85% are right-handed, and 2% are either ambidextrous or mixed-handed. Suppose we are performing analysis of how handedness affects the likelihood to sustain certain injuries. We would like to take a sample of 3,000 patients to perform preliminary analysis. Which sampling method would be appropriate and why? Briefly justify your answer.

   (b) (3 points) The Brown Corpus is a collection of English text samples which consists of over a million words sampled from 15 text categories in proportion to the amount published in each category. Suppose we would like to find the frequency of 5 word phrases in the English language, but due to time and hardware constraints we are unable to work with the entire data set and must take a sample on which we will perform our analysis. We decided to take a sample of 100,000 words. Which sampling method would be appropriate and why? Briefly justify your answer.

   (c) (3 points) In the task described in part (b) we notice that our results vary drastically if we draw a new sample. Now, instead of fixing our sample size, we would like to draw a sample such that the result converges to the true frequency of each phrase. Which sampling method would be appropriate and why? Briefly justify your answer.

6. (16 points) [**Data Transformation**] [**Graded by Chengyuan Liu**]

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

      i. (4 points) During the design of an artificial neural network, it is often desirable for the output of the network to be constrained from $[0, 1]$ such as for a classification task. Define a function that transforms an unbounded value $(-\infty, \infty)$ to one in the range $[0, 1]$. Please provide a function that monotonically increases.

      ii. (4 points) In polynomial interpolation, the choice of nodes can drastically affect the quality of your fit. Chebyshev nodes can be shown to minimize the interpolation error in the infinity norm. Chebyshev nodes are defined on the interval $x \in (-1, 1)$. However, if our data instead is in the domain $x \in (a, b)$ we must perform a transformation. Define a formula for the necessary data transformation. Additionally, perform this transformation on a body temperature variable ranged from 84 to 112 (mean = 98.6, standard deviation = 0.95). What is the new mean and standard deviation?

   (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) Sentiment analysis is the use of NLP to identify the emotions associated with a text. Sentiment analysis is commonly used to analyze product reviews and social media amongst other things. Knowledge-based approaches to sentiment analysis use words such as good, bad, happy, and sad to classify a text as positive or negative. Which word representation, one-hot encoding or distributed word vectors, would be more useful for a knowledge-based approach? Briefly justify your answer.

      ii. (4 points) You have been tasked with building a tool to identify synonyms and antonyms of words. Which word representation, one-hot encoding or distributed word vectors, would be more useful? Briefly justify your answer.

7. (30 points) [**Distance**] [**Graded by Md Mirajul Islam**] For this exercise, use the provided files "`Algerian_forest_fires_dataset_UPPDATE.csv`" (Here), which contains a list of 244 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 select few – namely, the *relative humidity* (RH) and *wind speed* (Ws) 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 a accessible representation e.g. a Pandas DataFrame. You may consider beginning by examining the provided file for possible problems when reading it in.

   Then, generate a scatter plot between the *relative humidity* and the *wind speed* of the observations. Label the axes (*relative humidity* should be x-axis and *wind speed* should be y-axis). Call this plot "relative humidity and wind speed". What general interpretation can you make from this plot?

   (b) (2 points) Define a data point called P such that $P = (\texttt{mean}(RH), \texttt{mean}(Ws))$. For the remaining parts, please use the transformed attributes.

   (c) (10 points) Compute the distance between P and the 244 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.