ALDA Fall 2021
HW1
08/18/2021

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

- **DUE DATE FOR SUBMISSION: 08/31/2021 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.**
- 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-fall2021-XXX` where `XXX` is your homework group number.
  For example, if your homework group is H2, then: `engr-ALDA-fall2021-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: `jwhostet`, `ggao5`, `azhang25`, and `cliu32`.** 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**] [**Designed by Angela Zhang, Graded by Chengyuan Liu**] 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) The pH of water that has been measured using the pH scale.

   (b) (1 point) Population counts in counties of North Carolina.

   (c) (1 point) Active Duty Enlisted Basic Military pay grade scale e.g. E-1, E-2, ..., E-9(for simplicity of the exercise, do not factor in years of service in your answer).

   (d) (1 point) The amount of Calories found on a nutrition label.

   (e) (1 point) The distance traveled on a road trip measured in miles.

   (f) (1 point) How much time until midnight?

   (g) (1 point) Percentage of daily vitamin C intake from a glass of orange juice.

   (h) (1 point) Existence and non-existence of cancerous tumors.

   (i) (1 point) Annual income (US currency).

   (j) (1 point) An angle measured in radians between 0 and 2π.

   (k) (1 point) Number of fishes in an aquarium.

   (l) (1 point) Required or Not Required.

   (m) (1 point) Level of Pain during Injury diagnosis (0: absent, 1: mild, 2: moderate, 3: severe, 4: incapacitating).

   (n) (1 point) Circumference of a pizza.

   (o) (1 point) Product Satisfactory Rating (5 - Excellent, 4 - Good, 3 - Fair, 2 - Poor, 1- Horrible).

   (p) (1 point) Temperature in degrees Fahrenheit.

   (q) (1 point) A person's weight measured in pounds.

   (r) (1 point) Correct or Incorrect.

   (s) (1 point) Number of pieces remaining to complete a jigsaw puzzle.

   (t) (1 point) Day of the month.

2. (15 points) [**Matrix Operations**] [**Ge Gao**] 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 identify matrix A.

   (b) (1 point) Change all elements in the 3^{rd} column of A to 2.

   (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 3^{rd} row, the sum of the diagonal and the sum of the 2^{nd} column in matrix A, respectively (your answer should be three numbers).

   (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 2^{nd} row of A, the second row of C is equal to the sum of the 4^{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 times 2, the second column of D is equal to the second column of C times 3 and so on.

   (i) (2 points) $X = [1, 3, 5, 7]^T$, $Y = [4, 3, 2, 1]^T$, $Z = [2, 4, 6, 8]^T$. 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 $x = [20, 1, 3, 5, 7, 9, 14]^T$ 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**] [**Designed by John Wesley Hostetter, Graded by Chengyuan Liu**] In this question, please summarize and explore data in the provided file "data\wine.csv". This data is the "Wine Data Set" and is a famous database donated by Stefan Aeberhard to the UCI Machine Learning Repository.

   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: *Alcohol*, *Malic acid*, *Ash*, *Alcalinity Of Ash*.

   (b) (3 points) Make a box-and-whisker plot for the attributes *Ash* and *Malic acid* 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 *Proanthocyanins* and *Proline*, respectively.

   (d) (4 points) Create a scatter matrix of the data. Include only the following features: *Flavanoids*, *Total phenols*, *Ash*, *Malic acid*, but use the *class* 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 *Proanthocyanins*, *Flavanoids* and *Total phenols* 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 *Ash* and *OD280/OD315 of diluted wines*, respectively. Give a brief analysis for the two plots.

4. (16 points) [**Short Answer Questions**] [**Angela Zhang**] 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 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 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) While conducting preliminary data exploration, you noticed that some patients' Electronic Healthcare Records (EHRs) had anomalous values that caught your attention. Specifically, you noticed that these patients' electrocardiogram results were significantly different from typical electrocardiogram behavior, but noticed that all these patients were in the most recent week of the collected data. You consult with your research partner, the oncologist, about your findings. Upon inspecting the electrocardiogram (EKG) machine that was used to collect the data, the oncologist discovers a fault within the machine's operation, that led to corrupt electrocardiogram results. Unfortunately, the results are not salvageable, as the corruption appears to occur randomly throughout the electrocardiogram readings. Were the corrupted electrocardiogram results an example of noisy data or were they outliers in the data? Briefly justify your answer.

5. (9 points) [**Sampling**] [**John Wesley Hostetter**] Answer the following questions:

   (a) (3 points) Exploration of the environment is a key principle to learning optimal actions with online reinforcement learning (RL). Imagine the RL agent reaches a state for which it has no prior knowledge to assist it in its decision making on what action it should take. For simplicity, assume that the environment is "safe" to explore within e.g. no human lives will be harmed or lost if the agent were to err. Let A denote the finite set of available actions to the agent when it reaches this novel state. Which sampling method would be appropriate and why? Briefly justify your answer.

   (b) (3 points) Assume you have a large collection of data (e.g. billions of logged human-computer interactions), and want to apply a machine learning algorithm to obtain a predictive model for some mundane task. Despite the size of the data, the data itself is not very complicated, and the domain it describes is rather trivial. In addition, the algorithm you would like to use has a time complexity of $O(n^3)$, and space complexity of $O(n^2)$, where n is the number of data observations. For the purpose of this exercise, assume your supervisor expects your algorithm to have *at least* 90% sensitivity on the validation data set. Which sampling method would be appropriate and why? Briefly justify your answer.

   (c) (3 points) An oncologist approaches your research laboratory to collaborate together on a cancer detection system. This system, when provided with a patient's electronic health record (EHR), would then output the predicted probability that the patient has cancer (irrespective on the type of cancer). Fortunately, this type of cancer is rare, among the 793,000 labeled EHRs, only 1,560 have cancer, where the rest do not. You are required to create three sets of balanced data from this collection. Which sampling method would be appropriate and why? Briefly justify your answer.

6. (16 points) Data Transformation. [**Created by Ge Gao and Graded by Chengyuan Liu**]

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

   i. (4 points) To learn a model to predict whether a patient suffers from organ failure, we need to consider two state features: heart rates and body temperature. Heart rates are distributed between 35 and 200 bpm (mean = 88, standard deviation = 18), and body temperatures are ranged from 84 to 112 (mean = 98.6, standard deviation = 0.95). 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.

   ii. (4 points) During the design of an artificial neural network, we sometimes need to transform a variable x that has a range of $(-\infty, \infty)$ to an open set $z \in (-1, 1)$. Note that z monotonically increases as x increases in this transformation. Please specify a proper function for such transformation.

   (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) Briefly describe a situation when one-hot encoding transformation is more desirable than distributed word vector transformation and explain the reason.

   ii. (4 points) Briefly describe a situation when distributed word vector transformation is more desirable than one-hot encoding transformation and explain the reason.

7. (30 points) [**Distance**] [**John Wesley Hostetter**] For this exercise, use the provided files "data\auto-mpg.data" and "data\auto-mpg.names" , which contains a list of 398 data instances. There are 9 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 *displacement* and *weight* 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 is not always readily available in .csv files, and sometimes must be appropriately formatted to facilitate data manipulation or analysis. Therefore, parse the .data file into a more accessible representation e.g. a Pandas DataFrame. You may consider beginning by examining the provided .data file for useful patterns to use with a regular expression.

   Then, generate a plot between the *displacement* and the *weight* of the observations. Label the axes (*displacement* should be x-axis and *weight* should be y-axis). Call this plot "Displacement and Weight Data". What general interpretation can you make from this plot?

   (b) (2 points) Compute the mean of the attributes *displacement* and *weight*. Define a data point called P such that $P = (\texttt{mean}(displacement), \texttt{mean}(weight))$.

   (c) (10 points) Compute the distance between P and the 398 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.