HOMEWORK 1: BACKGROUND MATERIAL

10-301/10-601 Introduction to Machine Learning (Summer 2022)
https://www.cs.cmu.edu/~hchai2/courses/10601/
OUT: Tuesday, May 17
DUE: Tuesday, May 24 at 1:00 PM
TAs: Ayush, Boyang (Jack), Brendon, Sana, Chu

START HERE: Instructions

- **Collaboration policy:** Collaboration on solving the homework is allowed, after you have thought about the problems on your own. It is also OK to get clarification (but not solutions) from books or online resources, again after you have thought about the problems on your own. There are two requirements: first, cite your collaborators fully and completely (e.g., "Jane explained to me what is asked in Question 2.1"). Second, write your solution *independently*: close the book and all of your notes, and send collaborators out of the room, so that the solution comes from you only. See the Academic Integrity Section on the course site for more information: https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus
- **Late Submission Policy:** See the late submission policy here: https://www.cs.cmu.edu/~hchai2/courses/10601/#Syllabus
- **Submitting your work:**
  - **Programming:** You will submit your code for programming questions on the homework to Gradescope (https://gradescope.com). After uploading your code, our grading scripts will autograde your assignment by running your program on a virtual machine (VM). When you are developing, check that the version number of the programming language environment (Python 3.9.6) and versions of permitted libraries (`numpy` 1.21.2) match those used on Gradescope. You have a **total of 10 Gradescope programming submissions.** Use them wisely. In order to not waste code submissions, we recommend debugging your implementation locally first before any Gradescope coding submission. The above is true for future assignments, but this one allows **unlimited submissions.**
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, we will be using Gradescope (https://gradescope.com/). Please use the provided template. Submissions can be handwritten onto the template, but should be labeled and clearly legible. If your writing is not legible, you will not be awarded marks. Alternatively, submissions can be written in LaTeX. Regrade requests can be made, however this gives the TA the opportunity to regrade your entire paper, meaning if additional mistakes are found then points will be deducted. Each derivation/proof should be completed on a separate page. For short answer questions you **should not** include your work in your solution. If you include your work in your solutions, your assignment may not be graded correctly by our AI assisted grader.
- **Materials:** The data that you will need in order to complete this assignment is posted along with the writeup and template on the course website.

- The LaTeXtemplate for this assignment is available here.

For multiple choice or select all that apply questions, shade in the box or circle in the template document corresponding to the correct answer(s) for each of the questions. For LaTeX users, replace `\choice` with `\CorrectChoice` to obtain a shaded box/circle, and don't change anything else.

Instructions for Specific Problem Types

For "Select One" questions, please fill in the appropriate bubble completely:

Select One: Who taught this course?

- ● Henry Chai
- ○ Marie Curie
- ○ Noam Chomsky

If you need to change your answer, you may cross out the previous answer and bubble in the new answer:

Select One: Who taught this course?

- ● Henry Chai
- ○ Marie Curie
- ⊗ Noam Chomsky

For "Select all that apply" questions, please fill in all appropriate squares completely:

Select all that apply: Which are scientists?

- ■ Stephen Hawking
- ■ Albert Einstein
- ■ Isaac Newton
- □ I don't know

Again, if you need to change your answer, you may cross out the previous answer(s) and bubble in the new answer(s):

Select all that apply: Which are scientists?

- ■ Stephen Hawking
- ■ Albert Einstein
- ■ Isaac Newton
- ⊠ I don't know

For questions where you must fill in a blank, please make sure your final answer is fully included in the given space. You may cross out answers or parts of answers, but the final answer must still be within the given space.

Fill in the blank: What is the course number?

10-601 10-6301

Written Questions (76 points)

1 Latex Bonus Point (1 points)

1. (1 point) **Select one:** Did you use LaTeX for the entire written portion of this homework?

   ◯ Yes

   ◯ No

2 Course Policies (8 points)

This section covers important course policies that every student should know and understand. These questions MUST be finished in order for the whole homework to be considered for grading.

1. (1 point) **Select one:** Assignments turned in late without prior approval will incur a daily penalty. How much is the penalty? Up to 1 day: ___ Up to 2 day: ___ Up to 3 day: ___ Up to 4 day: ___

   ◯ 5%, 10%, 15%, 20%

   ◯ 10%, 20%, 30%, 40%

   ◯ 25%, 50%, 75%, 100%

   ◯ 20%, 40%, 60%, 80%

2. (1 point) **Select one:** How many grace days do you have in total for all homework? Can you combine grace days with late days to extend a homework submission deadline by 4 days?

   ◯ As many as I want; Of course!

   ◯ 8; No

   ◯ 8; Yes

   ◯ 9; No

   ◯ 9; Yes

3. (1 point) **Select all that apply:** Seeking help from other students in understanding course materials needed to solve homework problems is ALLOWED under which of the following conditions?

   □ Any written notes are taken on an impermanent surface (e.g. whiteboard, chalkboard) and discarded before writing up one's solution alone.

   □ Learning is facilitated not circumvented; i.e., the purpose of seeking help is to learn and understand the problem instead of merely getting an answer

   □ Help both given and received is reported in collaboration questions in the homework

   □ The student updates his/her collaborative questions even if it is after submitting their own assignment

   □ None of the above

4. (1 point) **Select all that apply:** Which of the following is (are) strictly forbidden in solving and submitting homework?

   - □ Searching on the internet for solutions or sample codes
   - □ Consulting people outside this class who have seen or solved the problem before
   - □ Turning in someone else's homework
   - □ Using anyone else's, or allowing other classmates to use your computer or Gradescope account in connection with this course
   - □ None of the above

5. (1 point) **Select one:** If you solved your assignment completely on your own, you can skip the collaboration questions at the end of each homework.

   - ◯ True
   - ◯ False

6. (1 point) **Select all that apply:** What is (are) the consequence(s) of being caught cheating in this course?

   First time:

   - □ Failure of the course
   - □ AIV report to university authorities
   - □ Negative 100% on the assignment

   Second time:

   - □ Failure of the course
   - □ AIV report to university authorities
   - □ Negative 100% on the assignment

7. (1 point) **Select one:** Assume a difficult situation arises in the middle of the semester (e.g. medical, personal etc.) that might prevent you from submitting assignments on time or working as well as you would like. What should you do?

   - ◯ Email the education associates (EAs) for the course, your college liaison, and advisor (being sure to include the latter two in the case of a medical emergency) early so they can point you to the available resources on campus and make necessary arrangements
   - ◯ Do not speak to the course staff, try to finish the class, reach out to the course staff in the end of the semester explaining your special situation

8. (1 point) **Select one:** If you have an emergency or university approved travel and need to request an extension for one of the homework assignments in the course, what should you do?

   ◯ Email the professor of the course at least 3 days before the homework deadline

   ◯ Email the education associates (EAs) for the course at least 5 days before the homework deadline

   ◯ Post on Ed at least 4 days before the homework deadline

   ◯ Email the entire course staff the day before the homework deadline

3 Probability and Statistics (25 points)

Use the following data to answer questions 1. Consider data created by flipping a coin five times S = [1, 1, 0, 1, 0] , where 1 denotes that the coin turned up heads and 0 denotes that it turned up tails.

1. (2 points) **Select one:** What is the probability of observing any combination of this data (3 heads and 2 tails), assuming it was generated by flipping a coin X with an unequal probability of heads (1) and tails (0), where the distribution is $P(X = 1) = 0.6$, $P(X = 0) = 0.4$?

   ◯ $\frac{1024}{3125}$

   ◯ $\frac{216}{625}$

   ◯ $\frac{162}{625}$

   ◯ $\frac{1}{32}$

2. (1 point) **Select one:** For events A and B, where $A \cap B$ indicates A AND B, and $A \cup B$ indicates A OR B,

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

   ◯ True

   ◯ False

3. (1 point) **Select one:** For events A and B,

$$P(A_1 \cap A_2 \cap A_3) = P(A_3|A_2 \cap A_1)P(A_2|A_1)P(A_1)$$

   ◯ True

   ◯ False

4. (1 point) **Select one:** For some events $A_1, A_2, \ldots, A_n$,

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq P(A_1) + P(A_2) + \cdots + P(A_n)$$

   ◯ True

   ◯ False

5. (2 points) **Select one:** Whether your car is wet in the morning (W) is dependent on whether it rained last night (R) or not, however other factors may have lead to your car being wet. The following are probabilities of such events:

$$P(R) = 0.8$$
$$P(W|R) = 0.9$$
$$P(W|\neg R) = 0.3$$

What is the probability that your car is wet in the morning?

- ◯ 0.5
- ◯ 0.78
- ◯ 0.56
- ◯ 0.64

Use the following information to answer questions 6-7. Consider the following joint probability table where both X and Y are binary variables:

X	Y	Probability
0	0	0.1
0	1	0.4
1	0	0.2
1	1	0.3

6. (1 point) **Select one:** What is $P(X = 1|Y = 1)$?

- ◯ $\frac{2}{3}$
- ◯ $\frac{3}{7}$
- ◯ $\frac{4}{5}$
- ◯ $\frac{3}{5}$

7. (1 point) **Select one:** What is $P(Y = 0)$?

- ◯ 0.2
- ◯ 0.6
- ◯ 0.5
- ◯ 0.3

Use the following information to answer questions 8-11. Let X be a random variable with expected value $E[X] = 1$ and variance $Var[X] = 2$.

8. (1 point) **Select one:** What is $E[6X]$?

   ◯ 1

   ◯ 3

   ◯ 6

   ◯ 36

9. (1 point) **Select one:** What is $Var[2X]$?

   ◯ 1

   ◯ 4

   ◯ 8

   ◯ 16

10. (1 point) **Select one:** What is $Var[2X - 3]$?

    ◯ 3

    ◯ 4

    ◯ 5

    ◯ 8

11. (1 point) **Select one:** What is $E[X^2]$?

    ◯ 3

    ◯ 4

    ◯ 5

    ◯ 8

Use the following information to answer questions 12-15: Let A, B, and C be random variables with discrete probability distributions. Consider the following two joint probability tables: one relating A and B, and the other relating B and C.

$A \backslash B$	b_1	b_2	b_3
a_1	0.1	0.05	0.15
a_2	0.1	0.05	0.3
a_3	0.05	0.15	0.05

$B \backslash C$	c_1	c_2	c_3	c_4
b_1	0.02	0.14	0.06	0.03
b_2	0.03	0.05	0	0.17
b_3	0.35	0.04	0	0.11

12. (1 point) **Select all that apply:** Which of the following statements are necessarily **false**? Note X $\perp\!\!\!\perp$ Y indicates that random variable X is independent of random variable Y.

    ☐ A $\perp\!\!\!\perp$ B

    ☐ B $\perp\!\!\!\perp$ C

    ☐ A $\perp\!\!\!\perp$ C

    ☐ None of the above.

13. (2 points) What is $P(B = b_1 | A = a_2, C = c_4)$? If this value cannot be computed, write N/A.

    Your Answer

14. (2 points) What is $P(B = b_2 | A = a_3, C = c_3)$? If this value cannot be computed, write N/A.

    Your Answer

15. (2 points) **Select one:** True or False: $\sum_{i=1}^{3} P(B = b_i | C = c_1) = \sum_{j=1}^{4} P(C = c_j | B = b_1)$

    ◯ True

    ◯ False

16. (2 points) **Select one:** Consider two random variables X, Y. Assume that we have $P(X = x) = \frac{1}{2^x}$ for $x \in \mathbb{Z}_{\geq 1}$ (integers greater than or equal to 1) and $P(Y = y | X = x) = \frac{1}{n}$ for $y \in \{1, 2, ..., n\}$. Assume n is a fixed positive integer constant. What is $\mathbb{E}[Y]$?

    ◯ $\sum_{y=1}^{n} y \frac{1}{2^y}$

    ◯ $\sum_{y=1}^{n} y \frac{5}{3^y}$

    ◯ $\sum_{y=1}^{n} \frac{y}{n}$

    ◯ $\sum_{y=1}^{n} y$

17. (1 point) **Select one:** What is the mean, variance and entropy of a Bernoulli (p) random variable?

- ◯ $p,\ \ p(1-p),\ \ -(1-p)\log(1-p) - p\log(p)$
- ◯ $p(1-p),\ \ p,\ \ -(1-p)\log(1-p) - p\log(p)$
- ◯ $p,\ \ p(1-p),\ \ \log(1-p) - p\log(p)$
- ◯ The entropy of a Bernoulli variable is not defined.

18. (2 points) Please match the probability density function of the random variable X to its corresponding distribution name.

A) $P(X = x) = \frac{1}{\sqrt{(2\pi)^d |\sum|}} \exp\left(-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right)$

B) $P(X = x) = \lambda e^{-\lambda x}$ when $x \geq 0$; 0 otherwise

C) $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

D) $P(X = x) = \frac{1}{b-a}$ when $a \leq x \leq b$; 0 otherwise

E) $P(X = x) = p^x (1-p)^{1-x}$

Multivariate Gaussian:

Your Answer

Exponential:

Your Answer

Uniform:

Your Answer

Bernoulli:

Your Answer

Binomial:

Your Answer

4 Calculus (8 points)

1. (2 points) Evaluate the derivative of y with respect to x, where $y = \ln\left(\frac{4}{x^3} - x^2\right)$ at x = 1.

Your Answer

2. (2 points) **Select one:** Find the partial derivative of y with respect to x, where $y = 3x^2 \sin(z)e^{-x}$.

- ◯ $3x\sin(z)e^{-x}(2+x)$
- ◯ $-6x\sin(z)e^{-x}$
- ◯ $3x\sin(z)e^{-x}(2-x)$
- ◯ $6x\cos(z)e^{-x}$

3. (2 points) **Select one:** For the function $f(x) = 4x^3 - 5x^2 - 2x$ the value $x = -\frac{1}{6}$ sets the derivative to be 0. Additionally, the second order derivative of $f(x)$ at x = $-\frac{1}{6}$ is negative. What can you say about $f(x)$ at the point $x = -\frac{1}{6}$:

- ◯ a local minimum
- ◯ a local maximum
- ◯ a local minimum or a local maximum
- ◯ None of the above

4. (2 points) **Select one:** Suppose that $f(\mathbf{x}|\boldsymbol{\theta}) = \mathbf{x}^T\boldsymbol{\theta}$, where $\mathbf{x}, \boldsymbol{\theta} \in \mathbb{R}^n$. The function $g(\boldsymbol{\theta})$ is defined as $g(\boldsymbol{\theta}) = (f(\mathbf{x}^{(1)}|\boldsymbol{\theta}) - y^{(1)})^2$ for $\mathbf{x}^{(1)} \in \mathbb{R}^n$ and $y^{(1)} \in \mathbb{R}$. What is the function type of $g(\boldsymbol{\theta})$:

- ◯ $g : \mathbb{R}^n \to \mathbb{R}$
- ◯ $g : \mathbb{R} \to \mathbb{R}$
- ◯ $g : \mathbb{R} \to \mathbb{R}^n$
- ◯ $g : (\mathbb{R}^n \times \mathbb{R}^n) \to \mathbb{R}$

5 Vectors and Matrices (15 points)

1. (1 point) **Select one:** Consider the matrix $\mathbf{X}$ and the vectors $\mathbf{y}$ and $\mathbf{z}$ below: $\mathbf{X}=\begin{bmatrix}3 & 5\\ 7 & 9\end{bmatrix}$, $\mathbf{y}=\begin{bmatrix}3\\ 7\end{bmatrix}$, $\mathbf{z}=\begin{bmatrix}6\\ 5\end{bmatrix}$. What is the inner product (dot product) of the vectors **y** and **z**?

   ◯ $\begin{bmatrix}18 & 15\\ 42 & 35\end{bmatrix}$

   ◯ 57

   ◯ $\begin{bmatrix}18\\ 35\end{bmatrix}$

   ◯ 53

2. (1 point) **Select one:** Using the same values for **X**, **y**, and **z** as above, what is the product **Xy**?

   ◯ $\begin{bmatrix}9 & 35\\ 21 & 63\end{bmatrix}$

   ◯ $\begin{bmatrix}44\\ 84\end{bmatrix}$

   ◯ $\begin{bmatrix}58\\ 78\end{bmatrix}$

   ◯ $\begin{bmatrix}9 & 15\\ 49 & 63\end{bmatrix}$

3. (2 points) **Select all that apply:** Consider $\mathbf{u} = \begin{bmatrix}1\\ 2\\ 3\end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix}0 & 7\\ 4 & 5\\ -1 & 0\end{bmatrix}$. Which of these are valid operations?

   □ $\mathbf{u}^T\mathbf{V}$

   □ $\mathbf{V}^T\mathbf{u}$

   □ $\mathbf{uV}$

   □ $\mathbf{VV}$

   □ None of the above

4. (2 points) **Select one:** For the matrices $\mathbf{A} = \begin{bmatrix} 2 & 1 & 4 \\ -3 & 2 & 0 \\ 1 & 3 & -2 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 3 & 4 & 5 \\ 3 & -1 & 3 \\ 1 & 3 & -2 \end{bmatrix}$, what is the product $\mathbf{AB}$?

   ◯ $\begin{bmatrix} 13 & 19 & 5 \\ -3 & -14 & -9 \\ 4 & -4 & 18 \end{bmatrix}$

   ◯ $\begin{bmatrix} 13 & 19 & 28 \\ 19 & 9 & -7 \\ -10 & -2 & 13 \end{bmatrix}$

   ◯ $\begin{bmatrix} 20 & -20 & -28 \\ 3 & -14 & 9 \\ 3 & 2 & 13 \end{bmatrix}$

   ◯ $\begin{bmatrix} 13 & 19 & 5 \\ -3 & -14 & -9 \\ 10 & -5 & 18 \end{bmatrix}$

5. (1 point) **Select one:** The matrix $\mathbf{A}$ from the previous question has an inverse.

   ◯ True

   ◯ False

6. (2 points) **Select one:** Consider two vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$ and let $z = \mathbf{x}^T\mathbf{y}$. What is $\frac{\partial z}{\partial x_2}$?

   ◯ y_2

   ◯ x_2

   ◯ $\mathbf{x}$

   ◯ $\mathbf{y}$

7. (2 points) **Select one:** Given matrix $\mathbf{X} = \begin{bmatrix} 3 & 4 & 2 \\ 1 & 6 & 2 \\ 1 & 4 & 4 \end{bmatrix}$ and the column vector $\mathbf{y} = \begin{bmatrix} -6 \\ 1 \\ 1 \end{bmatrix}$, what is the eigenvalue of $\mathbf{X}$ associated with $\mathbf{y}$? (Recall an eigenvector of a matrix $\mathbf{A} \in \mathcal{R}^{n \times n}$ is a nonzero vector $\mathbf{v} \in \mathcal{R}^n$ such that $\mathbf{Av} = \lambda\mathbf{v}$ where we call the scalar λ the associated eigenvalue for $\mathbf{v}$.)

   ◯ -5

   ◯ -3

   ◯ 2

   ◯ 1.5

8. (2 points) **Select one:** Preparing for his linear algebra final, Joe is finding eigenvectors and eigenvalues for different matrices. Joe finds out that some matrix A (not given) has eigenvalues 4 and 5. To find the associated eigenvectors for eigenvalue 4, he then solves the equation $\mathbf{Av} = 4\mathbf{v}$ and finds two solutions, $\begin{bmatrix} 3 \\ 117 \\ 9 \end{bmatrix}$ and $\begin{bmatrix} 1 \\ 39 \\ 3 \end{bmatrix}$. He concludes that there are two distinct eigenvectors corresponding to the eigenvalue of 4. Which statement regarding his solution is true?

   ◯ The solution must be wrong because there cannot be multiple eigenvectors corresponding to a single eigenvalue.

   ◯ The solution must be wrong because the eigenvectors are linearly dependent.

   ◯ The solution is correct because there may be multiple eigenvectors corresponding to an eigenvalue, and eigenvectors of a matrix are linearly dependent.

   ◯ The solution is correct because both vectors are solutions to $\mathbf{Av} = 4\mathbf{v}$.

9. (2 points) **Select all that apply:** Consider $\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 \\ -7 & -1 & -4 \\ 3 & 0 & 1 \end{bmatrix}$ and $\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Which of the following is true for $\mathbf{A}$ and $\mathbf{x}$?

   □ 0 is an eigenvalue of $\mathbf{A}$

   □ 0 is an eigenvalue of $\mathbf{A}^T$

   □ $\mathbf{x}$ is an eigenvector of $\mathbf{A}$

   □ $\mathbf{x}$ is an eigenvector of $\mathbf{A}^T$

   □ None of the above

6 Geometry (5 points)

1. (2 points) **Select one:** What relationship does the vector $\mathbf{w}$ share with the line $\mathbf{w^T x} + b = 0$? (assume $\mathbf{x}$ and $\mathbf{w}$ are both two dimensional column vectors, and $\mathbf{w}^T$ indicates the transpose of the column vector $\mathbf{w}$.)

   ◯ parallel

   ◯ orthogonal

   ◯ depends on the value of b

   ◯ depends on the value of $\mathbf{x}$

2. (1 point) **Select one:** With reference to the above question, select the statement which best explains why $\mathbf{w}$ and $\mathbf{w^T x} + b = 0$ share the above relationship.

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$, where $\mathbf{x}'$ and $\mathbf{x}''$ are two points on the line $\mathbf{w^T x} + b = 0$, is 0

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$, where $\mathbf{x}'$ and $\mathbf{x}''$ are two points on the line $\mathbf{w^T x} + b = 0$, is 1

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$, where $\mathbf{x}'$ and $\mathbf{x}''$ are two points on the line $\mathbf{w^T x} + b = 0$, is b

3. (2 points) **Select one:** What is the distance from the origin to the line $\mathbf{w^T x} + b = 0$?

   (In the following answers, λ is some constant)

   ◯ $\frac{|b|}{||\mathbf{w}||}$

   ◯ $\frac{|b|}{\mathbf{w^T w}}\mathbf{w}$

   ◯ $\frac{2\lambda}{\mathbf{w}b}$

   ◯ $\frac{||\mathbf{w}||}{|b|}$

7 CS Foundations (14 points)

1. (1 point) **Select one:** If $f(n) = 3^n$ and $g(n) = 2^n$ which of the following are true?

   ◯ $f(n) \in O(g(n))$

   ◯ $g(n) \in O(f(n))$

   ◯ Both

   ◯ Neither

2. (1 point) **Select one:** If $f(n) = n^{10}$ and $g(n) = 10^n$ which of the following are true?

   ◯ $f(n) \in O(g(n))$

   ◯ $g(n) \in O(f(n))$

   ◯ Both

   ◯ Neither

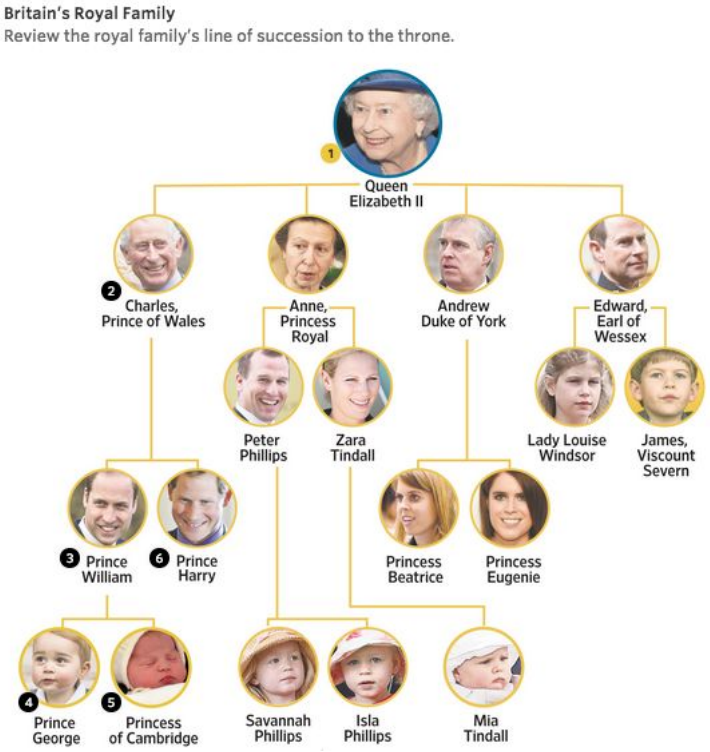

Figure 1: Britain's Royal Family

3. (2 points) **Select one:** Using the tree shown in Figure 1, how many nodes would depth-first-search visit in finding Mia Tindall (including her node)? Assuming we search left-to-right and top-down.

   ◯ 3

   ◯ 12

   ◯ 15

   ◯ 18

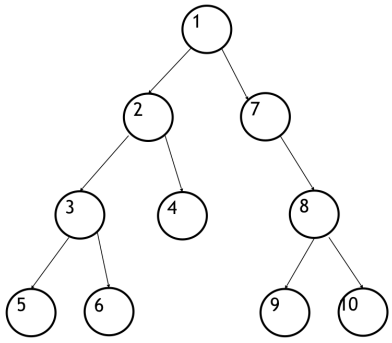

Figure 2: A Binary Tree with indexed nodes

4. (2 points) Figure 2 is a Binary Tree with indexed nodes. Assume root node is node 1. What is the node-visit order of **DFS** and **BFS** of the above Binary Tree?

   A depth-first search (DFS) traversal of a binary tree starts with visiting the root node, and recursively searches down the left subtree (i.e., the tree rooted at the left node) before going to search the right subtree (i.e., the tree rooted at the right node) until the traversal is done.
   Note: Alternatively, we can also look right subtree before left subtree too, for the question please consider left to right order!

   A breadth-first search (BFS) traversal of a binary tree visits every node (assuming a left-to-right order) on a level (with the same distance to the root) before going to a lower level until the traversal is done.

   The node-visit order of DFS is:

   Your Answer

   The node-visit order of BFS is:

   Your Answer

5. (2 points) Fill in the blanks in the pseudo code for key search using recursive depth-first search (DFS) traversal. *(Note: Please put your answer in the boxes below, not on the lines.)*

```
class TreeNode:
    def __init__(self, val):
        self.val = val
        self.leftNode = None
        self.rightNode = None

# (a) the left/right node is denoted as
#     node.leftNode/node.rightNode
# (b) left/right node are of type TreeNode
# (c) the value of the node is denoted as node.val
# (d) recursive DFS to search for the node
#     with value key in a binary tree
# (e) the left node is assumed to be searched
#     before the right node

def find_val(node, key):
    if node is None:
        return None

    if (1):
        return node

    else:
        result = (2)

        if result is None:
            result = (3)

        return (4)
```

Your pseudo code for missing field (**1**):

Your Answer

Your pseudo code for missing field (**2**):

Your Answer

Your pseudo code for missing field **(3)**:

Your Answer

Your pseudo code for missing field **(4)**:

Your Answer

Consider writing a recursive program to solve question 6:

A series of numbers is defined as:

$$M_n = \begin{cases} 2 & \text{if } n = 0 \\ 1 & \text{if } n = 1 \\ M_{n-1} + M_{n-2} & \text{if } n > 1 \end{cases}$$

6. (2 points) **Select one:** Which of the following is the numerical value for M_{32}?

- ◯ 3010349
- ◯ 3524578
- ◯ 4870847
- ◯ 7881196

Consider the following information to answer questions 7-8:

Given the functions of computing a Fibonacci number:

```
def fib_1(n):
    if n == 0 or n == 1:
        return 1
    return fib_1(n - 1) + fib_1(n - 2)

d = {}
d[0] = 1
d[1] = 1
def fib_2(n):
    if n in d.keys():
        return d[n]
    d[n] = fib_2(n - 1) + fib_2(n - 2)
    return d[n]
```

7. (2 points) **Select one:** Which of the following is the tightest upper bound on the time complexity of computing `fib_1(n)`?

   ◯ $O(n)$

   ◯ $O(n \log n)$

   ◯ $O(2^n)$

   ◯ $O(n!)$

8. (2 points) **Select one:** Which of the following is the tightest upper bound on the time complexity of computing `fib_2(n)`?

   ◯ $O(n)$

   ◯ $O(n \log n)$

   ◯ $O(2^n)$

   ◯ $O(n!)$

8 Collaboration Questions

After you have completed all other components of this assignment, report your answers to these questions regarding the collaboration policy. Details of the policy can be found here.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details.
2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details.
3. Did you find or come across code that implements any part of this assignment? If so, include full details.

Your Answer

9 Programming: Majority Vote Classifier [30 Points]

9.1 Introduction

The goal of this assignment is to ensure that you:

1. Have a way to edit and test your code (i.e. a text editor and compiler/interpreter)
2. Are familiar with submitting to Gradescope
3. Are familiar with file I/O and standard output in Python

Warning: This handout assumes that you are using a Unix command prompt (with `zsh, bash, csh` or similar). Windows commands may differ slightly.

9.2 Majority Vote Classifier

9.2.1 Algorithm

This assignment requires you to implement a Majority Vote Classifier. Your algorithm should calculate the most common label in the data, "predict" that label for each given point in the dataset, and calculate the error rate for the classifier's predictions. You may assume that the output class label is always binary.

The training procedure should store the label used for prediction at test time. In the case of a tie, output the value that is numerically higher (or comes *last* alphabetically). At test time, each example should be passed through the classifier. Its predicted label becomes the label most commonly occurring in the train set.

Looking ahead: This simple algorithm acts as a small component of the Decision *Tree* that you will implement in the next homework assignment. We hope that you will employ best practices when coding so that you can re-use your own code here in the next assignment. A Majority Vote Classifier is simply a decision tree of depth zero (it predicts a class label for the input instance based on the most commonly occurring label present in the data).

9.2.2 The Datasets

Materials Download the zip file from course website, which contains all the data that you will need in order to complete this assignment.

Datasets The handout contains two datasets. Each one contains attributes and labels and is already split into training and testing data. The first row of each `.tsv` file contains the name of each attribute, and *the class label is always the last column*.

1. **heart:** The first task is to predict whether a patient has been (or will be) diagnosed with heart disease, based on available patient information. The attributes (aka. features) are:
   (a) `sex`: The sex of the patient—1 if the patient is male, and 0 if the patient is female.
   (b) `chest_pain`: 1 if the patient has chest pain, and 0 otherwise.
   (c) `high_blood_sugar`: 1 if the patient has high blood sugar (>120 mg/dl fasting), or 0 otherwise.
   (d) `abnormal_ecg`: 1 if exercise induced angina in the patient, and 0 otherwise. Angina is a type of severe chest pain.
   (e) `flat_ST`: 1 if the patient's ST segment (a section of an ECG) was flat during exercise, or 0 if it had some slope.

(f) `fluoroscopy`: 1 if a physician used fluoroscopy, and 0 otherwise. Fluoroscopy is an imaging technique used to see the flow of blood through the heart.

(g) `thalassemia`: 1 if the patient is known to have thalassemia, and 0 otherwise. Thalassemia is a blood disorder that may impair the oxygen-carrying capacity of the patient's red blood cells.

(h) `heart_disease`: 1 if the patient was diagnosed with heart disease, and 0 otherwise. This is the class label you should predict.

The training data is in `heart_train.tsv`, and the test data in `heart_test.tsv`.

2. **education:** The second task is to predict the final grade for high school students. The attributes are student grades on 5 multiple choice assignments *M1* through *M5*, 4 programming assignments *P1* through *P4*, and the final exam *F*. Values of 1 indicate that a student received an A, and 0 indicates that the student did not receive an A. The training data is in `education_train.tsv`, and the test data in `education_test.tsv`.

The handout zip file also contains the predictions and metrics from a reference implementation of a Majority Vote Classifier for the **heart** and **education** datasets (see subfolder *example_output*). You can check your own output against these to see if your implementation is correct.[1]

Note: For simplicity, all attributes are discretized into just two categories. This applies to all the datasets in the handout, as well as the additional datasets on which we will evaluate your Majority Vote Classifier.

9.2.3 Command Line Arguments

The autograder runs and evaluates the output from the files generated, using the following command:

```
$ python majority_vote.py [args...]
```

Where above `[args...]` is a placeholder for five command-line arguments: `<train input> <test input> <train out> <test out> <metrics out>`. These arguments are described in detail below:

1. `<train input>`: path to the training input `.tsv` file
2. `<test input>`: path to the test input `.tsv` file
3. `<train out>`: path of output `.txt` file to which the predictions on the *training* data should be written
4. `<test out>`: path of output `.txt` file to which the predictions on the *test* data should be written
5. `<metrics out>`: path of the output `.txt` file to which metrics such as train and test error should be written

As an example, the following command line would run your program on the heart dataset. The train predictions would be written to `heart_train_labels.txt`, the test predictions to `heart_test_labels.txt`, and the metrics to `heart_metrics.txt`.

```
$ python majority_vote.py heart_train.tsv heart_test.tsv \
    heart_train_labels.txt heart_test_labels.txt heart_metrics.txt
```

[1] Yes, you read that correctly: we are giving you the correct answers.

9.2.4 Output: Labels Files

Your program should write two output `.txt` files containing the predictions of your model on training data (`<train out>`) and test data (`<test out>`). Each should contain the predicted labels for each example printed on a new line. Use '\n' to create a new line.

Your labels should exactly match those of a reference majority vote classifier implementation—this will be checked by the autograder by running your program and evaluating your output file against the reference solution.

The first few lines of an example output file is given below for the heart dataset:

```
0
0
0
0
0
0
...
```

9.2.5 Output: Metrics File

Generate another file where you should report the training error and testing error. This file should be written to the path specified by the command line argument `<metrics out>`. Your reported numbers should be within 0.0001 of the reference solution. You do not need to round your reported numbers! The autograder will automatically incorporate the right tolerance for float comparisons. The file should be formatted as follows:

```
error(train): 0.490000
error(test): 0.402062
```

9.3 Command Line Arguments

In this and future programming assignments, we will use command line arguments to run your programs with different parameters. Below, we provide some simple examples for how to do this in Python. In the examples below, suppose your program takes two arguments: an input file and an output file.

Python:

```
import sys

if __name__ == '__main__':
    infile = sys.argv[1]
    outfile = sys.argv[2]
    print("The input file is: %s" % (infile))
    print("The output file is: %s" % (outfile))
```

9.4 Code Submission

You must submit a file named `majority_vote.py`. The autograder is case sensitive. You must submit this file to the corresponding homework link on Gradescope.

Note: For this assignment, you may make arbitrarily many submissions to the autograder before the deadline, but only your last submission will be graded.