HOMEWORK 1: BACKGROUND *

10-301 / 10-601 INTRODUCTION TO MACHINE LEARNING (SPRING 2026)
http://www.cs.cmu.edu/~mgormley/courses/10601/

OUT: Monday, January 12th
DUE: Wednesday, January 21st
TAs: Alan, Changwook, Doris, Soham, Zachary, Neural the Narwhal

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://mlcourse.org/index.html#7-collaboration-and-academic-integrity-policies
- **Late Submission Policy:** See the late submission policy here: http://mlcourse.org/index.html#6-general-policies
- **Submitting your work:** You will use Gradescope to submit answers to all questions and code. Please follow instructions at the end of this PDF to correctly submit all your code to Gradescope.
  - **Written:** For written problems such as short answer, multiple choice, derivations, proofs, or plots, 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. Each derivation/proof should be completed in the boxes provided. You are responsible for ensuring that your submission contains exactly the same number of pages and the same alignment as our PDF template. If you do not follow the template, your assignment may not be graded correctly by our AI assisted grader and there will be a **2% penalty** (e.g., if the homework is out of 100 points, 2 points will be deducted from your final score).

    For this assignment only, if you answer at least 90% of the written questions correctly, you get full marks on the written portion of this assignment. For this assignment only, **we will offer two rounds of grading**. The first round of grading will happen immediately following the due date specified above. We will then release your grades to you and if you got less than 90% on the written questions, you will be allowed to submit once again by a second due date. The exact due date for the second round will be announced after we release the first round grades.
  - **Programming:** You will submit your code for programming questions on the homework to Gradescope. After uploading your code, our grading scripts will autograde your assignment by running your program on a virtual machine (VM). You are only permitted to use the Python Standard Library modules and `numpy`.

    Ensure that the version number of your programming language environment (i.e. Python 3.12.*) and versions of permitted libraries (i.e. `numpy` 2.2.4) match those used on Gradescope. You have 10 free Gradescope programming submissions, after which you will begin to lose points

*Compiled on 2026-01-12 at 17:22:11

from your total programming score. We recommend debugging your implementation on your local machine (or the Linux servers) and making sure your code is running correctly first before submitting your code to Gradescope.

- **Materials:** The data and reference output that you will need in order to complete this assignment is posted along with the writeup and template on the course website.

Instructions for Specific Problem Types

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

Select One: Who taught this course?

- ● Matt Gormley
- ○ Henry Chai
- ○ 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?

- ● Matt Gormley
- ○ Henry Chai
- ⊗ Noam Chomsky

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

Select all that apply: Which are instructors for this course?

- ■ Matt Gormley
- ■ Pat Virtue
- □ Henry Chai
- □ 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 the instructors for this course?

- ■ Matt Gormley
- ■ Pat Virtue
- ⊠ Henry Chai
- ⊠ 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-~~6~~301

Written Questions (66 points)

1 LaTeX Point and Template Alignment (1 points)

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

   ◯ Yes

   ◯ No

2. (0 points) **Select one:** I have ensured that my final submission is aligned with the original template given to me in the handout file and that I haven't deleted or resized any items or made any other modifications which will result in a misaligned template. I understand that incorrectly responding yes to this question will result in a penalty equivalent to 2% of the points on this assignment.
   Note: Failing to answer this question will not exempt you from the 2% misalignment penalty.

   ◯ Yes

2 Course Policies (10 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:** Assignment 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 more than 3 days?

   ◯ As many as I want; Of course!

   ◯ 6; No

   ◯ 6; Yes

   ◯ 4; No

   ◯ 4; Yes

3. (1 point) **Select one:** You may use grace days on HW1.

   ◯ True

   ◯ False

4. (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

5. (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

6. (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

7. What is (are) the consequence(s) of being caught cheating in this course?

   (a) (1 point) **Select all that apply:** First instance of cheating:

   - □ Failure of the course
   - □ AIV report to university authorities
   - □ Negative 100% on the assignment
   - □ None of the above

   (b) (1 point) **Select all that apply:** Second instance of cheating:

   - □ Failure of the course
   - □ AIV report to university authorities
   - □ Negative 100% on the assignment
   - □ None of the above

8. (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

9. (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 Piazza at least 4 days before the homework deadline

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

3 Probability and Statistics (13 points)

1. (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

2. (1 point) **Select one:** For events A_1, A_2, A_3,

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

   ◯ True

   ◯ False

3. (1 point) **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:

$$P(R) = 0.4$$
$$P(W|R) = 0.8$$
$$P(W|\neg R) = 0.2$$

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

   ◯ 0.64

   ◯ 0.56

   ◯ 0.44

   ◯ 0.4

4. 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

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

   ◯ $\frac{2}{3}$

   ◯ $\frac{3}{7}$

   ◯ $\frac{4}{5}$

   ◯ $\frac{3}{5}$

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

- ◯ 0.2
- ◯ 0.6
- ◯ 0.5
- ◯ 0.3

5. Let $\mathcal{D}$ be a distribution with mean 1 and variance 2. Now, consider 16 independently identically distributed random variables $X_1, X_2, \cdots, X_{16} \sim \mathcal{D}$. Let $Y = \frac{1}{16}\sum_{i=1}^{16} X_i$.

(a) (1 point) **Select one:** What is $\mathbb{E}[Y]$?

- ◯ 16
- ◯ 1
- ◯ $\frac{1}{16}$
- ◯ 8

(b) (1 point) **Select one:** What is $Var[Y]$?

- ◯ $\frac{1}{8}$
- ◯ 2
- ◯ 32
- ◯ $\frac{1}{16}$

6. 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

(a) (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.

- ☐ $A \perp\!\!\!\perp B$
- ☐ $B \perp\!\!\!\perp C$
- ☐ $C \perp\!\!\!\perp B$
- ☐ None of the above.

(b) (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

7. (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 $\mathbb{E}[Y]$?

   ○ $\sum_{y=1}^{n} y \frac{1}{2^y}$

   ○ $\sum_{y=1}^{n} y \frac{5}{3^y}$

   ○ $\sum_{y=1}^{n} \frac{y}{n}$

   ○ $\sum_{y=1}^{n} y$

8. (1 point) Please match the probability density function of the random variable X to its corresponding distribution name.

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

   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	Exponential	Uniform	Bernoulli	Binomial

4 Linear Algebra (9 points)

1. (1 point) **Select one:** The matrix $\mathbf{A} = \begin{bmatrix} 2 & 1 & 4 \\ -3 & 2 & 0 \\ 1 & 3 & -2 \end{bmatrix}$ has an inverse.

   ◯ True

   ◯ False

2. (2 points) **Select all that apply:** If $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$, then which of the following are true?

   □ A is invertible

   □ A has rank 2

   □ A has determinant 0

   □ A has trace 5

   □ None of the above

3. (2 points) **Select one:** Consider two vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$. Let $z = \mathbf{x}^T\mathbf{y}$. What is $\frac{\partial z}{\partial y_2}$?

   ◯ y_2

   ◯ x_2

   ◯ $\mathbf{x}$

   ◯ $\mathbf{y}$

4. (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 $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a nonzero vector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ where we call the scalar λ the associated eigenvalue for $\mathbf{v}$.)

   ◯ -5

   ◯ -3

   ◯ 2

   ◯ 1.5

5. (2 points) **Select one:** While preparing for her linear algebra final, Ada is finding eigenvectors and eigenvalues for different matrices. Ada finds out that some matrix $\mathbf{A}$ (not given) has eigenvalues -3 and 2. She also notices that both $\begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} -3 \\ 0 \\ 6 \end{bmatrix}$ are solutions to the equation $\mathbf{Av} = 2\mathbf{v}$, and concludes that both of these vectors are eigenvectors corresponding to the eigenvalue of 2. Which statement regarding her conclusion is true?

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

   ◯ It must be wrong because the eigenvectors are linearly dependent and thus cannot both be solutions to $\mathbf{Av} = 2\mathbf{v}$.

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

   ◯ It is correct because both vectors are solutions to $\mathbf{Av} = 2\mathbf{v}$.

5 Calculus (8 points)

1. (2 points) Evaluate the derivative of y with respect to x, where $y = \ln(\frac{4}{x^3} - x^2)$ 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 $g(\boldsymbol{\theta}) = \left(f\left(\mathbf{x}^{(1)}|\boldsymbol{\theta}\right) - y^{(1)}\right)^2$ for some $\mathbf{x}^{(1)} \in \mathbb{R}^n$ and $y^{(1)} \in \mathbb{R}$. What is the function type of $g(\boldsymbol{\theta})$:

   ◯ $g : \mathbb{R}^n \to \mathbb{R}$

   ◯ $g : \mathbb{R} \to \mathbb{R}$

   ◯ $g : \mathbb{R} \to \mathbb{R}^n$

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

6 Geometry (8 points)

1. Consider the vector $\mathbf{w}$ and the line $\mathbf{w^T x} + b = 0$. Assume $\mathbf{x}$ and $\mathbf{w}$ are both two-dimensional column vectors and that $\mathbf{w^T}$ indicates the transpose of the column vector $\mathbf{w}$.

   (a) (2 points) **Select one:** Consider any two points $\mathbf{x}'$ and $\mathbf{x}''$ on the line $\mathbf{w^T x} + b = 0$. Select the correct statement describing the relationship between $\mathbf{w^T}$ and $\mathbf{x}' - \mathbf{x}''$.

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$ is 0

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$ is 1

   ◯ The inner product $\mathbf{w^T}(\mathbf{x}' - \mathbf{x}'')$ is b

   (b) (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

2. (2 points) **Select one:** Let $\mathbf{w}, \mathbf{x}$ be d-dimensional vectors. What is the distance from the origin to the line $\mathbf{w^T x} + b = 0$?

   ◯ $\frac{|b|}{||\mathbf{w}||}$

   ◯ $\frac{d|b|}{||\mathbf{w}||}$

   ◯ $\frac{||\mathbf{w}||}{|b|}$

   ◯ $\frac{||\mathbf{w}||}{d|b|}$

3. (2 points) **Select one:** Suppose we have a vector $\mathbf{a}$ and a vector $\mathbf{b}$. Which of the following is the projection of $\mathbf{a}$ onto $\mathbf{b}$?

   ◯ $(\mathbf{a}^T\mathbf{b})\mathbf{b}$ ◯ $\frac{1}{||\mathbf{b}||^2}\mathbf{b}$ ◯ $\frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{b}||}\mathbf{b}$ ◯ $\frac{\mathbf{a}^T\mathbf{b}}{||\mathbf{b}||^2}\mathbf{b}$ ◯ $\frac{\cos(\mathbf{a}^T\mathbf{b})}{||\mathbf{b}||^2}\mathbf{b}$

7 CS Foundations (17 points)

1. **Big-O Notation:** For each pair (f, g) of functions below, select which of the following are true:

   (a) (1 point) **Select one:** $f(n) = \frac{n}{10}, g(n) = log_{10}(n)$

   ◯ $f(n) \in O(g(n))$ ◯ $g(n) \in O(f(n))$ ◯ Both ◯ Neither

   (b) (1 point) **Select one:** $f(n) = n^{50}, g(n) = 50^n$

   ◯ $f(n) \in O(g(n))$ ◯ $g(n) \in O(f(n))$ ◯ Both ◯ Neither

2. Consider the following functions for computing the nth Fibonacci number.

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

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

   (a) (1 point) **Select one:** Which of the following is the tightest upper bound on the time complexity of computing `fib_1(n)`?

   ◯ $O(n)$

   ◯ $O(n \log n)$

   ◯ $O(2^n)$

   ◯ $O(n!)$

(b) (1 point) **Select one:** Which of the following is the tightest upper bound on the time complexity of computing `fib_2(n)`?

- ◯ $O(n)$
- ◯ $O(n \log n)$
- ◯ $O(2^n)$
- ◯ $O(n!)$

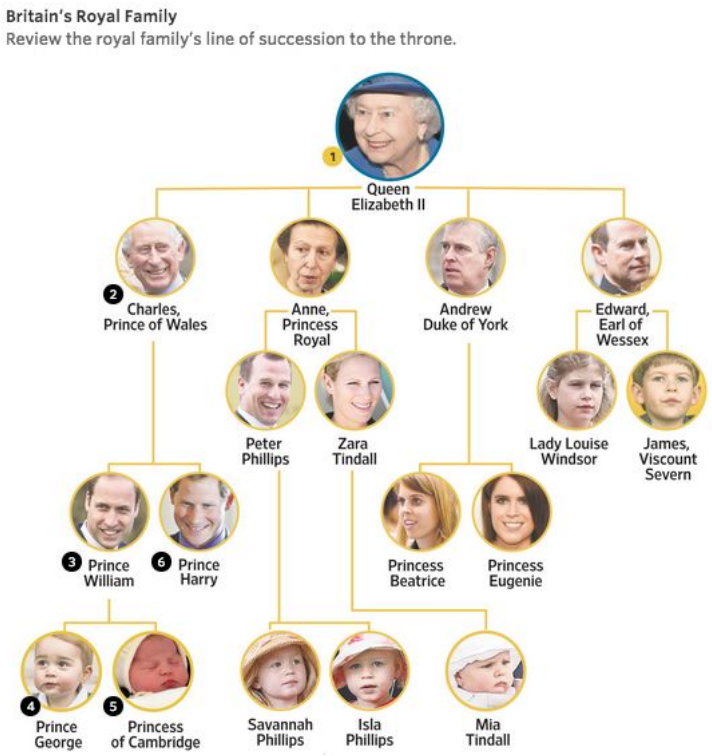

Figure 1: Britain's Royal Family

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

- ◯ 3
- ◯ 12
- ◯ 15
- ◯ 18

4. Figure 2 is a Binary Tree with indexed nodes. Assume root node is node 1.

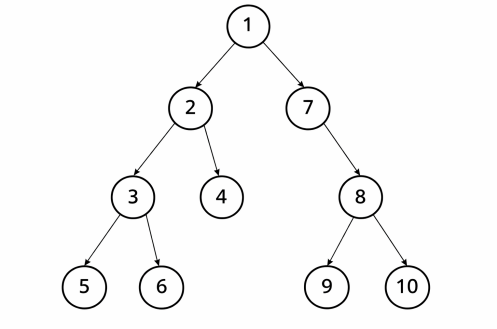

Figure 2: A Binary Tree with indexed nodes

(a) (1 point) What is the node-visit order of **DFS** 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!

The node-visit order of DFS is (separate values with commas, e.g., 1,2,3,4)

(b) (1 point) What is the node-visit order of **BFS** of the above Binary Tree?

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 BFS is (separate values with commas, e.g., 1,2,3,4)

5. (4 points) Fill in the blanks in the Python 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, key):
        self.key = key
        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 key of the node is denoted as node.key
# (d) the left node is 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)__
```

Python code for missing field **(1)**

Python code for missing field **(2)**

Python code for missing field **(3)**

Python code for missing field **(4)**

6. Consider an $M \times N$ grid where each cell has a cost associated with passing through it. From any given cell, you can either travel to the right (same row, next column) or down (same column, next row). You need to return the minimum possible cost of traveling from the start point $[0, 0]$ in the top left of the grid to the end point $[M-1, N-1]$ in the bottom right of the grid. In Figure 3, the minimum cost path is highlighted in gold and the cost is 24. Assume the cost of traveling outside the grid is infinite. You can use `float('inf')` to represent infinity.

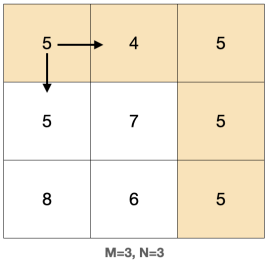

Figure 3: An example grid containing the costs for each cell and the directions one can travel from a cell.

We will first attempt to solve this using recursion.

```
# grid is of type List[List[int]]
m, n = len(grid), len(grid[0])


def minCostPath(r, c):
    if __(1)__:
            return __(2)__

    if r == m-1 and c == n-1:
        return grid[r][c]

    return __(3)__

minCostPath(0, 0)
```

(a) (3 points) Fill in the blanks to complete the above function.

Python code for missing field **(1)**

Python code for missing field **(2)**

Python code for missing field **(3)**

(b) (3 points) From Figure 4, we can deduce that using recursion is inefficient because we recalculate the costs of different partial routes from the same intermediate cells to the endpoint.

The following figure demonstrates the inefficiency of using recursion to find the minimum cost path through the grid because it recalculates the cost of partial routes. For example, we can see that the red and black paths share the partial route from $[1, 2]$ to $[2, 2]$. The cost of this partial route is calculated once for the red path and once again for the black path.

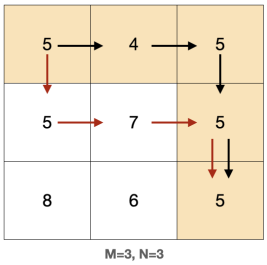

Figure 4

To solve this problem, we can utilize memoization by creating a dictionary `cache` with each intermediate `r, c` state and its corresponding value of the `minCostPath` function.

Fill in the blanks below. **Assume fields (1) and (2) use your same solutions from above**.

```
# grid is of type List[List[int]]
m, n = len(grid), len(grid[0])
cache = {}

def minCostPath(r, c):
        if __(1)__:
            return __(2)__

        if r == m-1 and c == n-1:
            return grid[r][c]

        key = (__(4)__,__(5)__)
        if key not in cache:
            cache[key] = __(6)__
        return cache[key]

minCostPath(0, 0)
```

Your Python code for missing field **(4)**

Your Python code for missing field **(5)**

Your Python code for missing field **(6)**

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 Setting Up Your Conda Environment

Use conda to setup your python environment. While this homework has very simple dependencies, it will be difficult to work with future assignments' python and package dependencies without it. So we've set up a guide for installation and setup **here.** The guide goes over installing Miniconda on your machine and setting up a conda environment for this class.

Future homeworks will assume that you have a working conda environment.

9.3 Majority Vote Classifier

9.3.1 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- ie. in a tie between Apples and Bananas, you would return Bananas). 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).

9.3.2 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.

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

9.3.3 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
```

9.3.4 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
...
```

9.3.5 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
```

9.4 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(f"The input file is: {infile}")
    print(f"The output file is: {outfile}")
```

9.5 Code Submission

You must submit a single file named `majority_vote.py`. **Any other files will be deleted.** 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.