

HOMWORK 6: LEARNING THEORY, MLE/MAP, FAIRNESS METRICS, AND SOCIETAL IMPACT *

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

OUT: Monday, October 27th
DUE: Saturday, November 1st
TAs: Alan, Jayden, Julie, Orelia, Neural the Narwhal

Summary Homework 6 covers topics on Learning Theory, MLE/MAP, Probabilistic Learning, Fairness Metrics, and Societal Impacts.

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.
 - **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 $\text{\LaTeX}$. 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).

*Compiled on 2026-03-15 at 11:47:00

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 (72 points)

1 $\text{\LaTeX}$ Point and Template Alignment (1 points)

1.1. (1 point) **Select one:** Did you use $\text{\LaTeX}$ for the entire written portion of this homework?

☐ Yes

☐ No

1.2. (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

1.3. (0 points) **Select one:** Did you fill out the [Exit Poll](#) for the previous HW? Completing the exit poll will count towards your participation grade.

☐ Yes

2 Learning Theory (15 points)

- 2.1. Neural the Narwhal is given a classification task to solve, and he decides to use a decision tree learner with 2 binary features X_1 and X_2 . On the other hand, you think that Neural should not have used a decision tree. Instead, you think he would get a better bound on generalization error by using logistic regression with 16 real-valued features plus a bias term. You want to use the framework of PAC learning to reason about which model is better.

You first train your logistic regression model on N examples to obtain a training error $\hat{R}$.

- 2.1.a. (1 point) Which of the following cases of PAC learning should you use for your logistic regression model?

- ☐ Finite and realizable
- ☐ Finite and agnostic
- ☐ Infinite and realizable
- ☐ Infinite and agnostic

- 2.1.b. (2 points) What is the upper bound on the true error R in terms of $\hat{R}$, δ , $\text{VC}(\mathcal{H})$ and N ? You may use big- $\mathcal{O}$ notation if necessary. Write only the final answer. Your work will *not* be graded.

Note: Your answer must be in terms of only the quantities listed above. Use the formulas given in the recitation handout to solve this question.

Your Answer

- 2.1.c. (1 point) What is the value of the VC dimension in part (b)? Provide a single numerical value.

Your Answer

2.1.d. (2 points) **Select one:** You want to argue your method has a better (smaller) bound on the true error as compared to the Neural's true error bound. Assume that you have obtained enough data points to satisfy the PAC criterion with the same ϵ and δ as Neural. Which of the following is true?

- ☐ Neural's model will always classify unseen data more accurately because it only needs 2 binary features and therefore is simpler.
- ☐ You must first regularize your model by removing 14 features to make any comparison at all.
- ☐ It is sufficient to show that the VC dimension of your classifier is higher than that of Neural's, therefore having a better bound for the true error.
- ☐ You must show that your model's total true error bound (combining training error and VC complexity) is smaller than Neural's.

2.2. (3 points) Consider the hypothesis space of functions that map M binary attributes to a binary label. A function f in this space can be characterized as $f : \{0, 1\}^M \rightarrow \{0, 1\}$. Neural the Narwhal says that regardless of the value of M , a hypothesis class containing all possible functions in this space can always shatter 2^M points. Is Neural wrong? If so, provide a counterexample. If Neural is right, briefly explain why in 1–2 *concise* sentences.

Hint: Decision tree-like structures are one way to help understand the possible hypothesis space.

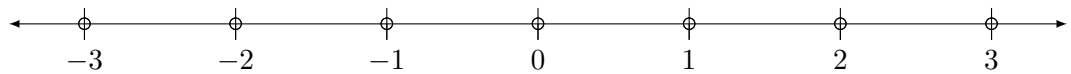
Your Answer

2.3. Consider an instance space $\mathcal{X}$ which is the set of real numbers.

2.3.a. (3 points) **Select one:** What is the VC dimension of hypothesis class H , where each hypothesis h in H is of the form “if $a < x < b$ or $c < x < d$ then $y = 1$; otherwise $y = 0$ ”? (i.e., H is an infinite hypothesis class where a, b, c , and d are arbitrary real numbers).

- ☐ 2
☐ 3
☐ 4
☐ 5
☐ 6

2.3.b. (3 points) Given the set of points in $\mathcal{X}$ below, construct a labeling of some subset of the points to show that any dimension larger than the VC dimension of H by *exactly* 1 is incorrect (e.g. if the VC dimension of H is 3, only fill in the answers for 4 of the points). Fill in the boxes such that for each point in your example, the corresponding label is either 0 or 1. For points you are not using in your example, write N/A (do *not* leave the answer box blank).



Answer for -3	Answer for -2	Answer for -1	
Answer for 0	Answer for 1	Answer for 2	Answer for 3

3 MLE/MAP (8 points)

- 3.1. (1 point) **True or False:** Recall that the MLE of a parameter θ given some dataset $\mathcal{D}$ is

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} p(\mathcal{D} \mid \theta)$$

while the MAP estimate of θ is

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\mathcal{D} \mid \theta)p(\theta)$$

For every dataset, assuming that the parameter space is bounded, there will always exist some prior distribution $p(\theta)$ for which the MAP estimate is equivalent to the MLE.

☐ True

☐ False

- 3.2. (1 point) **True or False:** Suppose you place a Beta prior over the parameter θ of the Bernoulli distribution, and attempt to learn θ from data. Further suppose an adversary chooses “bad” but finite hyperparameters for your Beta prior in order to confuse your learning algorithm. As the number of training examples grows to infinity, the MAP estimate of θ can still converge to the MLE estimate of θ .

☐ True

☐ False

- 3.3. (2 points) **Select one:** Let Γ be a random variable with the following probability density function (pdf):

$$f(\gamma) = \begin{cases} 2\gamma & \text{if } 0 \leq \gamma \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose another random variable Y , which is conditioning on Γ , follows an exponential distribution with $\lambda = 3\gamma$. Recall that the exponential distribution with parameter λ has the following pdf:

$$f_{\text{exp}}(y) = \begin{cases} \lambda e^{-\lambda y} & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the MAP estimate of γ given $Y = \frac{2}{3}$ is observed?

Your Answer

- 3.4. (4 points) Neural the Narwhal found a mystery coin and wants to know the probability of landing on heads by flipping this coin. He models the coin toss as sampling a value from $\text{Bernoulli}(\theta)$ where θ is the probability of heads. He flips the coin three times and the flips turned out to be heads, tails, and heads. An oracle tells him that $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$, and *no other values of θ should be considered*.

Find the MLE and MAP estimates of θ . Use the following prior distribution for the MAP estimate:

$$p(\theta) = \begin{cases} 0.9 & \text{if } \theta = 0 \\ 0.04 & \text{if } \theta = 0.25 \\ 0.03 & \text{if } \theta = 0.5 \\ 0.02 & \text{if } \theta = 0.75 \\ 0.01 & \text{if } \theta = 1 \end{cases}.$$

Again, remember that $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$, so the MLE and MAP should also be one of them.

MLE of θ	MAP of θ

4 Probabilistic Learning (17 points)

- 4.1. In a previous homework assignment, you have derived the closed form solution for linear regression. Now, we are coming back to linear regression, viewing it as a statistical model, and deriving the MLE and MAP estimate of the parameters in the following questions.

As a reminder, in MLE, we have

$$\hat{\theta}_{MLE} = \operatorname{argmax}_{\theta} p(D|\theta)$$

For MAP, we have

$$\hat{\theta}_{MAP} = \operatorname{argmax}_{\theta} p(\theta|D)$$

Assume we have data $D = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^N$, where $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_M^{(i)})$. So our data has N instances and each instance has M features. Each $y^{(i)}$ is generated given $\mathbf{x}^{(i)}$ with additive noise $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$, where σ^2 is fixed: that is, $y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)}$ where $\mathbf{w}$ is the parameter vector of linear regression.

- 4.1.a. (2 points) **Select one:** Given this assumption, what is the distribution of $y^{(i)}$?

- ☐ $y^{(i)} \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}^{(i)}, \sigma^2)$
- ☐ $y^{(i)} \sim \mathcal{N}(0, \sigma^2)$
- ☐ $y^{(i)} \sim \text{Uniform}(\mathbf{w}^T \mathbf{x}^{(i)} - \sigma, \mathbf{w}^T \mathbf{x}^{(i)} + \sigma)$
- ☐ None of the above

- 4.1.b. (2 points) **Select one:** The next step is to learn the MLE of the parameters of the linear regression model. Which expression below is the correct conditional log likelihood $\ell(\mathbf{w})$ with the given data?

- ☐ $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐ $\sum_{i=1}^N [\log(\sqrt{2\pi\sigma^2}) + \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐ $\sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐ $-\log(\sqrt{2\pi\sigma^2}) + \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

- 4.1.c. (2 points) **Select all that apply:** Then, the MLE of the parameters is just $\operatorname{argmax}_{\mathbf{w}} \ell(\mathbf{w})$. Among the following expressions, select ALL that can yield the correct MLE.

- ☐ $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐ $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐ $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$
- ☐ $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})]$
- ☐ $\operatorname{argmax}_{\mathbf{w}} \sum_{i=1}^N [-\frac{1}{2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2]$

☐ None of the above

4.1.d. (1 point) **Select one:** Suppose we fix $\sigma = 3.7$ when deriving the MLE of $\mathbf{w}$. How would the optimal $\mathbf{w}_{MLE}$ change if we instead fixed $\sigma = 5.1$? Briefly justify your answer.

- ☐ $\mathbf{w}_{MLE}$ would change because the variance affects the optimization scale.
- ☐ $\mathbf{w}_{MLE}$ would not change, since σ^2 only scales the loss and does not affect its minimizer.
- ☐ $\mathbf{w}_{MLE}$ would change only if σ is smaller than 1.
- ☐ None of the above.

4.2. Now we are moving on to learn the MAP estimate of the parameters of the linear regression model. Consider the same data D we used for the previous problem.

4.2.a. (2 points) **Select all that apply:** Which expression below is the correct optimization problem the MAP estimate is trying to solving? Recall that D refers to the data, and $\mathbf{w}$ to the regression parameters (weights).

- ☐ $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(D, \mathbf{w})$
- ☐ $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)}$
- ☐ $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} \frac{p(D, \mathbf{w})}{p(\mathbf{w})}$
- ☐ $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(D|\mathbf{w})p(\mathbf{w})$
- ☐ $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} p(\mathbf{w}|D)$
- ☐ None of the above

4.2.b. (2 points) **Select one:** Suppose we are using a Gaussian prior distribution with mean 0 and variance $\frac{1}{\lambda}$ for each element w_m of the parameter vector $\mathbf{w}$, i.e. $w_m \sim \mathcal{N}(0, \frac{1}{\lambda})$ ($1 \leq m \leq M$). Assume that $w_1, \dots, w_M$ are mutually independent of each other. Which expression below is the correct log joint-probability of the data and parameters $\log p(D, \mathbf{w})$?

- ☐ $\sum_{i=1}^N \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \right) - \sum_{m=1}^M \log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- ☐ $\sum_{i=1}^N \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}) \right) + \sum_{m=1}^M -\log(\sqrt{2\pi\lambda}) - \lambda(w_m)^2$
- ☐ $\sum_{i=1}^N \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)}) \right) - \sum_{m=1}^M \log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$
- ☐ $\sum_{i=1}^N \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 \right) + \sum_{m=1}^M -\log(\sqrt{\frac{2\pi}{\lambda}}) - \frac{\lambda}{2}(w_m)^2$

4.2.c. (2 points) **Select one:** For the same linear regression model with a Gaussian prior on the parameters as in the previous question, maximizing the log posterior probability $\ell_{MAP}(\mathbf{w})$ gives you the MAP estimate of the parameters. Which of the following is an equivalent definition of $\max_{\mathbf{w}} \ell_{MAP}(\mathbf{w})$?

- ☐ $\max_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$
- ☐ $\min_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$
- ☐ $\max_{\mathbf{w}} \sum_{i=1}^N \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 + \lambda\|\mathbf{w}\|_2^2$
- ☐ $\min_{\mathbf{w}} - \sum_{i=1}^N \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T \mathbf{x}^{(i)})^2 - \frac{\lambda}{2}\|\mathbf{w}\|_2^2$

4.2.d. (2 points) **Select one:** You found a MAP estimator that has a much higher test error than train error using some Gaussian prior. Identify the issue and a possible approach to fixing this.

- ☐ Overfitting; Increase the variance of the prior used
- ☐ Overfitting; Decrease the variance of the prior used
- ☐ Underfitting; Increase the variance of the prior used
- ☐ Underfitting; Decrease the variance of the prior used

4.3. (2 points) **Select one:** MAP estimation with what prior is equivalent to ℓ_1 regularization?

Note:

- The pdf of a uniform distribution over $[a, b]$ is $f(x) = \frac{1}{b-a}$ if $x \in [a, b]$ and 0 otherwise.
 - The pdf of an exponential distribution with rate parameter a is $f(x) = a \exp(-ax)$ for $x > 0$.
 - The pdf of a Laplace distribution with location parameter a and scale parameter b is $f(x) = \frac{1}{2b} \exp\left(\frac{-|x-a|}{b}\right)$ for all $x \in \mathbb{R}$.
- ☐ Uniform distribution over $[-1, 1]$
 - ☐ Uniform distribution over $[-\mathbf{w}^T \mathbf{x}^{(i)}, \mathbf{w}^T \mathbf{x}^{(i)}]$
 - ☐ Exponential distribution with rate parameter $a = \frac{1}{2}$
 - ☐ Exponential distribution with rate parameter $a = \mathbf{w}^T \mathbf{x}^{(i)}$
 - ☐ Laplace distribution with location parameter $a = 0$
 - ☐ Laplace distribution with location parameter $a = \mathbf{w}^T \mathbf{x}^{(i)}$

5 Fairness Metrics (12 points)

Neural works for the Bank of ML and is given the following dataset from another bank on whether or not to issue a loan to individuals. Each row in this dataset represents one individual's data, which includes their FICO credit score, their savings rate (percentage of their income that goes into their savings), and credit history in months. The data was collected in two different regions, region A and region B, as denoted in the first column. The "Label" column refers to the true label, where "1" refers to loan issued, and "0" refers to no loan issued. A csv file of this dataset could be found in the handout folder.

Region	FICO Score	Savings Rate (%)	Credit History (months)	Label
A	544.0625	28.0	21	1
A	489.0625	33.9	40	0
A	433.125	62.3	100	0
A	429.0625	56.7	203	1
A	417.8125	56.5	5	0
A	506.5625	32.7	75	1
A	400.625	60.7	216	0
A	836.875	10.7	86	1
A	471.875	36.2	92	1
A	402.8125	62.0	199	0
B	809.4285714	5.6	213	1
B	480.9375	40.2	72	1
B	505.0	31.1	20	0
B	438.4375	51.3	122	0
B	385.9375	76.2	89	0
B	505.625	34.7	39	1
B	514.0625	31.0	41	1
B	385.9375	76.2	89	0
B	446.25	44.5	51	0
B	428.75	55.6	215	1

5.1. Neural took the average value of the features (for example, the average value for the first data point is 197.69), and developed the following observation. In general, for all three features in this dataset, a high value indicates better credibility. Hence Neural trained the following decision stump on this dataset: if the average feature value is above the median (198.09), then we determine that the individual will receive the loan (prediction = 1). Otherwise, we decide that the individual will not receive the loan. **For parts (a), (b), (c) below, please round your answer to three decimal places.**

5.1.a. (1 point) Using the model that Neural proposed, what is the training error rate on the entire dataset?

Your Answer

5.1.b. (1 point) What is the training error rate for region A?

Your Answer

5.1.c. (1 point) What is the training error rate for region B?

Your Answer

5.1.d. (1 point) How many false positives were there in region A?

Your Answer

5.1.e. (1 point) How many false negatives were there in region A?

Your Answer

5.1.f. (1 point) How many false positives were there in region B?

Your Answer

5.1.g. (1 point) How many false negatives were there in region B?

Your Answer

5.25.2.a. (1 point) **True or False:** Using your responses to the previous question, we achieve statistical parity between regions A and B.

☐ True

☐ False

5.2.b. (1 point) Provide the positive rate for both regions A and B.

A Positive Rate	B Positive Rate

5.35.3.a. (1 point) **True or False:** We achieve equality of accuracy between regions A and B.

☐ True

☐ False

5.3.b. (1 point) Provide the accuracy rate for both regions A and B.

A Accuracy Rate	B Accuracy Rate

5.4. (1 point) **True or False:** We can achieve perfect independence between Regions A and B without changing the inherent properties of the dataset.

Note: This part refers to an arbitrary classifier, not the classifier used earlier.

☐ True

☐ False

6 Societal Impacts and Unintended Consequences (13 points)

Consider a large university's shared computing cluster, which has a fixed pool of 1,000 GPUs. The cluster administrators have a primary goal of maximizing job throughput (total jobs completed per day) to best serve the research community. As a secondary goal, the university is conscious of its carbon footprint and would like to reduce the cluster's total daily energy consumption.

- 6.1. (2 points) The cluster is currently managed by a scheduler with an old ML model, M_{old} . This model inefficiently predicts job runtimes, leading to low GPU utilization. Under this system, the cluster processes an average of $T_{old} = 5,000$ jobs per day. The average energy cost per job is $E_{old_job} = 2.0$ kWh.

Calculate the total daily energy consumption of the cluster, E_{old_total} , in kWh.

Your Answer

- 6.2. A team of 10-301 students develops a new ML-based scheduler, M_{new} . By using a Transformer to predict runtimes, it packs jobs with extreme efficiency. This new system dramatically improves the efficiency per job, reducing the average energy cost per job by 50% to $E_{new_job} = 1.0$ kWh.

- 6.2.a. (2 points) First, assume (for this subpart only) that the user demand for jobs remains constant, with $T_{new} = T_{old} = 5,000$ jobs per day. Calculate the new total daily energy consumption, E_{new_total} .

Your Answer

- 6.2.b. (2 points) Second, based on the scenario in this part (constant demand), explain in a sentence or two how the cluster administrators would most likely view this change, given their primary and secondary goals.

Your Answer

- 6.3. (2 points) The deployment of M_{new} has an unintended consequence. Because the scheduler is now so efficient, the average wait time for a user's job to start (the queue time) drops from 2 hours to just 5 minutes. This effectively lowers the "cost" of running computation.

Researchers and students quickly adapt their behavior. They were previously "cost-sensitive" and only ran critical experiments. Now, they run more experiments, launch large hyperparameter sweeps, and use the cluster for interactive debugging. This behavioral change induces a feedback loop: the total number of jobs submitted per day triples to $T_{feedback} = 15,000$ jobs/day.

Calculate the new total daily energy consumption, $E_{feedback_total}$, under these conditions. (Assume the efficiency per job remains $E_{new_job} = 1.0$ kWh).

Your Answer

- 6.4. This part concerns the unintended consequences from Part 3.

6.4.a. (1 point) **Select one:** How does the final total energy consumption $E_{feedback_total}$ from Part 3 compare to the original total consumption E_{old_total} from Part 1?

- ☐ Total energy consumption decreased by 50%, from 10,000 kWh to 5,000 kWh.
- ☐ Total energy consumption increased by 50%, from 10,000 kWh to 15,000 kWh.
- ☐ Total energy consumption tripled, from 5,000 kWh to 15,000 kWh.
- ☐ Total energy consumption remained unchanged.

6.4.b. (2 points) Explain (2–3 sentences) the unintended system-level consequence of deploying the "more efficient" ML model, M_{new} . Relate this to Goodhart's Law, as discussed in lecture.

Your Answer

6.4.c. (2 points) **Select all that apply:** Which of the following are distinct, viable mitigation strategies the university could implement to better align the system's outcome with the original goal of reducing total energy consumption, while still respecting the primary goal of maximizing throughput?

- ☐ Implement a non-ML policy such as a quota system, limiting the total energy a user can consume.
- ☐ Modify the ML model's objective function to jointly optimize for maximizing throughput AND minimizing total predicted energy consumption. For example, the model could learn to delay some jobs during peak-demand periods to manage the total energy load without sacrificing overall daily throughput.
- ☐ Introduce a non-ML policy of dynamic pricing, where users are "charged" (e.g., from a research budget) based on the energy their jobs consume.
- ☐ Roll back to the old, inefficient M_{old} model to increase queue times and discourage use.
- ☐ Purchase more GPUs to handle the 15,000 job/day demand, thereby reducing wait times even further.
- ☐ None of the above

7 Society, Ethics, and ML (6 points)

- 7.1. (6 points) This part of the assignment asks you to reflect on the societal and ethical dimensions of machine learning. All instructions, readings, and discussion prompts are located in a single pinned post on our class Piazza. Please share your answer in this format: @Q#_f# (Example: @114_f2).

hw6/fig/ethics_post.png

Your Answer

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