HOMEWORK 6: LEARNING THEORY, MLE/MAP, FAIRNESS METRICS, AND SOCIETAL IMPACT [*]

10-301 / 10-601 INTRODUCTION TO MACHINE LEARNING (SPRING 2025)
http://www.cs.cmu.edu/~mgormley/courses/10601/

OUT: Sunday, March 16th
DUE: Saturday, March 22nd
TAs: Mihir, Rohini, Joaquin, Zhifei, Sebastian, KeNNy

Summary Homework 6 covers topics on Learning Theory, MLE/MAP, Probabilistic Learning, Fairness Metrics, and Societal Impacts. The homework includes multiple choice, True/False, and short answer questions. There will be no consistency points in general, so please make sure to double check your answers to all parts of the questions!

START HERE: Instructions

- **Collaboration Policy**: Please read the collaboration policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **Late Submission Policy:** See the late submission policy here: http://www.cs.cmu.edu/~mgormley/courses/10601/syllabus.html
- **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 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 Sunday 14th September, 2025 at 01:13

Instructions for Specific Problem Types

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

Select One: Who taught this course?

- ● Matt Gormley
- ○ 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 instructors for this course?

- ■ Matt Gormley
- ■ Henry Chai
- □ Noam Chomsky
- □ 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
- ■ Henry Chai
- ☒ Noam Chomsky
- ☒ 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 (98 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

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 (21 points)

1. Neural the Narwhal is given a classification task to solve, which 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 it would be best to use logistic regression with 16 real-valued features in addition to a bias term. You want to use PAC learning to check whether you are correct. You first train your logistic regression model on N examples to obtain a training error $\hat{R}$.

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

   ○ Finite and realizable

   ○ Finite and agnostic

   ○ Infinite and realizable

   ● Infinite and agnostic

   (b) (2 points) What is the upper bound on the true error R in terms of $\hat{R}$, δ, 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 may not contain any other symbols.

   Your Answer

   $$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \ln\left(\frac{1}{\delta}\right)\right]}\right) \text{ for all hypotheses } h \text{ in } \mathcal{H}$$

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

   Your Answer

   17

(d) (2 points) **Select one:** You want to argue your method has a lower 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 lower bound for the true error.

● It is necessary to show that the training error you achieve is lower than the training error Neural achieves.

2. In lecture, we saw that we can use our sample complexity bounds to derive bounds on the true error for a particular algorithm. Consider the sample complexity bound for the infinite, agnostic case:

$$N = O\left(\frac{1}{\epsilon^2}\left[\text{VC}(\mathcal{H}) + \log\frac{1}{\delta}\right]\right).$$

(a) (3 points) What is the big-$\mathcal{O}$ bound of ϵ in terms of N, δ, and $\text{VC}(\mathcal{H})$?
Note: $A = \mathcal{O}(B)$ (for some value B) $\Leftrightarrow$ there exists a constant $c \in \mathbb{R}$ such that $A \leq cB$.

Your Answer

$$\epsilon = O\left(\sqrt{\frac{\text{VC}(\mathcal{H}) + \log\frac{1}{\delta}}{N}}\right)$$

(b) (3 points) Now, using the definition of ϵ (i.e. $|R(h) - \hat{R}(h)| \leq \epsilon$) and your answer to part a, prove that with probability at least $(1 - \delta)$:

$$R(h) \leq \hat{R}(h) + O\left(\sqrt{\frac{1}{N}\left[\text{VC}(\mathcal{H}) + \log\frac{1}{\delta}\right]}\right).$$

Your Answer

3. (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 $f : \{0,1\}^M \to \{0,1\}$. 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.

Your Answer

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

   (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

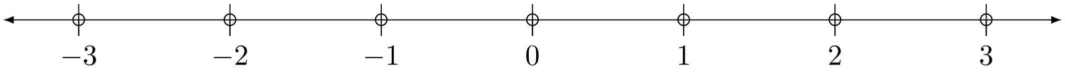
   ○ 3

   ○ 4

   ○ 5

   ○ 6

   (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)

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

2. (1 point) **True or False:** Suppose you place a Beta prior over the Bernoulli distribution, and attempt to learn the parameter θ of the Bernoulli distribution 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. (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:

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

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

Your Answer

1

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 Bernoulli(θ) 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
0.75	0.5

4 Probabilistic Learning (18 points)

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

   (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

   (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?

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

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

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

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

   (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

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.

   (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

   (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. $w_m \sim \mathcal{N}\left(0, \frac{1}{\lambda}\right)$ $(1 \leq m \leq M)$. Assume that $w_1, \cdots, w_M$ are mutually independent of each other. Which expression below is the correct log joint-probability of the data and parameters $\log p(D, \mathbf{w})$? Please show your work below.

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

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

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

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

Work

$$
\begin{aligned}
&\log p(D, \mathbf{w}) \\
&= \log p(D|\mathbf{w})p(\mathbf{w}) \\
&= \log \prod_{i=1}^{N} p(y^{(i)}, x^{(i)}|\mathbf{w}) \prod_{m=1}^{M} p(w_m) \\
&= \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)}-\mu)^2}{2\sigma^2}} \prod_{m=1}^{M} \frac{1}{\sqrt{2\pi/\lambda}} e^{-\frac{w_m^2}{2/\lambda}} \\
&= \sum_{i=1}^{N} \left(-\log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2}(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)})^2\right) + \sum_{m=1}^{M} \left(-\log\left(\sqrt{\frac{2\pi}{\lambda}}\right) - \frac{\lambda}{2} w_m^2\right)
\end{aligned}
$$

(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 $\max_{\mathbf{w}} \ell_{MAP}(\mathbf{w})$?

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

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

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

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

(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

3. (2 points) **Select one:** Suppose now the additive noise ϵ is different per datapoint. That is, each $y^{(i)}$ is generated given $\mathbf{x}^{(i)}$ with additive noise $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma_i^2)$, i.e. $y^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$. Unlike the standard regression model we have worked with until now, there is now an example specific variance σ_i^2. Maximizing the log-likelihood of this new model is equivalent to minimizing the *weighted* mean squared error with which of the following as the weights? Please show your work below.

○ $1/y^{(i)}$

● $1/\sigma_i^2$

○ $1/\|\mathbf{x}^{(i)}\|_2^2$

Work

$$\begin{aligned}
&\log p(D|\mathbf{w}) \\
&= \log \prod_{i=1}^{N} p(y^{(i)}, x^{(i)}|\mathbf{w}) \\
&= \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi {\sigma^{(i)}}^2}} e^{-\frac{(y^{(i)}-\mu)^2}{2{\sigma^{(i)}}^2}} \\
&= \sum_{i=1}^{N} \left(-\log\left(\sqrt{2\pi {\sigma^{(i)}}^2}\right) - \frac{1}{2{\sigma^{(i)}}^2}\left(y^{(i)} - \mathbf{w}^T\mathbf{x}^{(i)}\right)^2 \right)
\end{aligned}$$

4. (2 points) **Select one:** MAP estimation with what prior is equivalent to ℓ_1 regularization? Please show your work below.

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 $\left[-\mathbf{w}^T\mathbf{x}^{(i)}, \mathbf{w}^T\mathbf{x}^{(i)}\right]$

○ 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 $a = \mathbf{w}^T\mathbf{x}^{(i)}$

Work

$$\begin{aligned}
-\log p(w_m) &= -\log\left(\frac{1}{2b}\exp\left(\frac{-|w_m|}{b}\right)\right) \\
&= -\left[\log\left(\frac{1}{2b}\right) - \frac{|w_m|}{b}\right] \\
&= \log(2b) + \frac{1}{b}|w_m|
\end{aligned}$$

Summing over all M independent weights for the full parameter vector

$$\begin{aligned}
-\log p(\mathbf{w}) &= \sum_{m=1}^{M}\left(\log(2b) + \frac{1}{b}|w_m|\right) \\
&= M\log(2b) + \frac{1}{b}\sum_{m=1}^{M}|w_m| \\
&= M\log(2b) + \frac{1}{b}||\mathbf{w}||_1
\end{aligned}$$

5 Fairness Metrics (23 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 cities, city A and city 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

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**.

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

   Your Answer

   0.4

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

Your Answer

0.4

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

Your Answer

0.4

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

Your Answer

3

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

Your Answer

1

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

Your Answer

1

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

Your Answer

3

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

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

A Positive Rate	B Positive Rate
0.7	0.3

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

   ● True

   ○ False

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

A Accuracy Rate	B Accuracy Rate
0.6	0.6

4. (a) (1 point) **True or False**: We achieve equality of the ratio $\frac{FPR}{FNR}$ between regions A and B.

   ○ True

   ● False

   (b) (1 point) Provide the ratio for regions A and B.

Region A Ratio	Region B Ratio
3	0.333

5. (a) (1 point) **True or False**: We achieve equality of $\frac{PPV}{NPV}$ between regions A and B.

   ○ True

   ● False

   (b) (1 point) Provide the ratio for both regions.

Region A Ratio	Region B Ratio
0.857	1.167

6. (1 point) **True or False**: We can achieve perfect independence between Regions A and B without changing the inherent properties of the dataset.

○ True

● False

7. (1 point) **Select all that apply:** Suppose we are guaranteed that our dataset achieves equality of $\frac{FPR}{FNR}$ ratio between regions, i.e. $\frac{FPR_A}{FNR_A} = \frac{FPR_B}{FNR_B}$. Assuming only this condition is explicitly met, which of the following criteria are necessarily satisfied?

   □ Independence

   □ Separation

   □ Sufficiency

   ■ None of the above

8. Consider a scenario where Neural visits the local bank, hoping to take out a loan to visit Markov. The bank teller needs to make a binary decision about whether or not approve the loan on the bank's behalf. Recall that a Type I error occurs when you erroneously predict a positive label (false positive), and a Type II error is when you erroneously predict a negative label (false negative).

   (a) (1 point) **Select one:** From the perspective of the bank, making which type of error will have more significant consequences?

   ● Type I Error

   ○ Type II Error

   (b) (1 point) **Select one:** From the perspective of Neural, the bank making which type of error will have more significant consequences?

   ○ Type I Error

   ● Type II Error

9. Suppose you have a classifier with an imbalanced dataset where the positive class occurs much less frequently than the negative class.

   (a) (2 points) **Select one:** How does this imbalance typically affect precision and recall?

   ○ Precision decreases, recall increases

   ○ Precision increases, recall decreases

   ○ Both precision and recall increase

   ● Both precision and recall decrease

   (b) (2 points) **Select all that apply:** Which of the following strategies can help improve classification performance on this imbalanced dataset?

   ■ Using oversampling for the minority class

   ■ Using undersampling for the majority class

   ■ Adjusting the decision threshold of the classifier

   □ Ignoring the minority class to improve overall accuracy

   ■ Using evaluation metrics like F1-score instead of accuracy

6 Societal Impacts (27 points)

The fictional country, Xtopia, is in the midst of an epidemic. The Xtopian healthcare system has been under a great deal of strain in the past year due to a regional epidemic caused by an airborne virus called Xvid. The number of hospital beds is limited and as the result, healthcare professionals have to frequently make very difficult choices about which subset of Xvid patients can be hospitalized. Hospital care greatly increases the chance of recovering from the illness with no subsequent long-term health complications.

To save time and make these decisions more efficient and consistent, a team of ML practitioners have been brought in to automate the decision-making process. They have been given access to a data set consisting of the information about prior Xvid patients who sought hospital care along with the binary decision made about them by the hospital doctors ('+' indicates hospitalization and '-' indicates no hospitalization). The ML team has determined that the decision about each patient is highly correlated with his/her age as well as his/her prior utilization of medical insurance. This observation reflects the fact that Xtopian doctors are on average more likely to allocate scarce medical resources to the young and the vulnerable (i.e., those with prior medical conditions and comorbidities). Here, the insurance utilization serves as a proxy for severity of the patient's health conditions.

Figure 1 provides a snapshot of the Xvid training data and the predictive model that the ML team has come up with. Each instance corresponds to an individual patient, and each patient belongs to one of the two socially salient groups in Xtopia, indicated by blue and red.

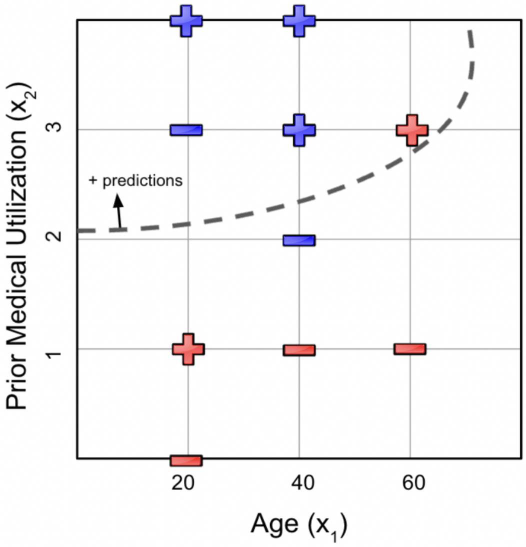


Figure 1: Xvid Training Data

Answer the following questions with respect to the above hypothetical context and data set.

1. (2 points) **Select all that apply:** Does the predictive model above satisfy the following notions of fairness across blue and red groups?

   □ False Negative Rate (FNR) parity

   □ False Positive Rate (FPR) parity

   □ Negative Predictive Value (NPV) parity

   □ Positive Predictive Value (PPV) parity

   □ Error parity

   ■ Statistical parity (or Selection rate parity)

   □ None of the above

2. (2 points) **Select all that apply:** Which of the above notions of fairness would be satisfied if the ML team could train a model with 0 true error (i.e., a model that always predicts the correct label for every patient)?

   ■ False Negative Rate (FNR) parity

   ■ False Positive Rate (FPR) parity

   ■ Negative Predictive Value (NPV) parity

   ■ Positive Predictive Value (PPV) parity

   ■ Error parity

   ■ Statistical parity (or Selection rate parity)

   □ None of the above

3. (2 points) **Select all that apply:** Which of the above notions of fairness would be satisfied in expectation by a random classifier (i.e., a model that makes a randomized prediction for every patient: with probability 0.5 the patient is hospitalized regardless of their attributes)?

   ■ False Negative Rate (FNR) parity

   ■ False Positive Rate (FPR) parity

   □ Negative Predictive Value (NPV) parity

   □ Positive Predictive Value (PPV) parity

   ■ Error parity

   ■ Statistical parity (or Selection rate parity)

   □ None of the above

4. (2 points) **Select all that apply:** From the perspective of a patient subject to the predictions made by this model, the violation of which of the parity conditions below would be most problematic?

   ■ False Negative Rate (FNR) parity. Violating FNR parity means some patients who truly need hospitalization are denied care, which can directly harm their health.

   ■ False Positive Rate (FPR) parity. Violating FPR parity means some patients may face unnecessary hospitalization.

   ■ Negative Predictive Value (NPV) parity. Violating NPV parity implies that prediction is less reliable for one group which may give false reassurance, potentially causing one to miss out on critical, timely care.

   ■ Positive Predictive Value (PPV) parity. Violating PPV parity implies the prediction is less reliable for one group, which undermines fairness in allocating scarce resources.

   □ None of the above

5. (2 points) **Causes of unfairness:** Name one potential cause of disparity in false negative rates across the two groups in the above context.

Your Answer

One potential cause of the disparity in false negative rates is that "insurance utilization" is a biased proxy for the severity of a patient's health condition, and it is less reliable for the red group than for the blue group.

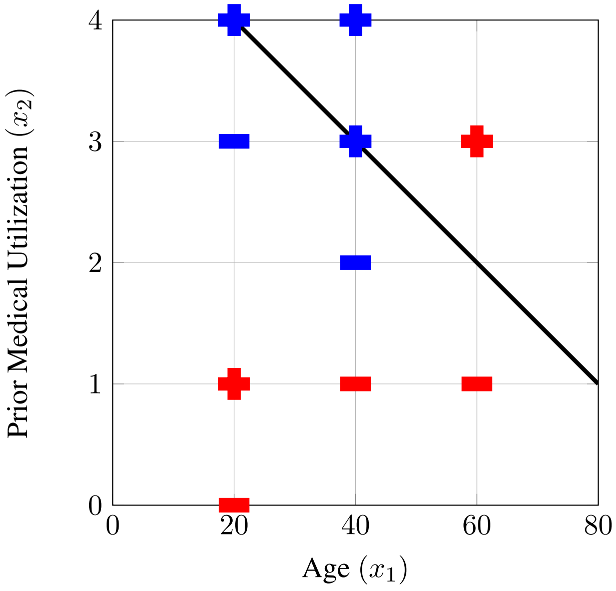
6. **Fairness interventions:** consider the following pre-processing method to improve statistical parity:

```
While the selection rate is unequal across the two groups:

(a) Pick the group with lowest selection rate.
(b) From this group in the training data, pick the data point
    closest to the decision boundary predicted as negative.
(c) Change the label of this instance to positive.
(d) Retrain the model on the modified training data by finding the
    highest accuracy classifier in the hypothesis class.
```

Suppose our hypothesis class is the class of all linear separators defined over $\mathbb{R}^2$.

(a) (1 point) The highest accuracy linear separator is shown in the figure below (assume that points on the decision boundary are characterized as '+'):

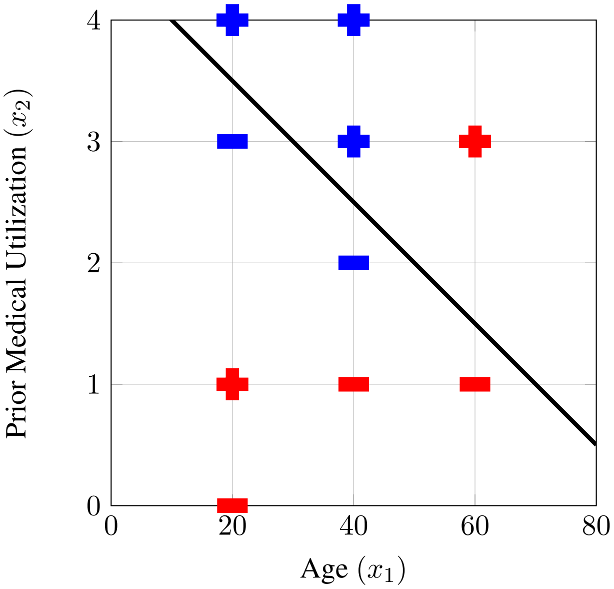


What are the coordinates of the data point whose label would be flipped first by this pre-processing method?

Your Answer

(60, 1)

(b) (2 points) After flipping the label of the point you identified in the previous question, plot the linear separator with the highest accuracy. For your convenience, we have provided a mechanism for you to input your answer by specifying the coordinates of two points on the decision boundary.
Note: Look at the LaTeX comments for instructions on how to redraw the decision boundry.



(c) (1 point) Using the linear decision boundary your plotted in the previous question, what are the coordinates of the data point whose label would be flipped next by this pre-processing method?

Your Answer

(40, 1)

(d) (1 point) **True or False:** The algorithm terminates at this point.

○ True

● False

7. **The fairness impossibility theorem:** Prove by contradiction that if prevalence rate, $r_s = P[Y = 1|S = s]$ across the two groups $s \in \{\text{blue}, \text{red}\}$ is different, then there does not exists a classifier that can satisfy PPV parity, FPR parity, and FNR parity simultaneously.

(a) (4 points) Verify that the following identify holds for any $s \in \{\text{blue}, \text{red}\}$:

$$FPR_s = \frac{r_s}{1 - r_s} \times (1 - FNR_s) \times \frac{(1 - PPV_s)}{PPV_s}$$

.

Your Answer

$$\begin{aligned}
\text{RHS} &= \frac{r_s}{1-r_s} \times (1 - FNR_s) \times \frac{1 - PPV_s}{PPV_s} \\
&= \frac{P(Y=1|s)}{P(Y=0|s)} \times TPR_s \times \frac{P(Y=0|\hat{Y}=1,s)}{P(Y=1|\hat{Y}=1,s)} \\
&= \frac{P(Y=1|s)}{P(Y=0|s)} \times TPR_s \times \frac{P(Y=0|\hat{Y}=1,s)P(\hat{Y}=1)}{P(Y=1|\hat{Y}=1,s)P(\hat{Y}=1)} \\
&= \frac{P(Y=1|s)}{P(Y=0|s)} \times P(\hat{Y}=1|Y=1,s) \times \frac{P(\hat{Y}=1|Y=0,s)P(Y=0|s)}{P(\hat{Y}=1|Y=1,s)P(Y=1|s)} \\
&= P(\hat{Y}=1|Y=0,s) \\
&= FPR_s
\end{aligned}$$

(b) (4 points) Show that the expression from part (a) can be rewritten as

$$1/r_s = 1 + \frac{(1 - FNR_s)}{FPR_s} \times \frac{(1 - PPV_s)}{PPV_s}$$

Your Answer

$$\begin{aligned}
FPR_s &= \frac{r_s}{1-r_s} \times (1 - FNR_s) \times \frac{(1 - PPV_s)}{PPV_s} \\
FPR_s - r_s \times FPR_s &= r_s \times (1 - FNR_s) \times \frac{(1 - PPV_s)}{PPV_s} \\
FPR_s &= r_s \times FPR_s + r_s \times (1 - FNR_s) \times \frac{(1 - PPV_s)}{PPV_s} \\
1 &= r_s + r_s \times \frac{(1 - FNR_s)}{FPR_s} \times \frac{(1 - PPV_s)}{PPV_s} \\
\frac{1}{r_s} &= 1 + \frac{(1 - FNR_s)}{FPR_s} \times \frac{(1 - PPV_s)}{PPV_s}
\end{aligned}$$

(c) (4 points) Finally, using results from parts (a) and (b), show that if FPR_s, FNR_s, and PPV_s are equal for $s \in \{\text{blue}, \text{red}\}$, then r_s must be equal for $s \in \{\text{blue}, \text{red}\}$, which is a contradiction.

Your Answer

We want to prove that if the prevalence rates (r_s) are different between the blue and red groups, then a classifier cannot satisfy FPR, FNR, and PPV parity at the same time.

Assume for contradiction that a classifier **does exist** that satisfies all three parity conditions, even when the prevalence rates are different. This means $FPR_{blue} = FPR_{red}$, $FNR_{blue} = FNR_{red}$, $PPV_{blue} = PPV_{red}$, and $r_{blue} \neq r_{red}$.
We apply the identity from part (b) to both the blue and red groups.

$$\frac{1}{r_{blue}} = 1 + \frac{1 - FNR_{blue}}{FPR_{blue}} \times \frac{1 - PPV_{blue}}{PPV_{blue}}$$
$$\frac{1}{r_{red}} = 1 + \frac{1 - FNR_{red}}{FPR_{red}} \times \frac{1 - PPV_{red}}{PPV_{red}}$$

Based on our assumption, all the terms on the right-hand side are equal for both groups. This implies that

$$\frac{1}{r_{blue}} = \frac{1}{r_{red}} \quad \text{or} \quad r_{blue} = r_{red}$$

This result directly contradicts our initial premise that the prevalence rates are different ($r_{blue} \neq r_{red}$).

7 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