

CS 4973: Responsible ML Midterm Quiz

Fall 2023

45 points

This quiz has 20 questions. You have the whole class (90 mins) to answer all the questions.

Your name:

Evaluation Metrics

Confusion Matrix: Definitions summarizes the performance of one binary classifier that we are all familiar with: a Covid test. Many common performance metrics are defined in terms of this confusion matrix. For example:

- Accuracy = $(TP + TN) / (TP + FN + FP + TN)$
- Positive Predictive Value = $TP / (TP + FP)$
- False Negative Rate = $FN / (TP + FN)$

Confusion Matrix: Definitions

	Predicted: Sick	Predicted: Well
Actual: Sick	True Positive (TP)	False Negative (FN)
Actual: Well	False Positive (FP)	True Negative (TN)

Confusion Matrix: All Patients

	Predicted: Sick	Predicted: Well
Actual: Sick	90	60
Actual: Well	10	40

1. Use *Confusion Matrix: All Patients* to compute the accuracy of the test across all patients. (Note: it is fine to leave your answers in this section as fractions) 1 point

2. Use *Confusion Matrix: All Patients* to compute the positive predictive value (PPV) of the test across all patients. Do we want to minimize or maximize the PPV? 2 points

3. Use *Confusion Matrix: All Patients* to compute the false negative rate (FNR) of the test across all patients. Do we want to minimize or maximize the FNR? 2 points

Fairness Metrics

Many common fairness metrics are comparisons of performance metrics across different subgroups. For example, the two confusion matrices below disaggregate the performance of the test across two important subgroups: children and seniors.

Confusion Matrix: Children

	Predicted: Sick	Predicted: Well
Actual: Sick	4	6
Actual: Well	1	39

Confusion Matrix: Seniors

	Predicted: Sick	Predicted: Well
Actual: Sick	20	20
Actual: Well	5	5

4. Compute the PPV difference across the two subgroups (children PPV - seniors PPV). 2 points

5. Compute the FNR difference across the two subgroups (children FNR - seniors FNR). 2 points

6. In this example, we chose to equalize one of the two fairness metric across children and seniors. Given the specific context of Covid testing, do you think we made the right choice about which metric to equalize? 1 point

7. What difference between children and seniors will make it fundamentally hard to achieve multiple different definitions of fairness? 2 points

Complicating Fairness Metrics

8. In class, we discussed multiple studies that measured race/ethnicity and gender in limited ways. What is one example of measuring these important variables in a limited way? What could a study miss if it measures race/ethnicity or gender in the way you described? 2 points

9. In class, we discussed how lots of work on machine learning fairness embeds a US perspective. Using a country of your choosing, describe two important contextual differences that a US-centric viewpoint misses or obscures. 2 points

Professor Ricks described three high-level ways in which algorithms can treat people unfairly a) purposes/goals, b) data collection practices, and c) decisions/outcomes.

Each of the next four statements describe a specific way in which an algorithm can treat people unfairly. Mark which of the three high-level categories it falls under.

10. OpenAI pays annotators in Kenya \$0.01 to label six different kinds of toxicity in a ChatGPT response. 2 points

Describe briefly why:

Mark only one oval.

- Purposes/Goals
- Data Collection Practices
- Decisions/Outcomes

11. A facial recognition system is trained only on images of white men from the US. 2 points

Describe briefly why:

Mark only one oval.

- Purposes/Goals
- Data Collection Practices
- Decisions/Outcomes

12. An algorithm predicts a person's sexuality from an image of their face. 2 points

Describe briefly why:

Mark only one oval.

- Purposes/Goals
- Data Collection Practices
- Decisions/Outcomes

13. A Covid test is optimized to minimize the false negative rate for seniors, the most at risk group. This causes the false negative rate to be higher for children. 2 points

Describe briefly why:

Mark only one oval.

- Purposes/Goals
- Data Collection Practices
- Decisions/Outcomes

Interpretability

14. Below is the formula and procedure for calculating Shapley Values. Say you have a model with three features (players). The outcome values for the different permutations are: $v(\{1\}) = 80$, $v(\{2\}) = 60$, $v(\{3\}) = 30$, $v(\{1,2\}) = 180$, $v(\{1,3\}) = 160$, $v(\{2,3\}) = 120$, $v(\{1,2,3\}) = 260$. 8 points

Given this information, fill in the rest of the table in the answer. The values for player 1 have already been filled in for you.


*Hint: Because Shapley is an **additive** explanation method, the sum of the Shapley values for features (players) 1, 2 and 3 should be equal to the value of the situation where all 3 features participate, i.e., $v(\{1,2,3\})$.*

1) Given a set N of players i , each of which can be attributed a value $N = \{1, 2, 3\}$,


2) We calculate a set of permutations R of N .

3) We then calculate the marginal contribution given by that feature in the following way:

$$\varphi_i(v) = \frac{1}{|N|!} \sum_R [v(P_i^R \cup \{i\}) - v(P_i^R)],$$



Set of features preceding i in order R , including i



Set of features preceding and excluding i

4) Where R is an ordering, given by permuting the values in set N , and P_i^R is the set of a players preceding i in the order R .

Answer:

$v(\{1\})$	$v(\{2\})$	$v(\{3\})$	$v(\{1,2\})$	$v(\{1,3\})$	$v(\{2,3\})$	$v(\{1,2,3\})$
80	60	30	180	160	120	260
		player 1's marginal contribution	player 2's marginal contribution	player 3's marginal contribution		
order	probability					
123	1/6	80				
132	1/6	80				
213	1/6	120				
231	1/6	140				
312	1/6	130				
321	1/6	140				
	sum	690				
	Shapley value	115				

Algorithm Auditing

In our week on algorithm auditing, we discussed three research designs that algorithm auditors often use: a) scraping audits, b) sock puppet audits, c) collaborative/crowdsourced audits.

Each of the next four statements describe a specific algorithm auditing research design. Mark which of the three high-level categories it falls under.

15. Use TikTok's research API to collect videos made by creators of different genders and races/ethnicities and the comments that were left on these videos. 2 points

Mark only one oval.

- Scraping Audit
- Sock Puppet Audit
- Collaborative Audit

16. Create multiple TikTok accounts and program each one to exclusively interact with (e.g. watch, like) a specific type of content (e.g. diet or mental-health related content). 2 points

Mark only one oval.

- Scraping Audit
- Sock Puppet Audit
- Collaborative Audit

17. Submit fashion queries to Pinterest, download the results page, and extract each pin. 2 points

Mark only one oval.

- Scraping Audit
- Sock Puppet Audit
- Collaborative Audit

18. Create a browser extension for Pinterest users that identifies when a search is made on Pinterest and uploads the pins on the results page to the research team's server. 2 points

Mark only one oval.

- Scraping Audit
- Sock Puppet Audit
- Collaborative Audit

Privacy

19. What is one of the primary reasons for the success of membership inference attacks in machine learning models? 2 points

Mark only one oval.

- Data underfitting
- Lack of auxiliary information
- Data overfitting
- Data diversity

20. Describe very briefly the mechanism of a membership inference attack using shadow models. Give an example of a realistic scenario where this type of attack could be used. 3 points
