Instructions

There are **25 total points**. When asked to provide your answer within a figure or table, be careful to not exceed box boundaries. Bubbles must be filled out completely: ● is correct, ✓ ⊙ ⊗ are incorrect All answers must be given within the provided circles, answer boxes, figures or tables.

1. [1 point]: Write your full name in the box to acknowledge the instructions.

Nick Feamster

Use Cases

2. [3 points]: Many networked systems and applications now rely on machine learning for performance modeling and prediction because modeling with closed-form equations has become too complex (and inaccurate). Give an example of a networked system—and associated *performance* prediction problem—where machine learning has been useful because closed-form modeling is not feasible.

(Answer inside the box)

Solution: Predicting web search response time is one example that we discussed in class. There are many possible answers.

3. [4 points]: We discussed applications of machine learning to security, such as detecting Internet scanning. Scanning for a web vulnerability (e.g., the Log4j vulnerability we explored in class) might look different than web traffic. (1) Why might a scan look different than regular Web traffic? (2) What is one feature you could encode in a machine learning model that could distinguish scanning from regular web traffic?

(Answer inside the box)

Solution: 1. Scans are looking for vulnerabilities and thus do not need to complete a full transaction. 2. Request rate (there are other possibilities, if justified!).

Data Acquisition

4. [2 points]: Suppose you want to extract all Netflix traffic from a traffic capture. Capturing all traffic to and from the IP address for `netflix.com` will yield all Netflix traffic streams. ◯ Yes ● No

5. [2 points]: In class, we used the domain name system (DNS) lookup traffic to identify Netflix traffic. This approach can work in practice but is imperfect. List one reason why DNS names may not always be practical for identifying traffic for services like Netflix. **(Answer box on next page!)**

Initials: ____

(Answer inside the box)

Solution: Domain names can change over time. DNS traffic is becoming increasingly encrypted, making it difficult to see domain name lookups and responses. Another reason is that the DNS names to these services can change.

6. [5 points]: What are the five header types in a network packet that make up a flow?

(Answer inside the box)

Solution: Source and destination IP address, source and destination port, protocol.

7. [3 points]: List three advantages to active Internet measurement over passive Internet measurement.

(Answer inside the box)

Solution: Direct measurement of desired effect, timing and frequency can be controlled, little to no privacy risks. (There was a whole slide on this in the board notes that we came up with in class, so anything from that slide, or anything reasonable, will suffice.)

Feature Engineering

8. [3 points]: Features should be characteristic of fundamental differences between classes, rather than simply characteristics of the dataset. Suppose you have a *single* packet trace from the University of Chicago campus network, where Log4j scans are being conducted at the same time as regular traffic. You decide to use *only incoming network traffic* to train a detection model, using features that include all of five the fields for incoming network traffic, and the detection model works really well. But, when your friends at Northwestern try to use your model, it doesn't work at all. What feature or features might be at fault, **and why**?

(Answer inside the box)

Solution: Destination IP address, because the destination IP addresses will be different at Northwestern. (Other fields, like source port, may be an acceptable answer if well-explained.)

Feedback

9. [1 point]: Interest (1=Boring!; 10=Amazing!): **5** Difficulty (1=Too easy; 10=Too hard): **5**

10. [1 point]: 1. One thing you like. 2. One suggestion for improvement:

(Answer inside the box)

Solution: More free food.

Initials: ____