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

Machine Learning Pipeline

2. [3 points]: One challenge with network traffic is that flow-based features are often different lengths (e.g., some flows last a few seconds, others may last minutes). (1) Give an example feature where different flows may have different feature lengths; (2) explain why this might be a problem for representing certain network features, and (3) explain how netml deals with normalizing the length of feature vectors.

(Answer inside the box)

Solution: 1. A feature representation of interarrival time between packets might give rise to this, for example, if two different flows have a different number of packets. 2. Many models require the features to be input as a matrix, with the same number of rows per sample/column. 3. Netml pads out shorter feature vectors with zeroes.

Supervised Learning

3. [2 points]: Linear regression models can only perform accurate predictions when there is a strict linear relationship between the input features and the target prediction. ○ Yes ● No

4. [3 points]: Despite their computational complexity, deep learning models often do not achieve better prediction accuracy than random forest on many common network prediction tasks. (We saw one example of this in class, for denial of service attack detection.) Why might this be the case?

(Answer inside the box)

Solution: Deep learning models are very good at learning complex non-linear relationships between features and targets. But, in many cases, input of semantic features can often produce better results if those features can in fact distinguish the classes. Deep learning models are also at some risk of overfitting, since they can learn complex relationships that may be in the training set but do not hold more generally.

Unsupervised Learning

5. [2 points]: Describe a common way of determining the appropriate value of principal components (k) when using principal component analysis to cluster data.

(Answer inside the box)

Solution: A scree plot (variance captured by each dimension) can help determine the point of “diminishing returns”.

Initials: ____

6. [4 points]: Provide two reasons that reducing the dimensionality of a dataset can be helpful, and one application of dimensionality reduction techniques (e.g., PCA) in networking.

(Answer inside the box)

Solution: 1. Visualization; 2. Reducing model complexity; 1. Anomaly detection/understanding structures in network traffic.

7. [3 points]: Describe (with illustration, if you like) an example dataset where density-based clustering may produce more "meaningful" clusters than k-means clustering.

(Answer inside the box)

Solution: Some examples have been provided in lecture, such as 2D data in circular shape, with a separate cluster of data points within that circular cluster.

Deployment Challenges

8. [2 points]: Describe (1) one advantage of using a bit-level data representation like nPrint as input to machine learning models for classification problems and (2) one example feature from network traffic that the nPrint representation does not efficiently represent.

(Answer inside the box)

Solution: 1. No need to engineer features in advance; 2. Any temporal based features (e.g., interarrival time) are great examples. Packet size could be an acceptable answer, as well (although technically that is represented inefficiently with padding).

9. [3 points]: Give an example where a semantic-free representation like nPrint could cause a model to learn a spurious correlation in a trained dataset (i.e., learning a feature that is a property of the dataset, not a truly distinguishing feature).

(Answer inside the box)

Solution: In class we discussed the example of learning TTL as an important feature, which is an example of learning the topology of an experiment, not anything fundamental.

Feedback

10. [2 points]: 1. Your favorite topic or activity in this course. 2. One topic you'd like to see covered that was not covered:

(Answer inside the box)

Solution: 1. Hands on. 2. Ethics and data privacy.

Initials: ____