Instructions

There are 35 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.

Solution:

Linear Regression and Regularization

2. [2 points]: Which of the following statements about polynomial basis expansion are correct? (Select all that apply.)

■ **It allows linear regression to model non-linear relationships**

◯ It requires changing the form of the regression function

■ **Higher degree polynomials increase risk of overfitting**

■ **It creates additional feature columns from existing features**

3. [3 points]: Explain why Ridge regression (L2 regularization) helps prevent overfitting by penalizing model complexity.

Solution: Ridge regression adds a penalty term (lambda times the sum of squared coefficients) to the error function. This penalty encourages coefficients to be small or zero, effectively reducing the number of features used. Fewer non-zero coefficients means a simpler model with fewer dimensions, which reduces overfitting by preventing the model from relying too heavily on many features.

Decision Trees and Ensemble Methods

4. [2 points]: Which of the following are ways that random forests introduce randomness compared to single decision trees? (Select all that apply.)

■ **Bootstrap sampling with replacement to create different training sets**

◯ Random selection of target variable

■ **Random feature selection at each split**

◯ Random initialization of tree depth

5. [2 points]: Decision trees are preferred over random forests when model interpretability is the primary concern.

Yes No

6. [2 points]: Explain why feature importance scores from random forests can be useful even if you don't use random forests as your final model.

Solution: Feature importance scores can guide exploratory data analysis by revealing which features are most predictive of the target variable. This helps understand the problem better and can inform feature engineering for other models. Even if you ultimately use a different classifier, knowing which features drive predictions helps with model debugging, feature selection, and understanding the underlying patterns in the data.

Neural Networks and Deep Learning

7. [2 points]: Which of the following are reasons why ReLU has become the most common activation function in hidden layers of neural networks? (Select all that apply.)

- ■ **It does not saturate for positive values, helping avoid vanishing gradients**
- ■ **It is computationally efficient to compute**
- ◯ It produces outputs in the range [0, 1] suitable for probabilities
- ■ **It creates sparse activations where many neurons output zero**

8. [2 points]: Explain why the vanishing gradient problem makes it difficult to train deep neural networks with sigmoid or tanh activation functions.

Solution: During backpropagation, gradients are multiplied together as they flow backward through layers. Sigmoid and tanh have gradients less than 1, especially in their saturated regions. When these small gradients are multiplied many times in deep networks, they approach zero. This means early layers receive extremely small gradients and their weights barely update, preventing the network from learning effectively.

9. [2 points]: Deep learning models operating on raw traffic input (like nPrint) always outperform traditional machine learning models with engineered features.

Yes **No**

10. [2 points]: Why or why not?

Solution: Deep learning excels at learning complex non-linear relationships from raw data, but traditional models like random forests can perform better when good semantic features are available. Traditional models may generalize better with limited training data and are less prone to overfitting when hand-crafted features effectively distinguish classes. The hands-on activities showed examples where random forests on engineered features matched or exceeded deep learning performance.

nPrint and Representation Learning

11. [2 points]: In nPrint's three-valued bitmap encoding, what do the values 1, 0, and -1 represent? (Select all correct mappings.)

- ■ **1 means the bit is set in the packet header**
- ■ **0 means the bit is not set in the packet header**
- ■ **-1 means the header field is not present in this packet**
- ◯ -1 means the bit value is unknown or corrupted

12. [2 points]: Explain why the -1 value is critical for nPrint's alignment problem. What would happen if nPrint only used 0 and 1?

Solution: The -1 value ensures every bit position has consistent meaning across all packets. Without -1, a UDP packet would have no TCP header fields, causing all subsequent fields to shift positions in the bitmap. For example, bit 320 might be a TCP option in one packet but a source IP address bit in a UDP packet. Models cannot learn consistent patterns when the same bit position means different things. The -1 placeholder keeps all fields aligned even when not present.

13. [2 points]: Which of the following are benefits of using nPrint for network traffic analysis? (Select all that apply.)

- ■ **Provides reproducibility across different research groups**
- ◯ Eliminates all risk of spurious correlations
- ■ **Enables automatic feature learning with deep learning models**
- ■ **Standardizes representation for comparing different approaches**

Dimensionality Reduction

14. [2 points]: Which dimensionality reduction technique is best suited for capturing non-linear relationships in data?

◯ PCA
■ **t-SNE**
◯ Ridge regression
◯ Random forest feature selection

15. [2 points]: Describe one advantage that autoencoders have over PCA or t-SNE for dimensionality reduction.

Solution: Autoencoders can work directly with raw, high-dimensional data without requiring manual feature engineering. Similar to how deep learning models learn features automatically, autoencoders learn the compression and decompression functions during training. This allows them to discover complex non-linear patterns that PCA (which assumes linearity) might miss, while requiring less manual feature design than traditional techniques.

Diffusion Models and Synthetic Traffic Generation

16. [2 points]: Which of the following are valid use cases for generating synthetic network traffic? (Select all that apply.)

■ **Training ML models when real labeled data is limited**
■ **Privacy-preserving data sharing without exposing real network topology**
◯ Replacing all real data collection to eliminate cost
■ **Augmenting datasets to handle class imbalance**

17. [2 points]: Explain the fidelity vs diversity trade-off in synthetic traffic generation. Why is finding the right balance important?

Solution: Fidelity measures how similar synthetic traffic is to real traffic, while diversity measures how much useful variation exists. Perfect fidelity (zero distance) means the synthetic data is just a copy with no new information. Maximum diversity (complete noise) means the synthetic data doesn't represent real traffic patterns. The right balance provides meaningful variations that preserve important traffic characteristics while adding useful diversity for model training and robustness. Too much fidelity offers no benefit, too much diversity hurts model performance.

Feedback

18. [1 point]: Interest (1=Boring!; 10=Amazing!): _______ Difficulty (1=Too easy; 10=Too hard): _______

19. [1 point]: 1. Your favorite topic or activity from meetings 11-16. 2. One topic you would have liked to see covered:

Solution: