Questions

Question 1: Model Selection with Data Leakage

You are evaluating four machine learning models for a classification task with the following error rates:

- Model A: Training error = 2%, Validation error = 15%, Test error = 14%
- Model B: Training error = 10%, Validation error = 11%, Test error = 12%
- Model C: Training error = 0%, Validation error = 0%, Test error = 25%
- Model D: Training error = 5%, Validation error = 6%, Test error = 7%

Upon investigation, you discover that Model C was trained with the validation set accidentally included in the training data.

- **Part A:** Which model would you select for deployment, and why? Analyze each model's performance, considering overfitting, underfitting, and the impact of data leakage.
- **Part B:** What does Model C's performance suggest about the training process and the dataset? Propose a method to detect such issues in future experiments.

Question 2: Diagnosing Underfitting in High-Capacity Models

You are training a deep neural network with 10 hidden layers, each containing 100 neurons, on a regression task with 1,000 samples. Despite the model's high capacity, both training and test errors remain high.

- **Part A:** What could be causing this underfitting? List at least two plausible reasons related to the model, training process, or data.
- **Part B:** How would you use learning curves to confirm your hypothesis? Describe the expected patterns in the curves for your proposed reasons.

- **Part C:** Suggest two specific modifications to the training process or model architecture to reduce underfitting, explaining how each addresses the issue.

Question 3: Evolving Model Capacity with Dataset Size

Imagine a regression problem where the dataset size increases from 100 to 10,000 samples. You are tasked with selecting the optimal degree of a polynomial regression model to minimize test error at each stage.

- **Part A:** Sketch a graph showing how the optimal polynomial degree might change as the dataset size increases. Explain the shape of this graph using the bias-variance tradeoff.
- **Part B:** If the true underlying function is a polynomial of degree 5, how would the optimal degree compare to 5 for small (e.g., 100 samples) and large (e.g., 10,000 samples) datasets? Justify your answer with reference to model capacity and generalization.

Question 4: Early Stopping vs. Dropout with Theoretical Tradeoffs

You're designing a neural network with two hidden layers to predict house prices, comparing early stopping and dropout (with a dropout rate of 0.3) as regularization strategies. The training dataset has 10,000 samples, but the validation set is only 500 samples, introducing uncertainty in performance estimates.

Task: Derive a theoretical comparison of how early stopping and dropout affect the model's expected generalization error, considering bias-variance decomposition. Incorporate the small validation set size into your analysis (e.g., how it impacts stopping decisions or dropout effectiveness). which method is preferable under specific data constraints (e.g., limited validation data or high model capacity)?

Question 5: The Batch Size Conundrum

A scientist is training a neural network with Batch Normalization to identify exoplanets from a small, noisy dataset with few samples per class. They notice that increasing the batch size from 8 to 64 reduces training loss but worsens generalization on a test set.

Task: why a larger batch size might harm generalization in this case. devise a clever strategy to improve generalization while retaining Batch Normalization, considering how batch statistics are computed and their impact on model robustness.

Question 6: The Initialization Puzzle

Two identical neural networks, Net X and Net Y, are set to explore a maze-solving task. Net X's weights are initialized from a normal distribution (mean 0, standard deviation 0.1), while Net Y's weights come from a uniform distribution (-0.5 to 0.5). Both use Batch Normalization.

Task: Predict which network, if either, will learn the maze more effectively and justify your reasoning. Explore how Batch Normalization interacts with these initialization choices and whether it fully neutralizes their differences in training performance.

Question 7: Image Classification of Flowers

Scenario:

You are participating in a science fair and have built a deep learning model to classify images of flowers into species such as roses, daisies, and tulips. Your dataset is limited to 100 images per species, collected from a controlled greenhouse environment. To make your model perform better on real-world images taken in varied lighting and angles, you decide to use data augmentation.

Task:

Describe three data augmentation techniques you could apply to your flower image dataset to increase its size and diversity. For each technique, explain how it would help your model generalize better to new images that might differ in orientation, brightness, or background.

Question 8: Speech Recognition in Noisy Environments

Scenario:

You are designing a voice-controlled robot for a robotics competition. The robot must recognize spoken commands like "forward," "stop," and "turn" in a noisy competition hall filled with cheering crowds and background chatter. Your initial audio dataset was recorded in a quiet room, so you plan to use data augmentation to adapt it to noisy conditions.

Task:

Propose two data augmentation techniques you could apply to your audio dataset to improve the robot’s ability to recognize commands in a noisy environment. For each technique, explain why it would help the model perform better under these challenging conditions.

Question 9: Comparing Augmentation Techniques for Animal Classification

Scenario:

In a wildlife monitoring project, you are training a model to classify trail camera images into categories like "deer," "fox," and "owl." Your dataset is small, with only 200 images per category, and the images vary widely in lighting and framing due to the camera’s automatic settings. You are deciding between two data augmentation techniques: random cropping and color jittering.

Task:

Compare and contrast random cropping and color jittering in terms of:

- How each technique might impact your model’s ability to accurately classify the animals.
- The specific types of variations they add to the training data.
- Which technique you would recommend for this wildlife project and why, considering the nature of the image variations.

Question 10: Feature Selection in House Price Prediction

Scenario:

You are part of a science olympiad team tasked with building a model to predict house prices for a local real estate competition. Your dataset includes features like number of bedrooms, square footage, age of the house, and distance to the city center. However, you suspect that not all features are equally important, and you want to simplify your model by focusing only on the most significant ones to make your predictions more efficient.

Question:

- Describe how you would use **L1 regularization** (Lasso) to identify and select the most important features for predicting house prices.
- Explain why **L1 regularization** is more suitable than **L2 regularization** (Ridge) for this feature selection task.
- What would happen to the coefficients of less important features when you apply L1 regularization, and how does this help your model?

Question 11: Preventing Overfitting in Plant Growth Analysis

Scenario:

In a biology experiment for the olympiad, your team is studying how environmental factors like sunlight, water amount, and soil quality affect the growth rate of plants. You've collected a small dataset with only 10 samples but have measurements for 15 different factors. After training a linear regression model, you notice it predicts perfectly on your collected data but fails miserably when tested on a new batch of plants from a different greenhouse.

Question:

- Explain what **overfitting** means in this situation and why it's likely happening with your model.
- Describe how **L2 regularization** (Ridge) can help improve your model's performance on the new batch of plants.

Question 12: Bias-Variance Tradeoff in Astronomy

Scenario:

Your olympiad team is analyzing a dataset of 50 stars, measuring 25 different properties (e.g., brightness, temperature, distance) to predict their ages. With so many features and so few samples, you're worried about how well your linear regression model will perform when predicting the ages of stars not in your dataset.

Question:

- Explain the **bias-variance tradeoff** and how it applies to this high-dimensional astronomy dataset.
- Discuss how **L1 regularization** and **L2 regularization** each influence the bias and variance of your model.
- In this scenario, where the number of features is large compared to the number of samples, which regularization technique would you choose to minimize prediction errors on new stars, and why?

Question 13: Image Recognition with Occlusion (Pattern Recognition and Robustness)

Scenario: You're given a printed image of a familiar object (e.g., a dog or the Eiffel Tower). Using a marker, you randomly cover 20% of the image with black dots or small stickers, simulating "dropout" of visual information. You then show this altered image to a group of classmates and ask them to identify it, repeating the process with different random patterns of occlusion.

Question: Test how well your classmates can recognize the image with 20% of it obscured compared to the original. Explain why recognition is still possible despite the missing parts and how this hands-on experiment relates to the machine learning technique of dropout, where some neurons are randomly ignored during training to enhance robustness.

Question 14: Medical Diagnosis Scenario

Scenario:

You are a data scientist working with a hospital to develop a machine learning model that can diagnose a rare disease based on patient symptoms and medical history. The dataset provided by the hospital has 10,000 patient records, but only

200 of them are confirmed cases of the rare disease. The rest are cases of more common diseases with similar symptoms.

Questions:

a) Explain why this dataset is considered imbalanced and discuss the potential problems this imbalance might cause when training a machine learning model.

b) Propose two different strategies to handle this imbalanced dataset. For each strategy, describe how it works and why it might be effective in this scenario.

c) Suppose you have access to additional datasets from other hospitals that contain more cases of the rare disease. How could you use these diverse datasets to improve the generalization of your model?

Question 15: Social Media Analysis Scenario

Scenario:

A social media platform wants to automatically detect and flag hate speech in user comments. They have a dataset of 100,000 comments, but only 2,000 are labeled as hate speech.

Questions:

a) Explain why this is an imbalanced dataset and how it might lead to a model that is biased towards predicting the majority class.

b) Propose a technique that adjusts the model's loss function to account for the class imbalance. Describe how this technique works and why it can be effective.