Exercises and Questions for Session 1: Introduction to Machine Learning

Here are some questions to self-evaluate your knowledge about the lesson. The answers are provided at the end of the document, but I recommend trying to answer them first without looking at the answers.

Knowledge Check Questions

1. What is supervised learning in Machine Learning?
2. What is the key difference between supervised and unsupervised learning?
3. Briefly explain the concept of "generalization" in Machine Learning.
4. What is a loss function, and why is it important?
5. Name one example of a classification problem and one of a regression problem.
6. What does "overfitting" mean, and how can it be avoided?
7. Mention two common dimensionality reduction techniques.
8. What are hyperparameters in a Machine Learning model?
9. What advantage does Machine Learning have over classical programming for complex tasks?
10. Explain what gradient descent is and its purpose.

Hands-On Exercises

1. **Data split:** Load the dataset `data_lesson_01.csv`. Answer the following questions:
   - What percentage of the data is missing?
   - Are there any data points that need to be removed? Justify your answer.
   - The last column is the target variable used for supervised learning. Is the target variable well-balanced?

   Based on your answers:

- Clean the dataset by handling missing values and removing any invalid data points.
- Create a random and well-balanced training-test split.
- Save each subset (training and test) as an independent file.

2. **Dataset Generation:** Write a Python program to create a synthetic dataset of size N (provided as a parameter to the program) with the following specifications:

   - **5 continuous variables:** Each following a uniform distribution ranging between 0 and 10.
   - **1 integer variable:** Following a binomial distribution with a mean of approximately 5 and a variance of 2.5.
   - **2 boolean variables:** Stored as strings (`'true'` and `'false'`), where `'true'` has a probability of 0.3.
   - **1 categorical variable:** Randomly chosen from the classes `'Bird'`, `'Dog'`, and `'Cat'`, with the following probabilities:
     - `'Cat'`: 50%
     - `'Bird'`: 25%
     - `'Dog'`: 25%

     Additionally, 2% of the entries in this categorical variable should be randomly converted to lowercase (`'cat'`, `'dog'`, `'bird'`).
   - **Generate and Save:** Create a dataset with $N = 10,000$ rows of data according to the specifications above, and save it in a CSV file named `synthetic_data.csv`.
   - **Validate the Dataset:**
     - Confirm that the summary statistics (e.g., mean, variance, probabilities) for each variable match the intended distributions.
     - For the categorical variable, verify that 2% of the entries are in lowercase and that the class proportions align with the given probabilities.
     - Provide a brief summary of the dataset (e.g., a table of summary statistics or visualizations of the distributions) in your submission.

Answers to Knowledge Check Questions

1. **What is supervised learning in Machine Learning?** Supervised learning involves training a model on labeled data, where each input has a corresponding output. The goal is to learn the mapping from inputs to outputs to make predictions on unseen data.

2. **What is the key difference between supervised and unsupervised learning?** Supervised learning uses labeled data with input-output pairs, while unsupervised learning works with unlabeled data to identify patterns or structures.

3. **Briefly explain the concept of "generalization" in Machine Learning.** Generalization is the ability of a model to perform well on unseen data, avoiding both underfitting (too simple) and overfitting (too complex).

4. **What is a loss function, and why is it important?** A loss function quantifies the error between predicted and actual outputs, guiding the optimization process to improve model accuracy.

5. **Name one example of a classification problem and one of a regression problem.** Classification: Email spam detection (spam vs. not spam). Regression: Predicting house prices based on size.

6. **What does "overfitting" mean, and how can it be avoided?** Overfitting occurs when a model memorizes the training data instead of learning generalizable patterns. It can be avoided through techniques like cross-validation, regularization, or using more data.

7. **Mention two common dimensionality reduction techniques.** Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

8. **What are hyperparameters in a Machine Learning model?** Hyperparameters are settings that define the structure of a model (e.g., number of layers in a neural network) or its training process (e.g., learning rate).

9. **What advantage does Machine Learning have over classical programming for complex tasks?** Machine Learning automates pattern discovery in data, making it effective for tasks where rules are too complex or numerous to explicitly program.

10. **Explain what gradient descent is and its purpose.** Gradient descent is an optimization algorithm used to minimize a loss function by iteratively adjusting model parameters in the direction of the steepest decrease in error.