Introduction to ML
Assignment 3

September 16, 2025

1 Conceptual Questions

This exercise should be sent as a pdf scan.

1. Give two real world example for:
   - Binary Classification Task
   - Multiclass Classification Task
   - Regression Task
2. Classifying into two classes, a model produces the following outputs:

y_i	$\hat{y}_i$
1	0.9
1	0.4
0	0.3
0	0.6

   Calculate the binary cross-entropy loss.
3. While improving a model complexity, if you notice that a model's training error is decreasing but the test error is getting higher, what does this indicate about the model's behavior?
4. When a model's training error is low but its test error is high, what term best describes this behavior?
   (a) Underfitting
   (b) Overfitting
   (c) Normalization
   (d) Convergence
5. A model's bias is the error due to:
   (a) Its sensitivity to small fluctuations in the training data.

(b) The difference between its average prediction and the true value.

(c) The randomness in the data collection process.

(d) Its inability to handle missing values.

6. The sum of a model's bias squared and its variance is equal to:

   (a) The Root Mean Squared Error (RMSE)

   (b) The Mean Absolute Error (MAE)

   (c) The Mean Squared Error (MSE)

   (d) The R-squared value

7. Suppose we have a dataset with categorical targets $Y = \{1, ..., K\}$.

   Let n_k be the size of the k-th category:

$$n_k = \sum_{i=1}^{n} \mathbb{I}[y_i = k], \quad \sum_{k=1}^{K} n_k = n.$$

   Consider a dummy model which always predicts category l, $1 \leq l < K$. What is the value of the error rate? For which l it is minimal?

8. The MSE for a constant model $f_\theta(x_i) = c$ is given by:

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - c)^2.$$

   Find the constant c that minimizes the MSE.

---

2 Properties of an Estimator in Machine Learning

This exercise explores the fundamental properties of estimators, which are crucial for understanding the performance and reliability of machine learning models. In this context, θ (theta) represents an unknown true parameter of the underlying data distribution that we are trying to estimate using our model. It could be, for instance, a regression coefficient, a class probability, or a population mean.

This exercise should be sent as a pdf scan.

2.1 Illustrate graphically a biased and an unbiased estimator for a parameter θ.

2.2 Illustrate graphically a consistent estimator for the parameter θ.

2.3 Assume there are two alternative estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for the parameter θ. Both estimators are unbiased, however, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$. Illustrate this in a graph.

2.4 Bias-Variance Decomposition

1. **Decompose the mean squared error (MSE) of an estimator into its components bias and variance.**

   Prove that for an estimator $\hat{\theta}$ of a parameter θ, the following relationship holds:

   $$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2$$

   where the bias is defined as $\mathrm{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$.

2. **Explain each component in the context of model performance.**

   - Explain what *bias* represents for a machine learning model.
   - Explain what *variance* represents for a machine learning model.

2.5 Using a graph, explain the possible bias-variance tradeoff of an estimator (draw two estimators). Include a comprehension question: Why does an estimator have a variance? Why is the estimator a random variable?

3 Data Scaling understanding (optional. I suggest to solve it if you struggle to understand scaling methods)

Consider the following vector of values:

$$x = \{4,\ 8,\ 6,\ 5,\ 3,\ 7,\ 9\}.$$

1. Compute the following descriptive statistics of x:
   (a) Mean
   $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
   (b) Standard deviation
   $$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
   (c) Interquartile range (IQR)
   $$IQR = Q_3 - Q_1$$
   where Q_1 is the first quartile and Q_3 the third quartile.
2. Apply the following scaling methods to the data by hand (use the formulas):
   - **Min–Max Scaling (Normalization):**
   $$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$
   - **Standardization (Z-score Scaling):**
   $$x' = \frac{x - \mu}{\sigma}$$
   where μ is the mean and σ is the standard deviation.
   - **Robust Scaling (useful for data with outliers):**
   $$x' = \frac{x - \text{median}(x)}{\text{IQR}}$$
   where $\text{IQR} = Q_3 - Q_1$ is the interquartile range.
3. For each scaled dataset, recompute the mean, standard deviation, and interquartile range.

4. Compare the statistics of the scaled data with those of the original vector. Which quantities are preserved, and which ones change?

5. **Thinking Question (no calculations needed):**

   In practice, we always **split the data into training and test sets before scaling**. That is, we fit the scaler (compute mean, standard deviation, minimum, maximum, etc.) using only the training set, and then apply the same transformation to both training and test data.

   (a) Why do you think scaling *before* the train/test split might cause problems?

   (b) How might the model be affected if information from the test set is used when scaling?

   (c) Imagine you accidentally scaled before splitting. Do you think your evaluation metrics (accuracy, RMSE, etc.) would become artificially higher or lower? Why?

---

4 Practical Exercises

In this exercise, we assume a very simple model that makes a constant prediction[1], represented by the parameter c. This model ignores all input features and simply predicts the same value for every data point. The objective is to find the specific value of c that minimizes the Mean Squared Error.

This exercise should be made in a jupyter notebook.

4.1 Generate synthetic data (Y) following any type of distributions.

```
import numpy as np
import matplotlib.pyplot as plt
```

4.2 Create an array of different values for the constant c (e.g., from 0 to 10). Calculate the MSE for each value of c.

4.3 Find the optimal c, compare it to the mean of your synthetic dataset

4.4 Plot the MSE values against the different values of c. Mark the optimal c on the plot.

4.5 Based on your final plot, what does the shape of the MSE curve tell you about the relationship between the constant c and the model's error? Why does the MSE increase as c moves away from the mean of your data?

4.6 In your opinion, why might a simple dummy model be useful even when we have access to more powerful and complex machine learning algorithms?

[1] Dummy model

Bonus: Write your own implementation of splitting the dataset on train and test using shuffling

```python
def train_test_split(X, y, test_size=0.2):
    """
    Split the dataset into training and testing sets.

    Parameters:
    X (numpy array): The feature matrix.
    y (numpy array): The target labels.
    test_size (float): The proportion of the dataset to include in the
        test split.

    Returns:
    X_train (numpy array): The training feature matrix.
    X_test (numpy array): The testing feature matrix.
    y_train (numpy array): The training labels.
    y_test (numpy array): The testing labels.
    """

    # YOUR CODE GOES HERE

    return X_train, X_test, y_train, y_test


# Example Usage
X = [[1, 2], [3, 4], [5, 6], [7, 8]]
y = [0, 1, 0, 1]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
```