ML Assignment Report 1

AI211 - Machine Learning

Name :- Kotha Dhakshin

Entry Number :- 2024AIB1009

Question 1 : Task 1 (Binary Classification with Threshold Hypothesis)

Dataset Description

In this task, we generated a one-dimensional dataset with binary labels $y \in \{0, 1\}$. The labels were sampled with equal probability. The feature values were generated from two different Gaussian distributions depending on the label. Since the two distributions overlap, samples from both classes appear in the same region of the x-axis. Because the class means are different, a threshold-based classifier is suitable for this dataset.

Observations for Each Subtask

Dataset Visualization

From the plot, we observed two main clusters corresponding to the two classes. However, there is an overlapping region where points from both classes are mixed. This shows that perfect separation using a single threshold is not possible.

Empirical Risk Minimizing Threshold

The empirical threshold was obtained by minimizing the training error. We observed that the learned threshold changes slightly with different random samples. Shifting the threshold reduces errors for one class but increases errors for the other, and the ERM threshold balances this trade-off.

Optimal Threshold & Generalization Error

The theoretical optimal threshold lies near the intersection of the two class distributions. The empirical threshold was close to this value but not exactly equal due to finite sample effects. As the sample size increased, the empirical threshold moved closer to the optimal one.

Empirical vs Generalization Error vs N

- For small N, the empirical error was unstable and sometimes very low, while the generalization error was high.
- As N increased, the generalization error decreased and both errors stabilized.
- Beyond a sufficiently large N, the learned threshold remained close to the optimal threshold, and the error rates converged to a constant value due to class overlap.

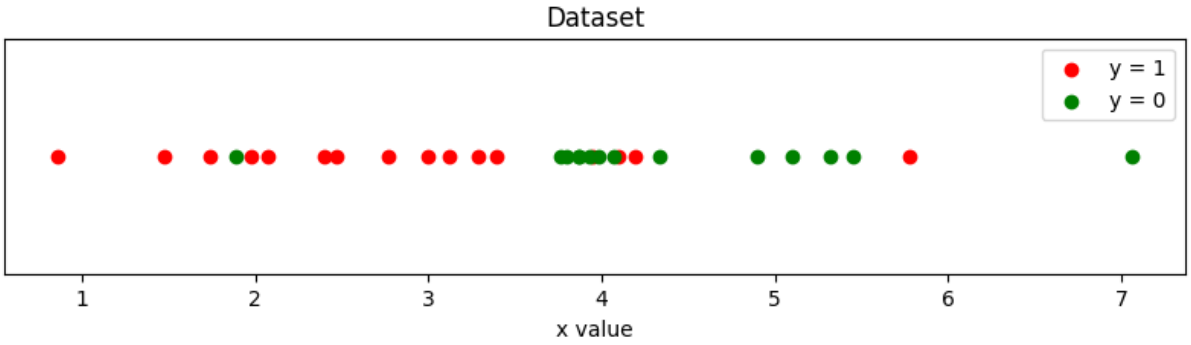

Figure 1: Visualization of the generated one-dimensional dataset for Question 1

Observations from Figure 1:

- The red points correspond to samples with label $y = 1$, while the green points correspond to samples with label $y = 0$.
- Most of the red points are concentrated toward the left side of the x-axis, whereas most of the green points are concentrated toward the right side. This pattern is expected because samples with $y = 1$ are drawn from a Gaussian distribution with a smaller mean compared to samples with $y = 0$.
- There exists a noticeable overlapping region between the two classes, especially around the middle values of x. This overlap occurs due to both classes having the same standard deviation, which causes their Gaussian distributions to intersect.
- A few red points appear at higher x values and a few green points appear at lower x values. These points act as outliers and arise naturally from random sampling in Gaussian distributions.
- Due to the overlapping nature of the two classes, it is not possible to perfectly separate the data using a single threshold. Any threshold-based classifier will result in some misclassification.
- However, since the majority of red points lie to the left and green points lie to the right, a threshold hypothesis is still a reasonable choice and is expected to achieve low (but non-zero) classification error.

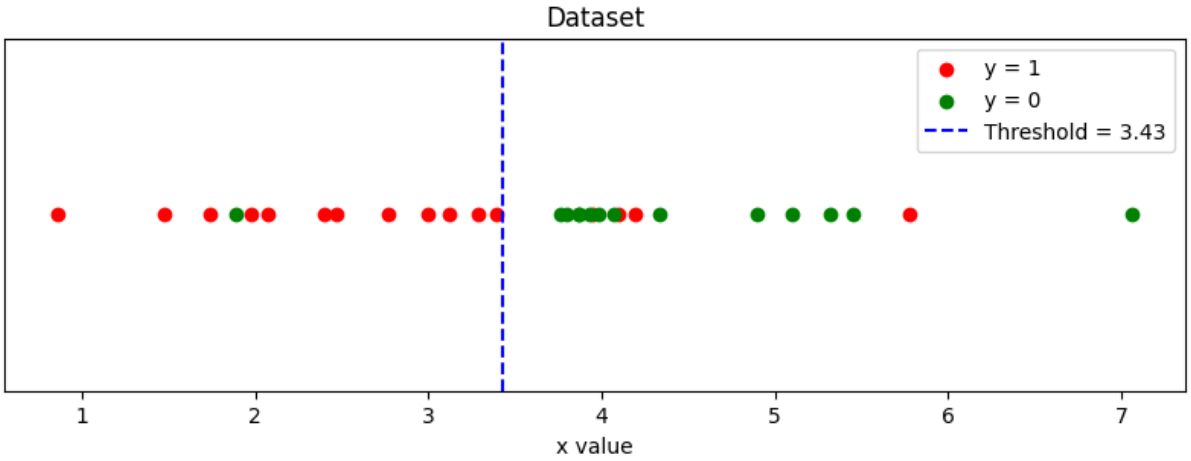

Figure 2: Dataset with empirically optimal threshold for Question 1

Observations from Figure 2:

- The dashed vertical line represents the threshold value θ obtained by minimizing the empirical risk on the given dataset.
- Points to the left of the threshold are classified as $y = 1$ (red), and points to the right are classified as $y = 0$ (green), following the chosen threshold hypothesis.
- Most red points lie on the left side of the threshold and most green points lie on the right, indicating that the learned threshold separates the two classes reasonably well.
- A few red points appear on the right side of the threshold and a few green points appear on the left side. These points are misclassified due to overlap between the two Gaussian distributions.
- The presence of these misclassified points shows that zero empirical error is not achievable, and the obtained threshold is a compromise that minimizes the total classification error.

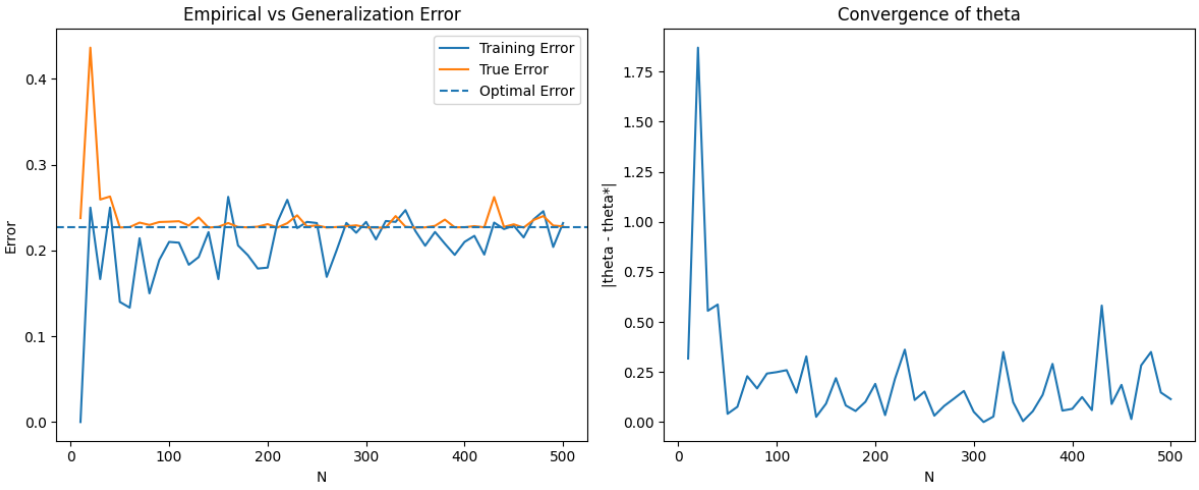

Figure 3: Empirical vs generalization error and convergence of threshold with increasing sample size

Observations from Figure 3:

- From the left plot, the empirical (training) error shows large fluctuations for small values of N, indicating high variance when the dataset size is small.
- As N increases, both the empirical error and the generalization (true) error stabilize and move closer to the optimal error level shown by the dashed horizontal line.
- The gap between empirical error and generalization error reduces as the number of samples increases, showing improved generalization with more data.
- From the right plot, the difference $|\theta - \theta^*|$ is large for small N, meaning the learned threshold is unstable when data is limited.
- As N increases, the value of $|\theta - \theta^*|$ decreases and remains small, indicating convergence of the empirical threshold towards the optimal threshold.
- From the plot, it can be observed that after a moderate value of N, the learned threshold stays within a small ϵ distance of θ^*, showing consistency of the estimator.

Question 2 : Task 2 (Polynomial Regression & Regularization)

Data Generation Explanation

The data was generated using a non-linear function with added Gaussian noise. The curve increases rapidly, so a simple linear model cannot fit it well. The added noise makes the problem more realistic by ensuring that points do not lie exactly on the curve.

Subtask-wise Observations

1. Analytical Polynomial Regression
We fitted polynomial models of degree $M = 2, 4, 5, 7, 10$ using the analytical solution.

- For low values of M, the model was too simple and could not capture the curvature of the data (underfitting).
- As M increased, the fit improved and followed the data more closely.
- For very high M, the curve became sensitive to noise, indicating overfitting.

2. Gradient Descent Polynomial Regression
Using gradient descent, we observed that convergence depended on the learning rate. After choosing a suitable learning rate, the fitted curves closely matched those obtained from the analytical solution, although gradient descent required many iterations.

3. Effect of More Data
With an increase in the number of data points:

- Overfitting reduced, even for higher polynomial degrees.
- The solutions obtained using analytical methods and gradient descent became more stable and similar.

4. Runtime vs Polynomial Degree
The analytical solution remained fast for the tested degrees. Gradient descent took more time and showed higher runtime as the polynomial degree increased due to iterative updates.

5. Runtime vs Data Size
As the number of data points increased, the analytical method became computationally expensive, while gradient descent scaled better with data size, making it more suitable for larger datasets.

Regularization Experiments

Ridge vs Lasso
We applied Ridge and Lasso regularization for $M = 5, 10, 15$.

- Ridge reduced overfitting by shrinking all coefficients smoothly.
- Lasso enforced sparsity by setting some coefficients to zero, simplifying the model.

Effect of λ

- Small values of λ led to overfitting.
- Large values of λ caused underfitting.
- An intermediate value of λ gave the lowest test error.

Fix $M = 10$, Vary n
As the number of samples increased, the need for strong regularization reduced since the model could learn the underlying pattern more reliably.

Sample Complexity Experiment

We observed that higher polynomial degrees required more data to achieve good performance. This is expected since more complex models need more samples to generalize well.

Non-Realizable Noise Case

With Poisson noise added, the error did not go to zero even for large M. This shows that some error is unavoidable due to noise, and increasing model complexity alone cannot eliminate it.

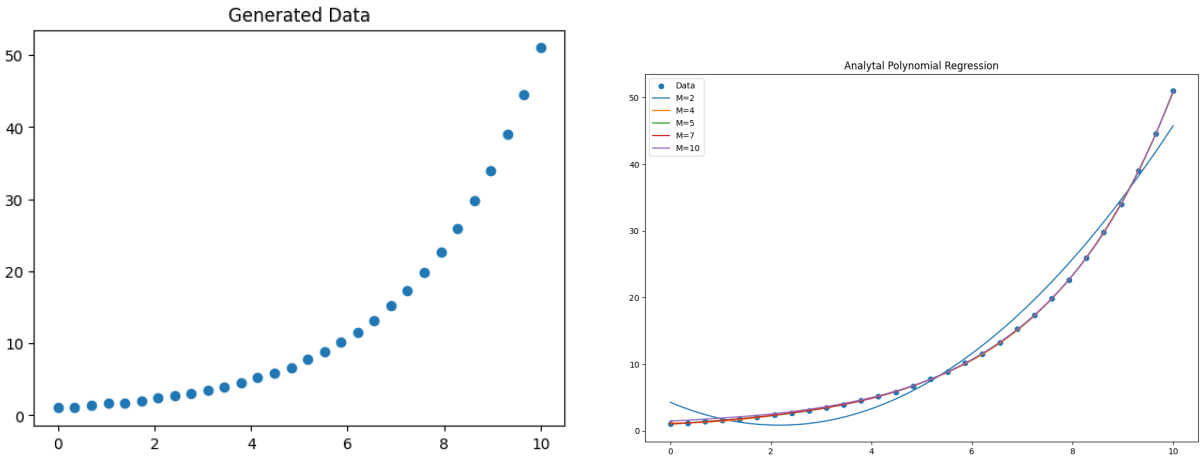

Figure 4: (Left) Generated data from the target function with noise. (Right) Analytical polynomial regression fits for different degrees M.

Observations from Figure 4:

- The generated data follows a clear non-linear increasing trend, mainly dominated by the exponential term in the target function. Small variations around the curve are visible due to the added Gaussian noise.
- For low polynomial degree ($M = 2$), the fitted curve is unable to capture the sharp increase in the data, showing clear underfitting.
- As the polynomial degree increases ($M = 4$ and $M = 5$), the fitted curves start matching the overall shape of the data much better.
- Higher-degree polynomials ($M = 7$ and $M = 10$) closely follow the data points across the entire range, indicating a good fit for the given dataset.
- From the plot, it can be observed that increasing the polynomial degree improves the approximation of the target function, but very high degrees may increase the risk of overfitting if noise dominates.

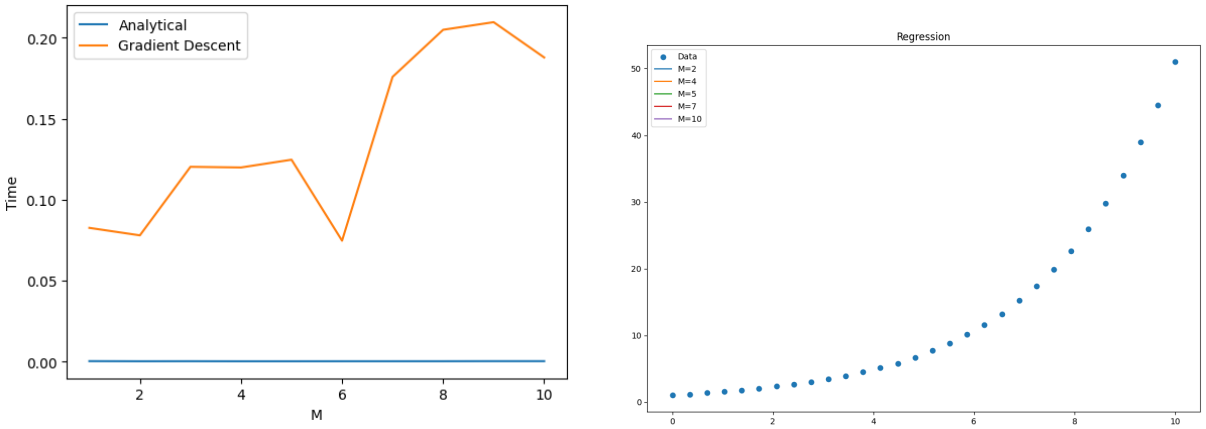

Figure 5: (Left) Runtime comparison of analytical solution and gradient descent for different polynomial degrees M. (Right) Polynomial regression curves obtained using gradient descent for different values of M.

Observations from Figure 5:

- From the left plot, the runtime of the analytical method remains almost constant and very small for different values of M, since it directly computes the solution using matrix operations.
- The runtime of gradient descent is significantly higher and increases with the polynomial degree M, as more parameters require iterative updates over many epochs.
- From the right plot, for low polynomial degrees, the gradient descent model underfits the data and fails to capture the sharp increase in the target function.
- As M increases, the fitted curves using gradient descent align better with the data points, showing improved approximation of the underlying function.
- Overall, gradient descent is computationally more expensive but flexible, while the analytical solution is faster for small datasets and lower polynomial degrees.

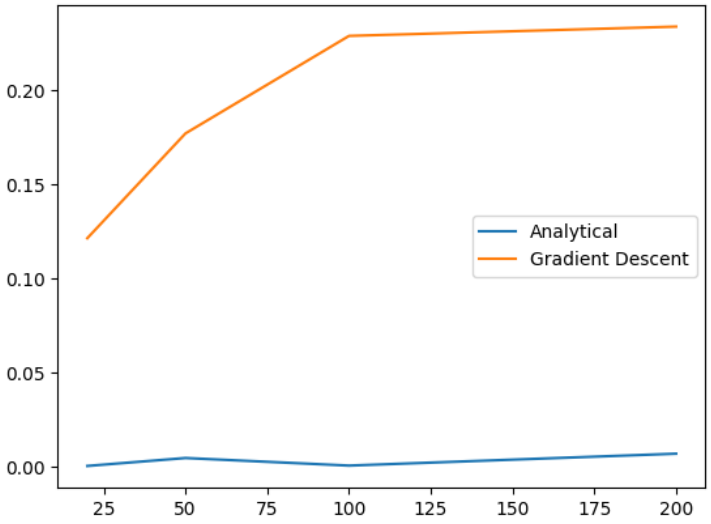

Figure 6: Runtime comparison of analytical solution and gradient descent with respect to number of data points

Observations from Figure 6:

- The analytical method shows very low runtime for all values of n, since the solution is computed directly using matrix operations.
- The runtime of gradient descent increases as the number of data points increases, because each iteration requires computing gradients over more samples.
- For small datasets, both methods are computationally feasible, but gradient descent already takes noticeably more time compared to the analytical approach.
- This plot shows that while the analytical solution is faster for small-scale problems, gradient descent becomes more expensive as the dataset size grows.

Question 3 : Task 3 (ERM, VC Dimension & Agnostic Learning)

Hypothesis Class Explanation

The hypothesis class consists of rectangular decision regions in a 2D space. Each hypothesis is defined by two parameters that form a rectangle. Points inside this rectangle are labeled as positive, and points outside are labeled as negative. In the realizable case, the true labeling function is also a rectangle of this form.

Subtask 3a : ERM Finder

The ERM algorithm searches for the rectangle that minimizes classification error on the training data. We restricted the search to boundaries defined by the data points, since changing the boundary between two points does not affect the error. In the realizable setting, the ERM was able to find a rectangle that achieved zero empirical error.

Subtask 3b : Error vs Sample Size

We studied how error changes with the number of samples N.

- The empirical error remained close to zero because the target hypothesis is realizable.
- For small N, the true error was high due to poor approximation of the true rectangle.
- As N increased, the estimated rectangle better matched the true rectangle and the true error decreased.
- The variability in error reduced as more samples were used.

Subtask 3c : Sample Complexity

The theoretical sample complexity bound was much larger than what we observed in practice. Empirically, we achieved low error with fewer samples. This difference occurs because theoretical bounds are conservative and consider worst-case scenarios.

Subtask 3d : Noisy Labels

When label noise was introduced:

- Zero error was no longer achievable because labels were inconsistent.
- The ERM still reduced error as much as possible but had to tolerate some misclassifications.
- The error converged to a non-zero value, representing the minimum achievable risk due to noise.

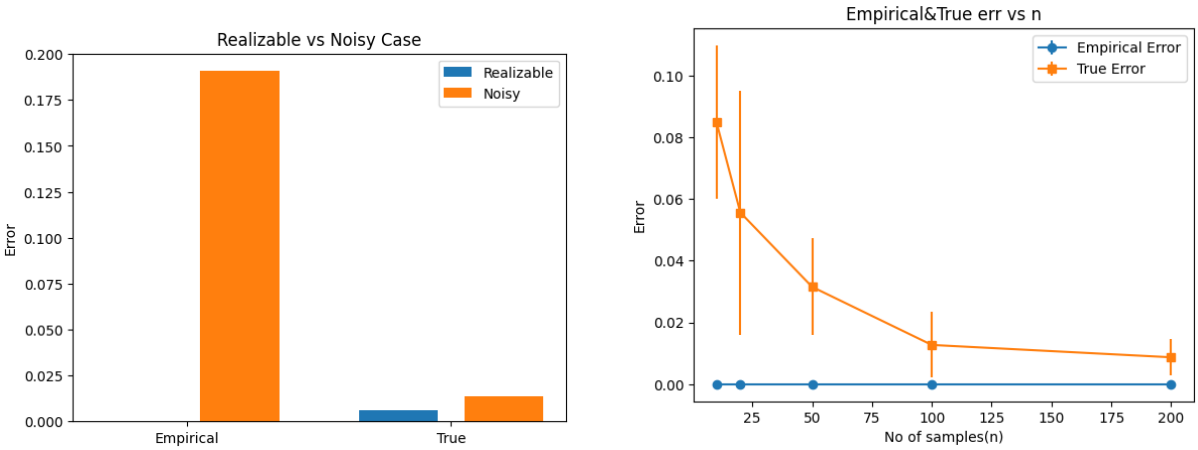

Figure 7: (Left) Comparison of empirical and true errors in realizable and noisy settings. (Right) Empirical and true error convergence with increasing sample size n.

Observations from Figure 7:

- From the left plot, in the realizable case, the empirical error is almost zero and the true error is also very small, indicating that the hypothesis class is capable of representing the true labeling function.
- In the noisy case, the empirical error is significantly higher and the true error does not go to zero. This shows that due to label noise, perfect classification is not possible even with the best hypothesis.
- From the right plot, for small values of n, the true error is high and shows large variation, indicating insufficient data for reliable learning.
- As the number of samples increases, the true error consistently decreases and becomes more stable, showing convergence towards the minimum achievable risk.
- The empirical error remains close to zero in the realizable case, while the true error reduces gradually with more samples, demonstrating improved generalization with increasing data.