Machine Learning Assignment 4 Report

111062117, Hsiang-Sheng Huang

November 10, 2024

1. Differences between Sigmoid and Softmax Activation Functions

- **Sigmoid** is used for binary classification. It outputs a probability between 0 and 1:

$$\sigma(Z) = \begin{cases} \frac{1}{1+e^{-Z}}, & Z \geq 0 \\ \frac{e^{Z}}{1+e^{Z}}, & \text{otherwise} \end{cases}$$

- **Softmax** is used for multi-class classification. It outputs a probability distribution:

$$\sigma(\vec{Z})_i = \frac{e^{Z_i - b}}{\sum_{j=1}^{C} e^{Z_j - b}}, \quad b = \max_{j=1}^{C} Z_j$$

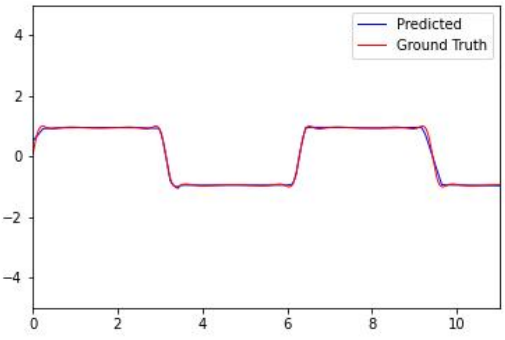
2. Reasons for Loss Oscillation

- **High Learning Rate**: Causes the model to overshoot, resulting in loss fluctuations.
- **Batch Size Variation**: Different batches lead to varying gradient estimates, causing minor oscillations.

3. Effect of Learning Rate and Batch Size on Training Time

- **Learning Rate**: Higher learning rates can reduce training time by making larger updates, but they increase the risk of overshooting or divergence. Lower learning rates tend to increase training time but usually lead to more stable convergence.
- **Batch Size**:
  - **Larger Batch Sizes**: Larger batches provide a more accurate estimate of the gradient, leading to faster convergence in terms of epochs. However, they require more memory and may not always lead to faster training in terms of wall-clock time due to computation limits.
  - **Smaller Batch Sizes**: Smaller batches make more frequent updates. However, the gradient estimates are noisier, which can lead to oscillations in the loss but might help in avoiding local minima by adding more exploration.

4. Regression Results


Predicted
Ground Truth