COMP24112 Lab Report

Yashika Gandhi

April 30, 2024

1 Linear Classification via Gradient Descent

1.1

Given:

$$O = C\sum_{i=1}^{N} \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)\right) + \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

The hinge loss term inside the summation is defined as:

$$L_i = \max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0)\right)$$

The gradient of O with respect to $\mathbf{w}$ can be computed using the chain rule. To differentiate the hinge loss L_i we need to consider two cases:

1. When $1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 0$:

$$\frac{\partial L_i}{\partial \mathbf{w}} = -Cy_i\mathbf{x}_i$$

2. When $1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) \leq 0$:

$$\frac{\partial L_i}{\partial \mathbf{w}} = 0$$

The gradient of the regularization term $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ is:

$$\frac{\partial}{\partial \mathbf{w}}\left(\frac{1}{2}\mathbf{w}^T\mathbf{w}\right) = \mathbf{w}$$

So, the gradient of O with respect to $\mathbf{w}$ can be written as:

$$\frac{\partial O}{\partial \mathbf{w}} = C\sum_{i=1}^{N} \left\{ \begin{array}{ll} -y_i\mathbf{x}_i & \text{if } 1 - y_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 0 \\ 0 & \text{otherwise} \end{array} \right\} + \mathbf{w}$$

1.2

The figure from section 2.1, which plots the cost (or loss) versus iteration indicates that the optimization algorithm is effectively minimizing the cost function for training iterations through its decreasing graph. The other figure, which depicts the accuracy per iteration, can be used to infer the training and testing sets' classification accuracies. Since this figure shows an increasing trend, it suggests that as the training progresses, the model's accuracy on both the training and testing sets is improving. Based on its performance on the testing set, this improvement in accuracy suggests that the model is learning to generalize well to unseen data.

1.3

With smaller learning rates ($\eta = 0.0001$ and $\eta = 0.001$), the model achieves high accuracy, precision, recall, and F1 score on the testing set. This suggests that the model generalizes well to unseen data and performs effectively in classifying instances. As the learning rate increases ($\eta > 0.01$), the model's performance starts to deteriorate, with a decrease in accuracy, precision, recall, and F1 score. This indicates that the model's ability to generalize diminishes as it becomes less stable and more prone to overfitting.

2 Air Quality Analysis by Neural Network

2.1

Based on the model selection results, the conclusions drawn are as follows:

1. The best parameters selected by the grid search are {'activation': 'relu', 'hidden_layer_sizes': (100, 100)}. This indicates that the rectified linear unit (ReLU) activation function and a neural network architecture with two hidden layers, each containing 100 neurons, provide the best performance for the regression task.

2. The cross-validation MSE, obtained from the grid search, is 0.0236. It serves as an estimate of the model's generalization error on unseen data.

3. The standard deviation of the cross-validation MSE is 0.0083, which suggests more consistency in model performance.

4. The MSE on the testing set is 0.0227, and the R^2 score is 0.6791. A lower MSE and higher R^2 score indicate better predictive performance of the model on unseen data.

Based on these results, it can be concluded that the selected neural network model with the specified parameters performs reasonably well for the regression task.

2.2

The Adam optimizer demonstrates lower training loss and training MSE compared to SGD throughout the training process.However, the final test MSE for Adam is relatively higher compared to SGD, suggesting potential overfitting or suboptimal generalization to the testing set.The $\hat{R2}$ score for Adam on the testing set is 0.4864, indicating moderate predictive performance. Overall, both Adam and SGD show effectiveness in minimizing the training loss and training MSE over iterations. However, SGD demonstrates better generalization to unseen data, as evidenced by its lower final test MSE and higher $\hat{R2}$ score compared to Adam.