Learning rate

Chain rule

Weight updating process

Vanishing gradiant problem – occurs because derivative of sigmoidal function is very less and range between 0 to 0.25. as the number of hidden layer increases value of gradient decent become so less. Thus vanishing gradient problem arises. Tanh (derivative between 0 to 1)also show same issue. Thus we do not use the sigmoidal function in the hidden layers. Its solved by ReLU because the derivative is 0 or 1

Exploding Gradient problem

Overfitting problem due to multiple hidden layers – solution – dropout and regularization – dropout is decided by the p value- p value can be decided by the hyperparameter optimization ss

Forward propagation

Backward propagation

One iteration

One epoch

Activation Functions

1. Sigmoid Activation Functions - output (0 to 1) BINAR CLASSIFICATION
2. Hyperbolic Tangent (tanh) Activation Functions - output (-1 to 1) HIDDEN LAYERS
3. Rectifier Linear Unit (ReLU) Activation Functions- negative values are considered as 0
4. Leaky ReLU Activation Functions

* Solves the problem of dead neuron problem of ReLU
* Derivative of the values of leaky RelU will be 1 or a value close to zero but not completely zero
* Thus the dead neuron problem does not occur and the slope value gets upgraded in back propagation

1. Parametric ReLU (PReLU) Activation Functions
2. Exponential Linear Unit (ELU) Activation Functions
3. Scaled Exponential Linear Unit (SELU) Activation Functions
4. Softmax Activation Functions

Loss functions

Classification Loss Functions:

1. Binary Crossentropy Loss – binary classification problems
2. Categorical Crossentropy Loss – Multiclass classification problems
3. Sparse Categorical Crossentropy Loss – when true labels are integers (not one hot encoded)
4. Hinge Loss – SVMs

Regression Loss Functions:

1. Mean Squared Error (MSE) Loss
2. Mean Absolute Error (MAE) Loss
3. Huber Loss - Combines characteristics of MSE and MAE. Less sensitive to outliers than MSE.

Optimizers

Gradient Descent – based Optimizers

1. Stochastic Gradient Descent (SGD)
2. Mini-batch Gradient Descent
3. SGD with Momentum
4. Nesterov Accelerated Gradient (NAG)

Adaptive Learning Rate Optimizers

1. Adagrad
2. RMSprop
3. Adam (Adaptive Moment Estimation)
4. Adadelta

Weight initialization techniques

1. Zero Initialization
2. Random Initialization
3. He Initialization (He Normal, He Uniform)
4. Xavier/Glorot Initialization (Xavier Normal, Xavier Uniform)
5. LeCun Initialization
6. Orthogonal Initialization
7. Identity Initialization
8. Scaled Exponential Linear Units (SELU) Initialization
9. Randomized Leaky ReLU Initialization
10. Uniform Initialization
11. Glorot Uniform Initialization (Xavier Uniform)

Parameters you can control

1. Number of hidden layers
2. Number of neurons in specific hidden layer
3. Activations functions
4. Optimizer
5. Learning rate
6. Number of epochs
7. Batch size
8. Weight initialization
9. Regularization (L1/L2/Dropout)