CODE BASICS

DEEP LEARNING TAKEAWAYS

codebasics.io

Chapter:

Neural Networks: Training

Training through Back Propagation

1. For training a neural network, we use a supervised training dataset. Feed samples one by one, calculate error and then back propagate it to adjust weights
2. The main objective of training is to find out right weights for the neural network. It is like adjusting nobs on a sound board to get expected audio output
3. Error backpropagation uses partial derivative to measure how much is the specific weight contributing to an error and based on that adjustments are made to that weight so that error is reduced in the next iteration
4. One epoch is feeding all the records in your dataset through the network once during a training process
5. MSE (Mean squared error) is one of the many cost functions used to measure error

Gradient Descent: Theoretical Foundation

1. Gradient Descent is a technique used in neural networks and statistical ML algorithms to find out optimal value of weights that will result into minimal prediction error. That optimal point is also known as global minimum
2. It uses gradient (or partial derivative) of error with respect to weights to perform weight adjustment
3. Learning rate is a hyper parameter that we need to supply in gradient descent

Batch GD vs Mini Batch GD vs SGD

1 Following table summarizes the differences between Batch GD, Mini Batch GD and SGD

Feature	Gradient Descent (GD)	Mini-Batch Gradient Descent	Stochastic Gradient Descent (SGD)
Data Used per Update	Entire dataset	Small batch (subset) of data	Single data point
Convergence Speed	Slow due to large data calculations	Faster than GD, slower than SGD	Fastest due to frequent updates
Computational Efficiency	High cost per iteration	Balanced cost per iteration	Low cost per iteration
Memory Usage	High (entire dataset in memory)	Moderate (batch in memory)	Low (single data point in memory)
Convergence Stability	Stable but may get stuck in local minima	Balanced stability and speed	Less stable, high variance in updates
Noise in Gradient Updates	Low (smooth updates)	Moderate (controlled by batch size)	High (due to single data point sampling)
Suitability for Large Datasets	Not ideal	Well-suited	Very well-suited
Example Use Case	Small datasets	Most real-world applications	Large datasets, online learning scenarios