NPTEL ONLINE CERTIFICATION COURSES

Course Name: Deep Learning

Faculty Name: Prof. P. K. Biswas

Department : E & ECE, IIT Kharagpur

Topic

Lecture 16: Optimization

CONCEPTS COVERED

Concepts Covered:

- Multiclass SVM Loss Function
- Optimization
- Stochastic Gradient descent
- Batch Optimization
- Mini-Batch Optimization

Optimizing Loss Function

$$L = \frac{1}{N}\sum_{i}\sum_{j \neq y_i}\left[\max(0, W_j^t X_i - W_{y_i}^t X_i + \nabla\right] + \lambda \sum_{k}\sum_{l} W_{kl}^2$$

$$\nabla_{W_{y_i}} = -\frac{1}{N}\sum_{i}\sum_{j \neq y_i}[X_i \,|\, (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \eta W_{y_i}$$

$$\nabla_{W_j} = \frac{1}{N}\sum_{i}\sum_{j \neq y_i}[X_i \,|\, (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \xi W_j$$

Optimizing Loss Function

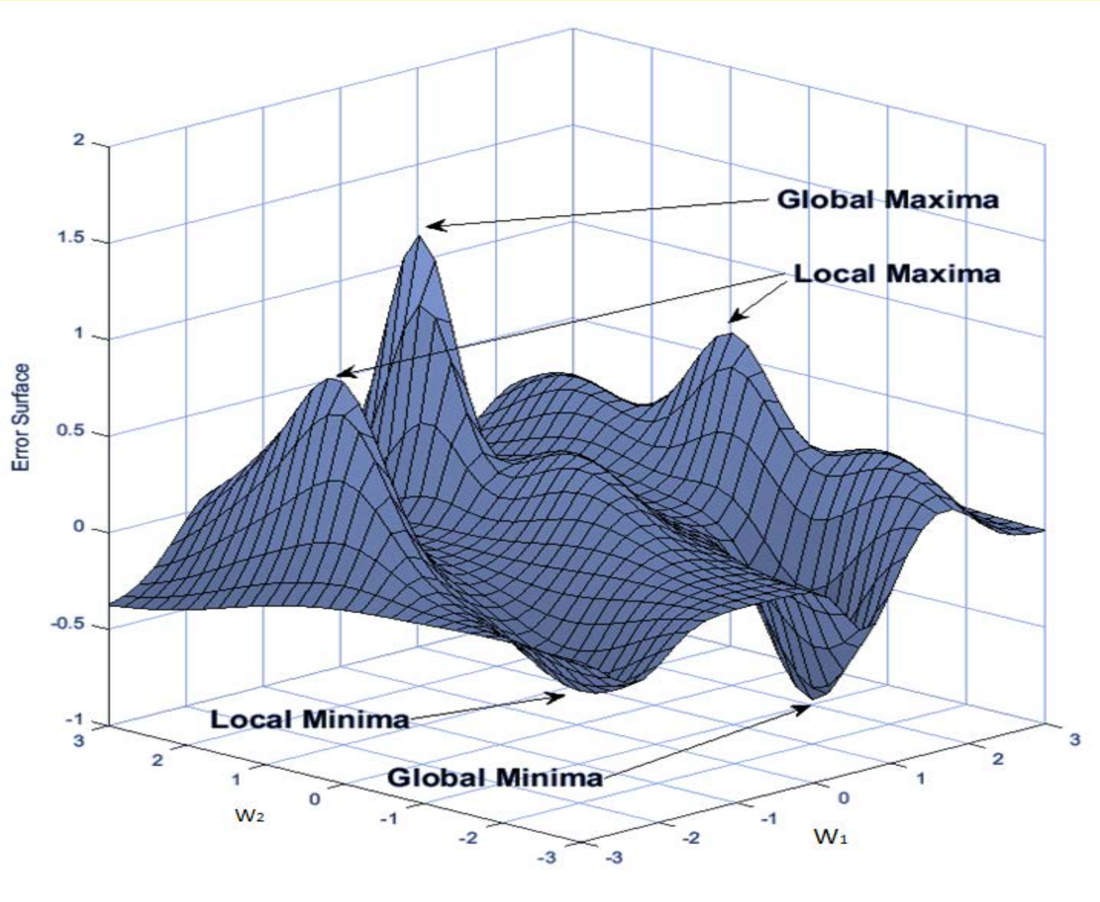
$$\nabla_{W_{y_i}} = -\frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \eta W_{y_i} \qquad \nabla_{W_j} = \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)] + \xi W_j$$

Gradient descent

$$W_{y_i}(k+1) \leftarrow (1-\eta) W_{y_i}(k) + \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)]$$

$$W_j(k+1) = (1-\xi) W_j(k) - \frac{1}{N}\sum_i \sum_{j \neq y_i} [X_i \mid (W_j^t X_i - W_{y_i}^t X_i + \nabla > 0)]$$

Local and Global Minima



Stochastic/ Batch/ Mini batch Optimization

Stochastic Gradient Descent

Upsides

- The frequent updates immediately give an insight into the performance of the model and the rate of improvement.
- This variant of gradient descent may be the simplest to understand and implement.
- The increased model update frequency can result in faster learning on some problems.
- The noisy update process can allow the model to avoid local minima (e.g. premature convergence).

Stochastic Gradient Descent

Downsides

- Updating the model so frequently is more computationally expensive than other configurations of gradient descent, taking significantly longer to train models on large datasets.

- The frequent updates can result in a noisy gradient signal, which may cause the model parameters and in turn the model error to jump around (have a higher variance over training epochs).

- The noisy learning process down the error gradient can also make it hard for the algorithm to settle on an error minimum for the model.

Batch Gradient Descent

Upsides

- Fewer updates to the model means this variant of gradient descent is more computationally efficient than stochastic gradient descent.
- The decreased update frequency results in a more stable error gradient and may result in a more stable convergence on some problems.
- The separation of the calculation of prediction errors and the model update lends the algorithm to parallel processing based implementations.

Batch Gradient Descent

Downsides

- The more stable error gradient may result in premature convergence of the model to a less optimal set of parameters.
- The updates at the end of the training epoch require the additional complexity of accumulating prediction errors across all training examples.
- It requires the entire training dataset in memory and available to the algorithm.
- Model updates, and in turn training speed, may become very slow for large datasets.

Mini-Batch Gradient Descent

Upsides

- The model update frequency is higher than batch gradient descent which allows for a more robust convergence, avoiding local minima.

- The batched updates provide a computationally more efficient process than stochastic gradient descent.

- The batching allows both the efficiency of not having all training data in memory and algorithm implementations.

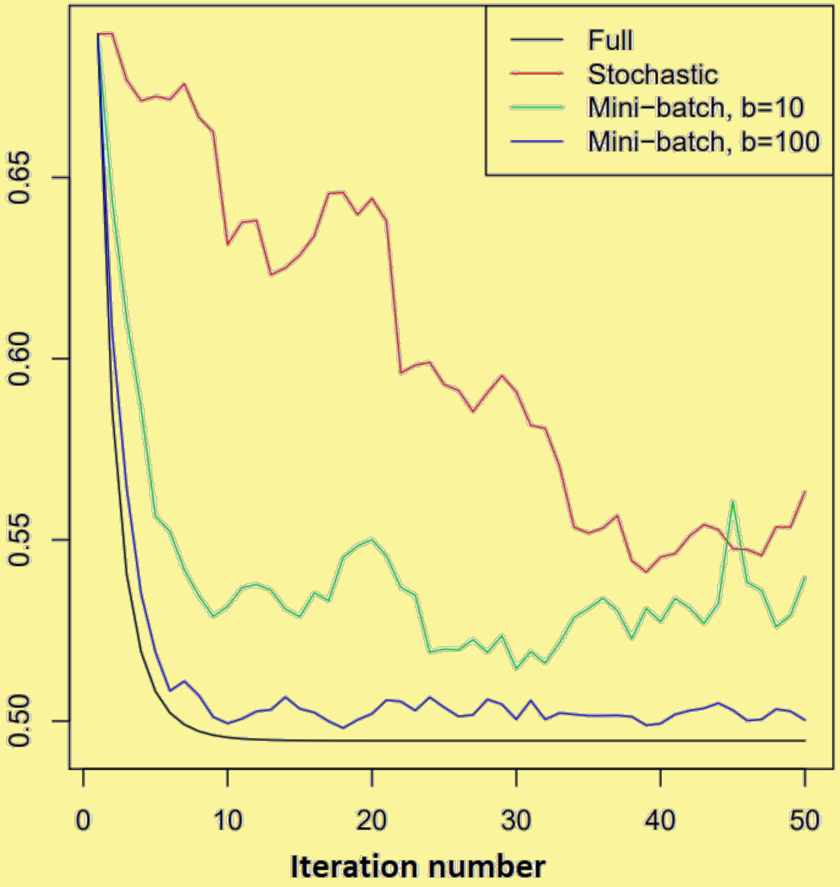
Mini-Batch Gradient Descent

Downsides

- Mini-batch requires the configuration of an additional "mini-batch size" hyper parameter for the learning algorithm.

- Error information must be accumulated across mini-batches of training examples like batch gradient descent.

Error minimization with iterations



NPTEL ONLINE CERTIFICATION COURSES

Thank you