Deep Neural Networks

Irdi Balla

08/10/2017

1 Linear Models

$$XW + b = Y- > softmax(Y) \tag{1}$$

1.1 Disadvantages

- (N + 1)K total parameters. In practice you might want more. (N is the size of X and K the number of labels)
- It's linear
  - can represent inputs that interact in an additive way but not those that interact in a multiplicative way

1.2 Advantages

- GPUs were designed for big matrix multiplications
- Linear Models are numerically stable
  - small changes in input cannot yield big changes in output
  - derivatives are constant

1.3 What we want

- Keep parameters inside big linear functions
- But the entire model should be non-linear

2 Rectified Linear Units (ReLUs)

ReLUs are the simplest non-linear model. Example of a ReLU:

$$y = 0 for x < 0, x for x >= 0 \tag{2}$$

3 2-layer Neural Network

$$XW + b- > ReLu * W + b = Y \tag{3}$$

The first layer is hidden and consists of weights and biases which are then passed through the ReLU. The output is fed to the second layer which again consists of weights and biases applied to the intermediate results.

4 Backward and forward propagation

- Back-propagation: Calculating derivatives by going backwards, starting from the result, and using the chain rule. Back-propagation makes calculating derivatives easy and efficient as long as they are simple.
- Forward-Propagation is running the model on the other direction, from the input to the prediction.

5 Stochastic Gradient Descent (SGD)

For every single batch of training data a forward and a back-propagation is run. This gives us gradients for every weight. They are used to update the weights through a learning rate. After enough repetitions we want our model to increase accuracy. We have to do this while preventing over-fitting. To prevent that we could:

- Termaite Early: stop training when the validation set performance stops increasing
- Regularize: Apply artificial constraints that reduce the number of free parameters without increasing the difficulty in optimization. L2-Regularization is a type of regularization. L2-Reg adds a term to the loss to penalize large weights.

$$L = loss + \beta \frac{1}{2} ||\omega||^2 \tag{4}$$

5.1 Dropout

Dropout is a regularization technique. The values that go from one layer to another are called activations. Using dropout for every train example half of the activations are dropped. This way the model cannot rely on any activation to be present and thus it has to learn a redundant representation. After evaluation the activations are averaged