CS480/680
Lecture 10: June 10, 2019

Multi-layer Neural Networks,

Error Backpropagation

[D] Chapt. 10, [HTF] Chapt. 11, [B] Sec. 5.2, 5.3, [M] Sec. 16.5, [RN] Sec. 18.7

Quick Recap: Linear Models

Linear Regression

Linear Classification

Quick Recap: Non-linear Models

Non-linear classification

Non-linear regression

Non-linear Models

- **Convenient modeling assumption:** linearity
- **Extension:** non-linearity can be obtained by mapping $\boldsymbol{x}$ to a non-linear feature space $\phi(\boldsymbol{x})$
- **Limit:** the basis functions $\phi_i(\boldsymbol{x})$ are chosen a priori and are fixed

- **Question:** can we work with unrestricted non-linear models?

Flexible Non-Linear Models

- Idea 1: Select basis functions that correspond to the training data and retain only a subset of them (e.g., **Support Vector Machines**)

- Idea 2: Learn non-linear basis functions (e.g., **Multi-layer Neural Networks**)

Two-Layer Architecture

- Feed-forward neural network

- Hidden units: $z_j = h_1(\boldsymbol{w}_j^{(1)}\bar{\boldsymbol{x}})$
- Output units: $y_k = h_2(\boldsymbol{w}_k^{(2)}\bar{\boldsymbol{z}})$
- Overall: $y_k = h_2\left(\sum_j w_{kj}^{(2)} h_1\left(\sum_i w_{ji}^{(1)} x_i\right)\right)$

Common activation functions h

- Threshold: $h(a) = \begin{cases} 1 & a \geq 0 \\ -1 & a < 0 \end{cases}$

- Sigmoid: $h(a) = \sigma(a) = \frac{1}{1+e^{-a}}$

- Gaussian: $h(a) = e^{-\frac{1}{2}\left(\frac{a-\mu}{\sigma}\right)^2}$

- Tanh: $h(a) = \tanh(a) = \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$

- Identity: $h(a) = a$

Adaptive non-linear basis functions

- Non-linear regression
  - h_1: non-linear function and h_2: identity

- Non-linear classification
  - h_2: non-linear function and h_2: sigmoid

Weight training

- Parameters: $< \boldsymbol{W}^{(1)}, \boldsymbol{W}^{(2)}, \ldots >$
- Objectives:
  - **Error minimization**
    - **Backpropagation (aka “backprop”)**
  - Maximum likelihood
  - Maximum a posteriori
  - Bayesian learning

Least squared error

- Error function

$$E(\boldsymbol{W}) = \frac{1}{2}\sum_n E_n(\boldsymbol{W})^2 = \frac{1}{2}\sum_n ||f(\boldsymbol{x_n}, \boldsymbol{W}) - y_n||_2^2$$

- When $f(\boldsymbol{x}, \boldsymbol{W}) = \underbrace{\sum_j w_{kj}^{(2)}}_{\text{Linear combo}} \underbrace{\sigma\left(\sum_i w_{ji}^{(1)} x_i\right)}_{\text{Non-linear basis functions}}$

then we are optimizing a linear combination of non-linear basis functions

Sequential Gradient Descent

- For each example $(\boldsymbol{x}_n, y_n)$ adjust the weights as follows:

$$w_{ji} \leftarrow w_{ji} - \eta \frac{\partial E_n}{\partial w_{ji}}$$

- How can we compute the gradient efficiently given an arbitrary network structure?
- Answer: **backpropagation algorithm**
- Today: **automatic differentiation**

Backpropagation Algorithm

- Two phases:
  - Forward phase: compute output z_j of each unit j
  - Backward phase: compute delta δ_j at each unit j

Forward phase

- Propagate inputs forward to compute the output of each unit
- Output z_j at unit j:

$$z_j = h(a_j) \quad \text{where} \quad \mathrm{a}_j = \sum_i w_{ji} z_i$$

Backward phase

- Use chain rule to recursively compute gradient
  - For each weight w_{ji}: $\frac{\partial E_n}{\partial w_{ji}} = \frac{\partial E_n}{\partial a_j}\frac{\partial a_j}{\partial w_{ji}} = \delta_j z_i$
  - Let $\delta_j \equiv \frac{\partial E_n}{\partial a_j}$ then

$$\delta_j = \begin{cases} h'(a_j)\left(z_j - y_j\right) & \text{base case: } j \text{ is an output unit} \\ h'(a_j)\sum_k w_{kj}\delta_k & \text{recursion: } j \text{ is a hidden unit} \end{cases}$$

  - Since $a_j = \sum_i w_{ji} z_i$ then $\frac{\partial a_j}{\partial w_{ji}} = z_i$

Simple Example

- Consider a network with two layers:
  - Hidden nodes: $h(a) = \tanh(a) = \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$
    - Tip: $tanh'(a) = 1 - (tanh(a))^2$
  - Output node: $h(a) = a$

- Objective: squared error

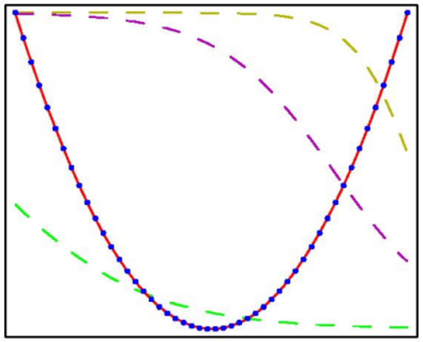
Simple Example

- Forward propagation:
  - Hidden units: $a_j =$ $z_j =$
  - Output units: $a_k =$ $z_k =$
- Backward propagation:
  - Output units: $\delta_k =$
  - Hidden units: $\delta_j =$
- Gradients:
  - Hidden layers: $\frac{\partial E_n}{\partial w_{ji}} =$
  - Output layer: $\frac{\partial E_n}{\partial w_{kj}} =$

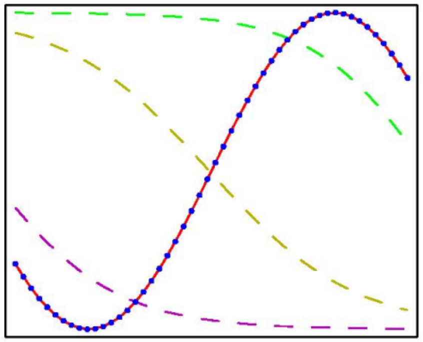
Non-linear regression examples

- Two layer network:
  - 3 tanh hidden units and 1 identity output unit

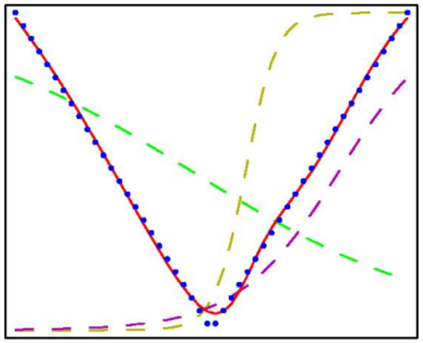
$y = x^2$



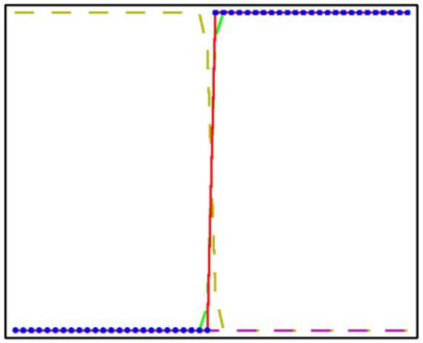
$y = \sin x$



$y = |x|$



$y = \int_{-\infty}^{x} \delta(t) dt$



Analysis

- Efficiency:
  - Fast gradient computation: linear in number of weights
- Convergence:
  - Slow convergence (linear rate)
  - May get trapped in local optima
- Prone to overfitting
  - Solutions: early stopping, regularization (add $||w||_2^2$ penalty term to objective), dropout

Slow convergence

- Gradient direction is not always ideal
- Picture

Adaptive Gradients

- Idea: adjust the learning rate of each dimension separately
- **AdaGrad:**

$$r_t \leftarrow r_{t-1} + \left(\frac{\partial E_n}{\partial w_{ji}}\right)^2 \quad \text{(sum of squares of partial derivative)}$$

$$w_{ji} \leftarrow w_{ji} - \frac{\eta}{\sqrt{r_t}} \frac{\partial E_n}{\partial w_{ji}} \quad \text{(update rule)}$$

- Problem: learning rate $\frac{\eta}{\sqrt{r_t}}$ decays too quickly

RMSprop

- Idea: divide by root mean square (RMS) (instead of root of the sum) of partial derivatives
- **RMSprop:**

$$r_t \leftarrow \boxed{\alpha} r_{t-1} + \boxed{(1-\alpha)} \left(\frac{\partial E_n}{\partial w_{ji}}\right)^2 \quad \text{(where } 0 \le \alpha \le 1\text{)}$$

$$w_{ji} \leftarrow w_{ji} - \frac{\eta}{\sqrt{r_t}} \frac{\partial E_n}{\partial w_{ji}} \quad \text{(update rule)}$$

- Problem: gradient lacks momentum

Adaptive moment estimation

- Idea: replace gradient by its moving average to induce momentum
- **Adam**:

$$r_t \leftarrow \alpha r_{t-1} + (1-\alpha)\left(\frac{\partial E_n}{\partial w_{ji}}\right)^2 \quad \text{(where } 0 \le \alpha \le 1\text{)}$$

$$\boxed{s_t \leftarrow \beta s_{t-1} + (1-\beta)\left(\frac{\partial E_n}{\partial w_{ji}}\right)} \quad \text{(where } 0 \le \beta \le 1\text{)}$$

$$w_{ji} \leftarrow w_{ji} - \frac{\eta}{\sqrt{r_t}} s_t \quad \text{(update rule)}$$

Empirical Comparison

- From Kingma & Ba (ICLR-2015):