Vietnam National University, Ho Chi Minh City

University of Science

Faculty of Information Techonology

Introduction to Machine Learning

Learning Algorithm

Duc Nguyen

November 2, 2022

Contents

Logistic Regression

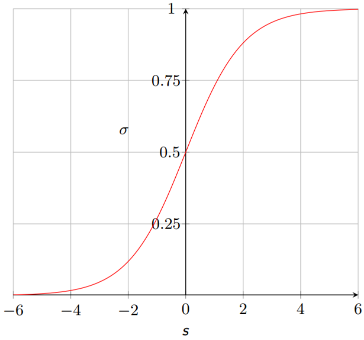
Logistic Function

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

Properties:

$$\sigma(-s) = 1 - \sigma(s)$$

$$\sigma'(s) = \sigma(s)(1 - \sigma(s))$$



Logistic Regression

Problem statement:

- Objective function f is a probability function

$$f : \mathbb{R}^D \to [0, 1]$$

- Hypothesis set $h_{\mathbf{w}}(\mathbf{x}) = \sigma(\mathbf{w}^\mathsf{T}\mathbf{x})$ and probability function

$$P(y|\mathbf{x}, \mathbf{w}) = \begin{cases} h_{\mathbf{w}}(\mathbf{x}) & \text{if } y = 1 \\ 1 - h_{\mathbf{w}}(\mathbf{x}) & \text{if } y = 0 \end{cases}$$

Logistic Regression

Model evaluation

- Likelihood of $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ...(\mathbf{x}_n, y_n)\}$:

$$\prod_{n=1}^{N} P(y_n|\mathbf{x}_n, \mathbf{w})$$

- Maximum a likelihood estimation:

$$\text{Maximize } \prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n, \mathbf{w}\right)$$

$$\Leftrightarrow \text{Minimize } -log \prod_{n=1}^{N} P\left(y_n|\mathbf{x}_n, \mathbf{w}\right)$$

Logistic

- Error function:

$$E(h_{\mathbf{w}}) = -\sum_{n=1}^{N}(y_n log(h_{\mathbf{w}}(\mathbf{x}_n)) + (1 - y_n) log(1 - h_{\mathbf{w}}(\mathbf{x_n})))$$

- **Learning objective:** minimize $E(h_{\mathbf{w}})$
- But how ????

Basic Optimization Problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\texttt{subject to } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}$$

- x is a design point.
- Element in x can be adjusted to minimize the objective function f.
- Any value of x from among all points in the feasible set $\boldsymbol{\mathcal{X}}$ that minimizes the objective function is called a solution.

Local Descent

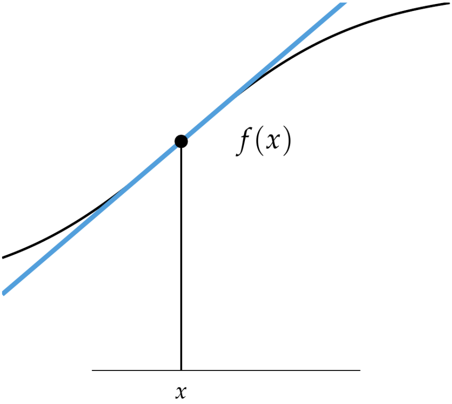
- A common approach for solving an optimization problem is to incrementally improve a design point x by taking a steps that minimizes the objective value based on a local model
  1. Check whether $\boldsymbol{x}^{(k})$ satisfies the terminal conditions.
  2. Determine the descent direction $\boldsymbol{d}^{(k)}$ using local information.
  3. Determine the step size or learning rate $\alpha^{(k)}$
  4. Compute the next design point:

$$\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \alpha^{(k)} \boldsymbol{d}^{(k)}$$

Gradient Descent

Derivatives I

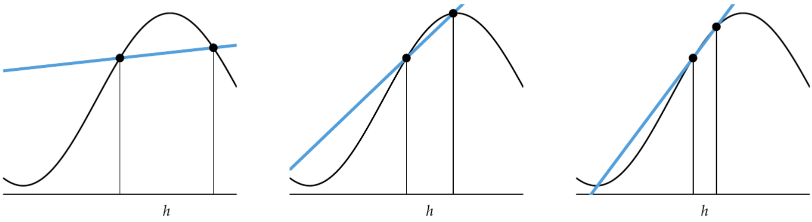
- The derivative $f'(x)$ of a function f of a single variable x is the rate at which the value of f changes at x.
- The value of the derivative equals the slope of the **tangent line**.



Derivatives II

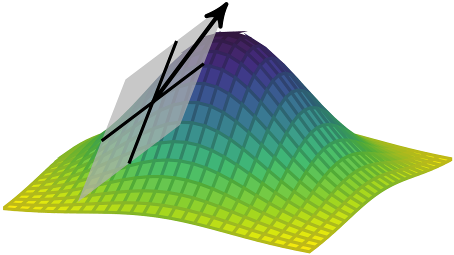
- Derivative can be used provide a linear approximation of the function near x

$$f(x + \Delta x) \approx f(x) + f'(x)\Delta x$$



Derivatives in Multiple Dimensions I

- The **gradient** is the generalization of the derivative to multivariate functions.
- It captures the local slope of the function, allowing us to predict the effect of taking a small step from a point in any direction.



Derivatives in Multiple Dimensions II

- The gradient points in the direction of steepest ascent of the **tagent hyperplane**
- A hyperplane in an n-dimensional space is the set of points that satisfies

$$w_1 x_1 + \dots w_n x_n + w_0 = 0$$

$$\boldsymbol{w}^\mathsf{T} \boldsymbol{x} = 0$$

- The gradient of f at $\boldsymbol{x}$ denoted as $\nabla f(\boldsymbol{x})$ is a vector

$$\nabla f(\boldsymbol{x}) = \left[\frac{\partial f(\boldsymbol{x})}{\partial x_0}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right]$$

Gradient Descent

- An intuitive choice for descent direction $\boldsymbol{d}$ is the steepest descent.
- The direction for steepest descent is the direction opposite the gradient ∇f

$$\boldsymbol{g}^{(k)} = \nabla f\left(\boldsymbol{x}^{(k)}\right)$$

- Typically, we normalize the direction of steepest descent

$$\boldsymbol{d}^{(k)} = -\frac{\boldsymbol{g}^{(k)}}{\|\boldsymbol{g}^{(k)}\|}$$

Logistic Regression (cont.)

Cross-entropy: $J(w) = -\left(ylog(z) + (1-y)log(1-z)\right)$

Chain rule: $\dfrac{\partial J(w)}{\partial w} = \dfrac{\partial J(w)}{\partial z}\dfrac{\partial z}{\partial h}\dfrac{\partial h}{\partial w}$

$$\frac{\partial J(w)}{\partial z} = -\left(\frac{y}{z} - \frac{1-y}{1-z}\right) = \frac{z-y}{z(1-z)}$$

$$\frac{\partial z}{\partial h} = z(1-z), \frac{\partial h}{\partial w} = X \rightarrow \frac{\partial J(w)}{\partial w} = X^T(z-y)$$

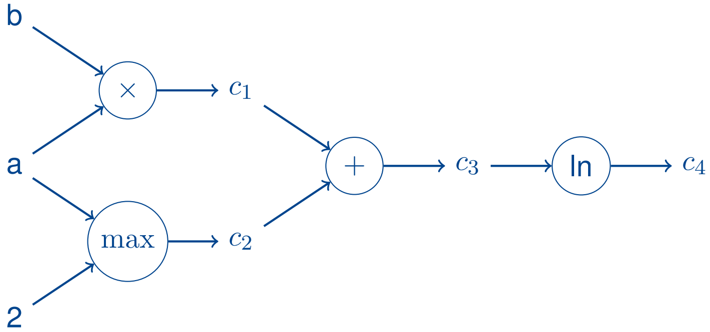
Automatic Differentiation

Automatic Differentiation

- Key to automatic differentiation is the application of Chain rule:

$$\frac{d}{dx}f(g(x)) = \frac{d}{dx}(f \circ g)(x) = \frac{df}{dg}\frac{dg}{dx}$$

Forward Accumulation



$$\frac{df}{dx} = \frac{df}{dc_4}\frac{dc_4}{dx} = \frac{df}{dc_4}\left(\frac{dc_4}{dc_3}\frac{dc_3}{dx}\right) = \frac{df}{dc_4}\left(\frac{dc_4}{dc_3}\left(\frac{dc_3}{dc_2}\frac{dc_2}{dx} + \frac{dc_3}{dc_1}\frac{dc_1}{dx}\right)\right)$$

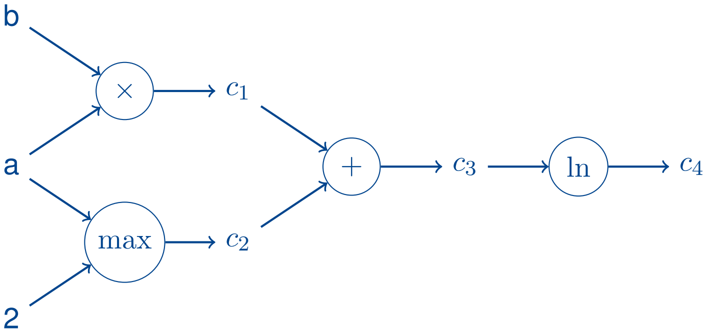
Forward Accumulation

$b = 2$

$\dot{b} = 0$

$a = 3$

$\dot{a} = 1$



Forward Accumulation

$b = 2$

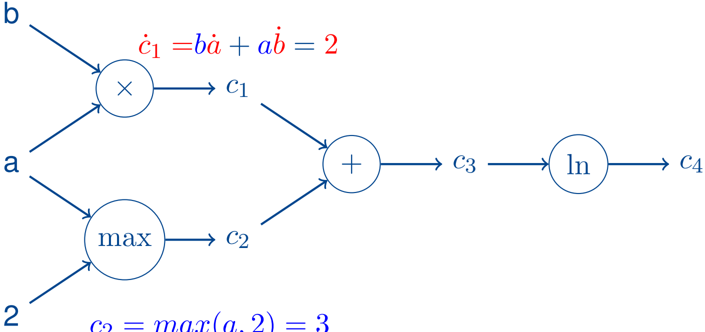
$\dot{b} = 0$

$a = 3$

$\dot{a} = 1$

$c_1 = a \times b = 6$

$\dot{c}_1 = b\dot{a} + a\dot{b} = 2$



$c_2 = max(a, 2) = 3$

$$\dot{c}_2 = \begin{cases} 0, & \text{if } 2 > a \\ 0, & \text{if } 2 < a \end{cases} = 1$$

Forward Accumulation

$b = 2$

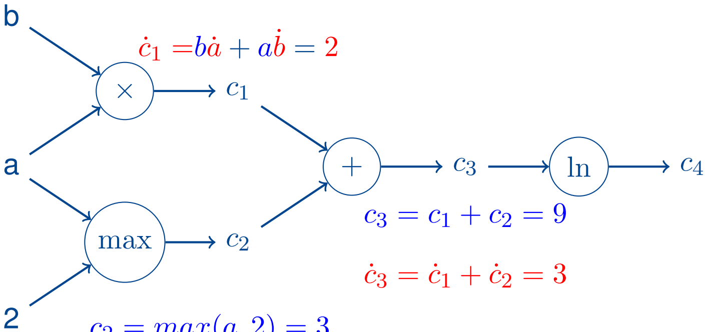
$\dot{b} = 0$

$a = 3$

$\dot{a} = 1$

$c_1 = a \times b = 6$

$\dot{c}_1 = b\dot{a} + a\dot{b} = 2$



$c_3 = c_1 + c_2 = 9$

$\dot{c}_3 = \dot{c}_1 + \dot{c}_2 = 3$

$c_2 = max(a, 2) = 3$

$$\dot{c}_2 = \begin{cases} 0, & \text{if } 2 > a \\ 0, & \text{if } 2 < a \end{cases} = 1$$

Forward Accumulation

$b = 2$

$\dot{b} = 0$

$a = 3$

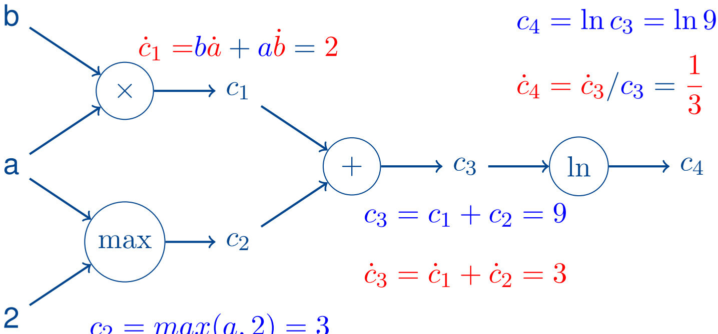
$\dot{a} = 1$

$c_1 = a \times b = 6$

$\dot{c}_1 = b\dot{a} + a\dot{b} = 2$

$c_4 = \ln c_3 = \ln 9$

$\dot{c}_4 = \dot{c}_3 / c_3 = \dfrac{1}{3}$



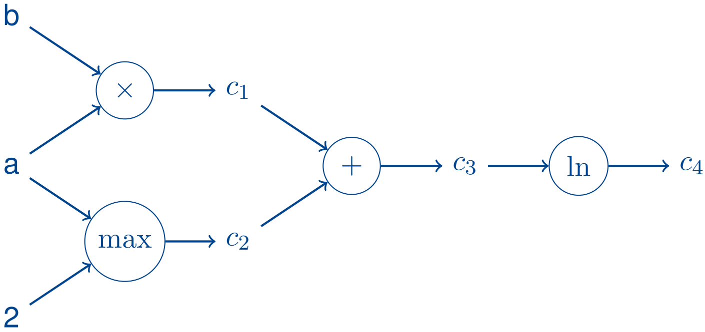
$c_3 = c_1 + c_2 = 9$

$\dot{c}_3 = \dot{c}_1 + \dot{c}_2 = 3$

$c_2 = max(a, 2) = 3$

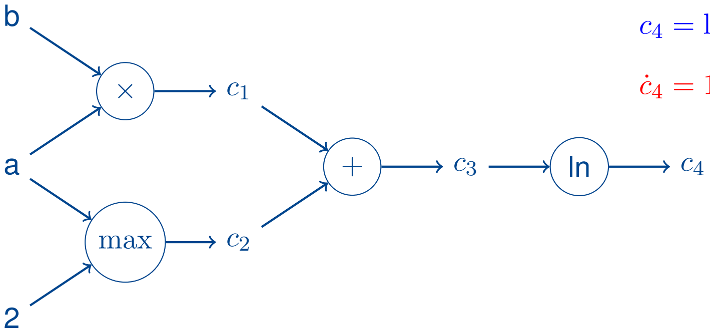
$$\dot{c}_2 = \begin{cases} 0, & \text{if } 2 > a \\ 0, & \text{if } 2 < a \end{cases} = 1$$

Reverse Accumulation



$$\frac{df}{dx} = \frac{df}{dc_4}\frac{dc_4}{dx} = \left(\frac{df}{dc_3}\frac{dc_3}{dc_4}\right)\frac{dc_4}{dx} = \left(\left(\frac{df}{dc_2}\frac{dc_2}{dc_3} + \frac{df}{dc_1}\frac{dc_1}{dc_3}\right)\frac{dc_3}{dc_4}\right)\frac{dc_4}{dx}$$

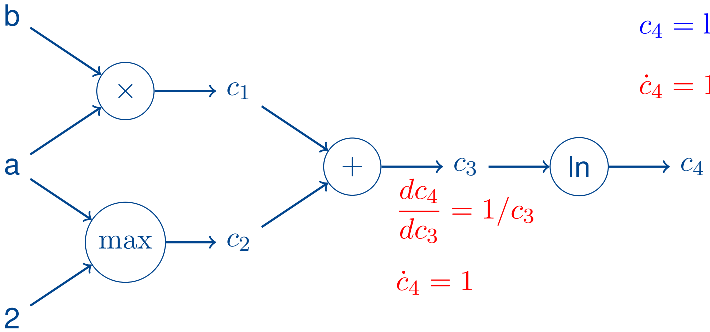
Reverse Accumulation



$c_4 = \ln 9$

$\dot{c}_4 = 1$

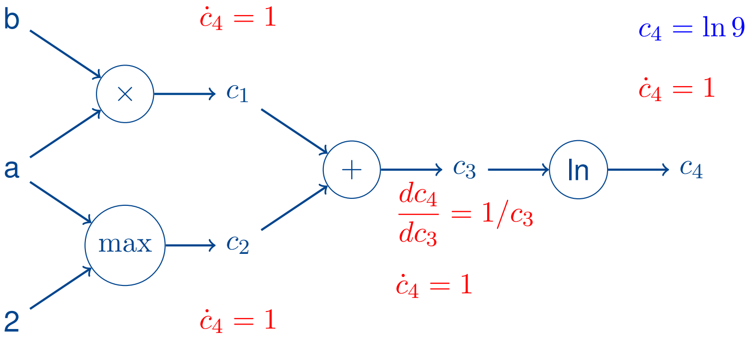
Reverse Accumulation



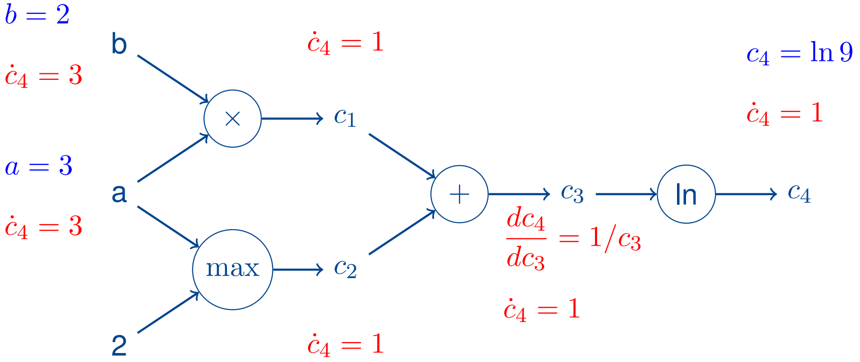
$c_4 = \ln 9$

$\dot{c}_4 = 1$

Reverse Accumulation



Reverse Accumulation



Revision

Revision

- Learning Model: Hypothesis Set, Learning Algorithm.
- Hypothesis Set: **Linear combination** of nonlinear function.
- Learning Algorithm: Gradient Descent.
- Features of training data are the **most important factors** of a learning model.

References

Len Bui.

Course: Introduction to machine learning.

2021.

Gilbert Strang.

Linear algebra and learning from data.

Cambridge Press.