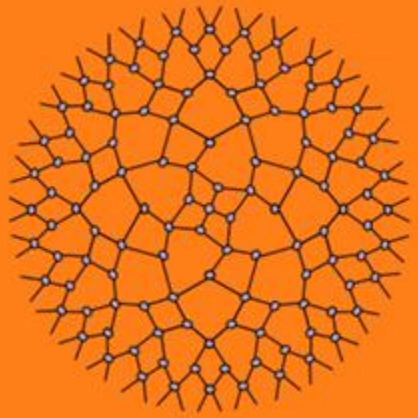
ML Algorithms

NEURAL NETWORKS



Class

A Detailed Look At Neural Networks

Topic

Gradient Descent;
Partial Derivatives

Gradient Descent

Data

Features	Response
$\boldsymbol{x_1} = x_{11}, x_{12}, \dots x_{1\text{m}}$	y_1
$\boldsymbol{x_2} = x_{21}, x_{22}, \dots x_{2\text{m}}$	y_2
. . .	. . .
$\boldsymbol{x_n} = x_{n1}, x_{n2}, \dots x_{\text{nm}}$	y_n

- Parameters $w_0, \; w_1, w_2, \dots., w_m$
- Cost Function: $C = C(\boldsymbol{w})$
- Partial Derivatives: $\dfrac{\partial C(w)}{\partial w_j}$

Partial Derivatives

Partial derivatives are useful for multivariate functions

Functions with multiple variables: A partial derivative of the function with respect to any one of the variables, measures the change in the value of the function for a small change in the value of that variable, keeping all other variables constant

Partial Derivatives

How does a function change when one variable is changed (others remaining fixed)?

<u>Example</u>

- $f(x_1, x_2) = 20 \ + a \, x_1 + b{x_2}^2$
- $\frac{\partial f}{\partial x_1} = a$
- $\frac{\partial f}{\partial x_2} = 2bx_2$
- $\nabla f \ = \binom{a}{2bx_2}$

Partial Derivatives

How does a function change when one variable is changed (others remaining fixed)?

<u>Example</u>

- $f(x_1, x_2) = 20 \; + a\, x_1 + b{x_2}^2$
- $\dfrac{\partial f}{\partial x_1} = a$
- $\dfrac{\partial f}{\partial x_2} = 2bx_2$
- $\boxed{\nabla f \; = \begin{pmatrix} a \\ 2bx_2 \end{pmatrix}}$

Gradient of the multivariate function 'f'

Partial Derivatives for Cost Function

The cost function is the sum of all the costs accrued in all the data points

- $C = \sum_i C_i(\boldsymbol{w})$
- $C_i(\boldsymbol{w}) = y_i \log \frac{1}{1+\exp(\boldsymbol{w}.\boldsymbol{x_i})} - (1 - y_i) \log \left(1 - \frac{1}{1+\exp(\boldsymbol{w}.\boldsymbol{x_i})}\right)$
- $\frac{\partial C(w)}{\partial w_j} = \sum_i (h(x_i) - y_i)\, x_{ij}$

Gradient Descent

- Initialize $\boldsymbol{w}$ to some random values
- Example: $\boldsymbol{w} = [1,4,7,0,\ldots]$
- Repeat
  - ❑ $w_j := w_j - \boldsymbol{\eta}\,\frac{\partial C(w)}{\partial w_j}$ for all $\boldsymbol{j}$
  - ❑ Equivalently $\mathrm{w} := w - \boldsymbol{\eta}\nabla C$ ($\boldsymbol{\eta}$: Learning Rate)
- Once initialized, the weights are continuously updated
- Updating=(Current value of the weights) - (quantity)

Gradient Descent

Updating=(Current value of the weights) — (**quantity**)

Eta/learning rate/step size

Partial derivative of the cost function

- Small *eta* value=slow process
- Large *eta* value=Algorithm running indefinitely

The gradient descent reaches the global minima of a function if a small value of *eta*

Gradient Descent: Example

Data

x1	x2	y
-2	0.5	0
2.5	-2	1

Initialize weights

$$\boxed{\begin{aligned} w_0 &= -2 \\ w_1 &= 1.5 \\ w_2 &= 3.5 \end{aligned}}$$

Gradient Descent: Example

Data

x1	x2	y	h(x)	Cost
-2	0.5	0	0.037	0.017
2.5	-2	1	0.0052	2.282

Gradient Descent: Example

Data

x1	x2	y	h(x)	Cost
-2	0.5	0	0.037	0.017
2.5	-2	1	0.0052	2.282

- $C([-2, 1.5, 3.5]) = 2.299$
- $\frac{\partial C(w)}{\partial w_0} = (0.037 \; - 0\;) \; X \; 1 + (0.0052 \; - 1) \; \mathrm{X} \; 1$ = 0.96
- $\frac{\partial C(w)}{\partial w_1} = (0.037 \; - 0\;) \; X \; (-2) + (0.0052 \; - 1) \; \mathrm{X} \; 2.5$ = - 2.56
- $\frac{\partial C(w)}{\partial w_2} = (0.037 \; - 0\;) \; X \; 0.5 + (0.0052 \; - 1) \; \mathrm{X} \; (-2)$ = 2.008

Gradient Descent: Example

Gradient Descent: Step 1

- $w_0 \coloneqq -2 - 0.01\, X\, (-0.96) = -1.99$
- $w_1 \coloneqq 1.5 - 0.01\, X\, (-2.56) = 1.53$
- $w_2 \coloneqq 3.5 - 0.01\, X\, (-2.008) = 3.48$

New Value of Cost Function: 0.016 (Verify)

Gradient Descent: Example

Gradient Descent: Step 2

- $w_0 := -2 - 0.01\, X\, (-0.96) = -1.99$
- $w_1 := 1.5 - 0.01\, X\, (-2.56) = 1.53$
- $w_2 := 3.5 - 0.01\, X\, (-2.008) = 3.48$

New Value of Cost Function: 0.016 (Verify)

Gradient Descent: Example

Gradient Descent: Step 3

- $w_0 \coloneqq -2 \ - 0.01\, X\, (-0.96) = -1.99$
- $w_1 \coloneqq 1.5 \ - 0.01\, X\, (-2.56) = 1.53$
- $w_2 \coloneqq 3.5 \ - 0.01\, X\, (-2.008)$ =3.48

New Value of Cost Function: 0.016 (Verify)

Gradient Descent

- Repeat these steps until the weights stop experiencing significant change
- Gradient descent will always find the optimum value
- Gradient descent can be slow
- Algorithms could converge slowly, in spite of reasonable learning rate value
- Alternative: Stochastic Gradient Descent

Recap

- Gradient Descent
- Partial Derivatives
- Gradient Descent: Example

JIGSAW ACADEMY

THE ONLINE SCHOOL OF ANALYTICS