Deep Learning

Vijaya Saradhi

IIT Guwahati

Mon, 14th Sept 2020

Knowledge Representation

Definition

Stored information or models used by a person or a machine to interpret, predict and appropriately respond to the outside world.

Knowledge Representation

Discussion

Knowledge of the world consists of two kinds of information:

- Prior Information the known facts.
- Class related prior information example: 20% of emails belong to spam;
- Feature related prior information example 2: 90% of spam emails contain the word "Free Free Free"

Knowledge Representation

Four main points

- Rule 1 Similar inputs from similar classes should produce similar representations inside the network
- Rule 2 Inputs to be categorized as separate classes should be given widely different representation in the network
- Rule 3 Importance to specific features is given throguh involving large number of neurons
- Rule 4 Prior information is achieved through design of neural network.

Introduction

- Obtain best result under given circumstance
- In engineering discipline the goal is to minimize the effort required or maximize the desired benefit
- These are expressed as function of certain decision variables
- Optimization can be defined as the process of finding conditions that gives maximum or minimum value of a function

Introduction

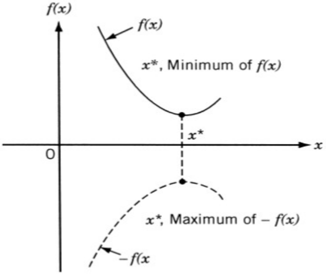
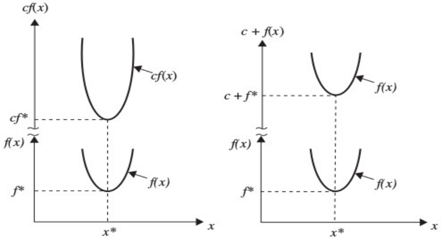


Figure 1.1 Minimum of $f(x)$ is same as maximum of $-f(x)$.



Statement Of Optimization Problem

- Optimization problem

$$
\begin{aligned}
\underset{\mathbf{x}}{minimize}\ f \quad & f(\mathbf{x}) \\
\text{subject to} \quad & g_j(\mathbf{x}) \leq 0 \ \ \forall \ \ j = 1, 2, \cdots, m \\
& l_j(\mathbf{x}) = 0 \ \ \forall \ \ j = 1, 2, \cdots, p
\end{aligned}
$$

- $\mathbf{x}$: Design variables/ design vector
- $f(\mathbf{x})$: objective function
- $g_j(\mathbf{x})$ inequality constraints
- $l_j(\mathbf{x})$ equality constraints
- Constrained optimization problem

Variations

- Design variables:
  - Single variable/Multi-variable
  - Continuous values/integer values
- objective function
  - Linear
  - Non-linear
  - Convex
  - Single objective/multi objective
  - Unimodal/multimodal
- Constraints
  - No constraints
  - only $l_j(.)$ which are linear
  - both $g_j(.)$ and $l_j(.)$
  - Convex

Variables

Single Variable

$$f(x) = (x^2 - 2x + 7)$$

$$f(x) = x^2 + \frac{54}{x}$$

Multi Variable

$$f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$$

$$f(x_1, x_2) = x_1 - x_2 + 2 \times x_1^2 + 2 \times x_1 \times x_2 + x_2^2$$

Continuous vs Integer

Continuous

$$f(x) = x^2 + \frac{54}{x}$$
$$s.t. \;\; x \in \mathbb{R}$$

Continuous

Integer

$$f(x) = (x^2 - 2x + 7)$$
$$s.t. \;\; x \in \mathbb{N}$$

Objective function

Linear

$$\textit{Minimize } f(x_1, x_2) = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\
&a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\
&\vdots \\
&a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m
\end{aligned}$$

Objective function

Matrix Form

Let

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

let

$$\mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

Objective function

Objective function matrix Form

$$Minimize f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$$

Objective function

Constraints in matrix form

Let $\mathbf{A}$ =

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

Let $\mathbf{b}$ =

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

Objective function

Constraints in matrix form

$\mathbf{Ax} = \mathbf{b}$

Objective function

Linear objective

$$\begin{aligned}&Minimize f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}\\&subjec\ to\ \mathbf{A}\mathbf{x} = \mathbf{b}\\&\mathbf{x} \geq 0\end{aligned}$$

Inequality Constraints

Example

$$\text{Minimize } f(x_1, x_2) = x_1^2 + x_2^2$$
$$\text{subject to } x_1 + 2x_2 \leq 15$$
$$1 \leq x_1 \leq 10$$
$$1 \leq x_2 \leq 10$$

Nature of objective functions

- When there are no constraints present the problem is an unconstrained optimization
- When there are constrains present the problem is known as constrained optimization
- Linear Optimization When $f(\mathbf{x})$ is linear and only linear constraints are present
- Non Linear Optimization when $f(\mathbf{x})$ is nonlinear
- Convex Optimization When $f(\mathbf{x})$ is convex and constraints are linear

Optimization Definition

Local optimal

f(x) has a minimum at $x = x^*$ if $f(x^*) \leq f(x^* + h)$ for all sufficiently small positive and negative values of h.
f(x) has a maximum at $x = x^*$ if $f(x^*) \geq f(x^* + h)$ for all sufficiently small positive and negative values of h.

Global optimal

$x = x^*$ found in the interval [a, b] such that x^* minimizes $f(x)$

Introduction

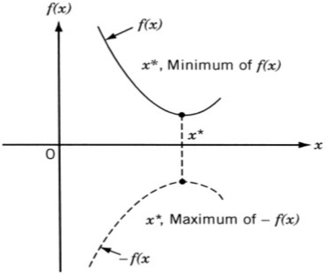
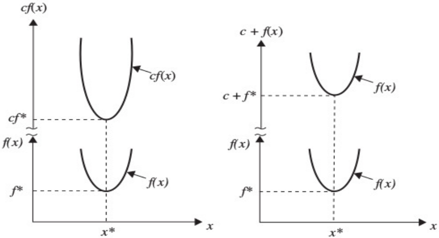


Figure 1.1 Minimum of $f(x)$ is same as maximum of $-f(x)$.



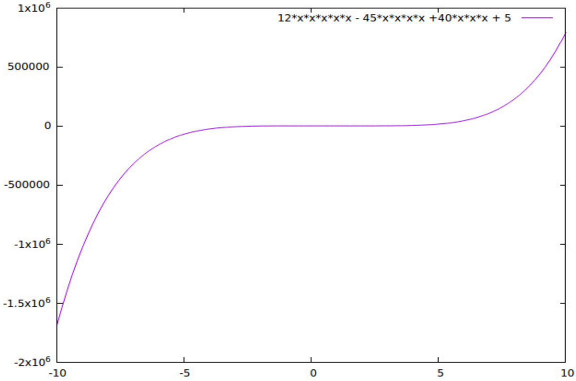
Single Variable

Necessary Condition

if $f(x)$ is defined in the interval [a, b] and has a local minimum at $x = x^*$; let the first order derivative of $f(x)$ exists at $x = x^*$ then

$$\frac{df(x)}{dx} = 0$$

Example


12*x*x*x*x*x - 45*x*x*x*x +40*x*x*x + 5
1x10^6
500000
0
-500000
-1x10^6
-1.5x10^6
-2x10^6
-10
-5
0
5
10

Example

$$f^{'}(x) = 60(x^4 - 3x^3 + 2x^2) = 60x^2(x-1)(x-2)$$

$f^{'}(x) = 0$ at x = 0, 1 and 2.

Multi Variable

Necessary Condition

Let $\mathbf{x} = (x_1, x_2, \cdots, x_n)$

If f($\mathbf{x}$) has a maximum or minimum point at $\mathbf{x} = \mathbf{x}^*$. Assume partial derivatives of $f(\mathbf{x})$ exists at $\mathbf{x}^*$ then

$$\frac{\partial f(\mathbf{x})}{\partial x_1}\bigg|_{x_1=x_1^*} = \frac{\partial f(\mathbf{x})}{\partial x_2}\bigg|_{x_2=x_2^*} = \cdots = \frac{\partial f(\mathbf{x})}{\partial x_n}\bigg|_{x_n=x_n^*} = 0$$

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\bigg|_{\mathbf{x}=\mathbf{x}^*} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}\Bigg|_{\mathbf{x}=\mathbf{x}^*} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

Example

$$f(x_1, x_2) = x_1^3 + x_2^3 + 2x_1^2 + 4x_2^2 + 6$$

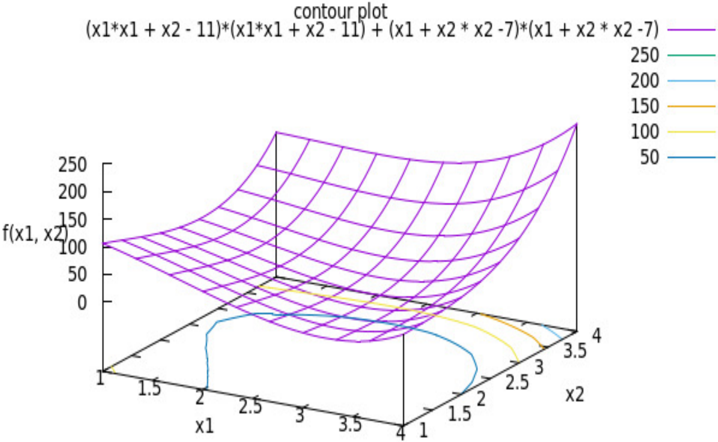
Necessary Condition

$$\frac{\partial f(x_1,x_2)}{\partial x_1} = 3x_1^2 + 4x_1 = x_1(3x_1 + 4) = 0$$

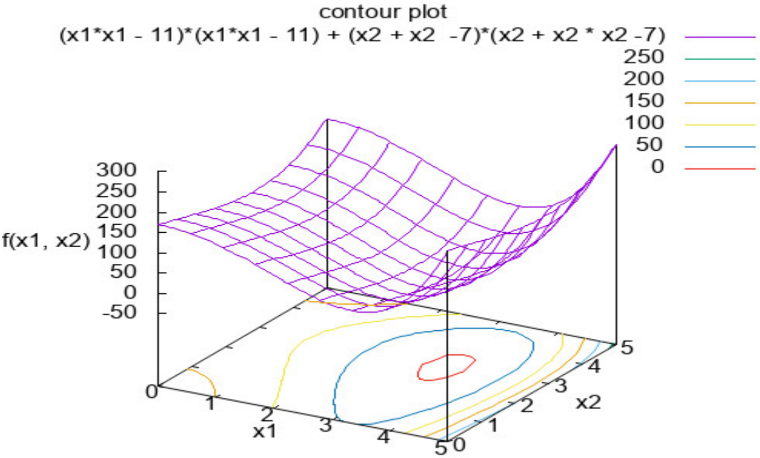
$$\frac{\partial f(x_1,x_2)}{\partial x_2} = 3x_2^2 + 8x_2 = x_2(3x_2 + 8) = 0$$

These equations satisfy at (0, 0), $(0, -\frac{8}{3})$, $(-\frac{4}{3}, 0)$ and $(-\frac{4}{3}, -\frac{8}{3})$

Contours


contour plot
(x1*x1 + x2 - 11)*(x1*x1 + x2 - 11) + (x1 + x2 * x2 -7)*(x1 + x2 * x2 -7)
250
200
150
100
50
f(x1, x2)
250
200
150
100
50
0
1
1.5
2
2.5
3
3.5
4
x1
1
1.5
2
2.5
3
3.5
4
x2

Contours

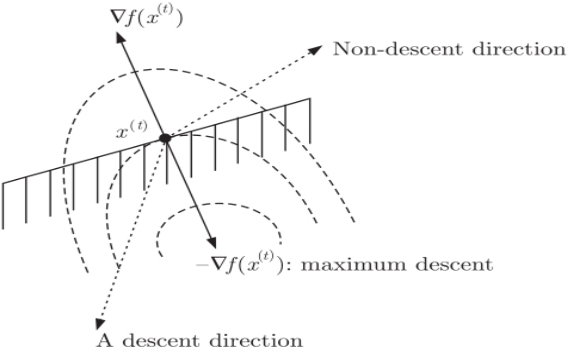


contour plot
(x1*x1 - 11)*(x1*x1 - 11) + (x2 + x2 -7)*(x2 + x2 * x2 -7)
250
200
150
100
50
0
f(x1, x2)
300
250
200
150
100
50
0
-50
0
1
2
3
4
5
x1
0
1
2
3
4
5
x2

Descent Direction

Definition

A search direction $\mathbf{d}^t$ is a descent direction at point $\mathbf{x}^t$ if the condition $\bigtriangledown f(\mathbf{x}^t).\mathbf{d}^t \leq 0$ is satisfied



Descent Direction

Condition

$$
\begin{aligned}
f(\mathbf{x}^{(t+1)}) &< f(\mathbf{x}^{t}) \\
&< f(\mathbf{x}^{t} + \alpha \bigtriangledown f(\mathbf{x}^{t}).\mathbf{d}^{t})
\end{aligned} \tag{1}
$$

That is function value at new point $\mathbf{x}^{(t+1)}$ is less than function value at the current point $\mathbf{x}^{(t)}$

Maximum Descent Direction

Condition

When $\mathbf{d}^t = -\bigtriangledown f(\mathbf{x}^t)$ maxium decrease in function value is obtained
Let $\mathbf{d}^t = (1,0)^T$ Example: $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$

Let $\mathbf{x}^t = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

Let $\mathbf{d}^t = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$\bigtriangledown f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} -46 \\ -38 \end{pmatrix}$

$(-46 - 38)\begin{pmatrix} 1 \\ 0 \end{pmatrix} = -46$

Maximum Descent Direction

Condition

When $\mathbf{d}^t = -\bigtriangledown f(\mathbf{x}^t)$ maxium decrease in function value is obtained

Example: $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$

Let $\mathbf{x}^t = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$

When $\mathbf{d}^t = -\bigtriangledown f(\mathbf{x}^t) = \begin{pmatrix} 46 \\ 38 \end{pmatrix}$

$\bigtriangledown f\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} -46 \\ -38 \end{pmatrix}$

$(-46 - 38)\begin{pmatrix} 46 \\ 38 \end{pmatrix} = -3560$

Gradient Descent

Algorithm

Step 1 Choose: No. of iterations, $\mathbf{x}^{(0)}, \epsilon_1, \epsilon_2$; set $k = 0$

Step 2 Calculate $\triangledown f(\mathbf{x}^{(k)})$

Step 3 if $\| \triangledown f(\mathbf{x}^{(k)})\| \leq \epsilon_1$ then *terminate*

Step 4 Perform *uni-directional search* to find $\alpha^{(k)}$ using ϵ_2

- such that $f(\mathbf{x}^{(k+1)}) = f(\mathbf{x}^{(k)} - \alpha^{(k)} \triangledown f(\mathbf{x}^{(k)}))$ is minimum
- Terminate when $\triangledown f(\mathbf{x}^{(k+1)}). \triangledown f(\mathbf{x}^{(k)}) \leq \epsilon_2$

Step 5 Increment k = k + 1; Repeat steps 2 to 5

Gradient Descent

Example

minimize. $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 + 7)^2$

Example

Step 1 Let k $= 0$; $\mathbf{x}^0 = (0, 0)^T$; $\epsilon_1 = \epsilon_2 = 0.001$

Gradient Descent

Example

minimize. $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 + 7)^2$

Example

Step 1 Let k $= 0$; $\mathbf{x}^0 = (0, 0)^T$; $\epsilon_1 = \epsilon_2 = 0.001$

Step 2 $\bigtriangledown f(\mathbf{x}^{(0)}) = (-14, -22)^T$;
$\| \bigtriangledown f(\mathbf{x}^{(0)}) \| = ((-14)^2 + (-22)^2) = 680 > \epsilon_1$

Gradient Descent

Example

minimize. $f(x_1, x_2) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 + 7)^2$

Example

Step 1 Let k $= 0$; $\mathbf{x}^0 = (0,0)^T$; $\epsilon_1 = \epsilon_2 = 0.001$

Step 2 $\bigtriangledown f(\mathbf{x}^{(0)}) = (-14, -22)^T$;
$\| \bigtriangledown f(\mathbf{x}^{(0)})\| = ((-14)^2 + (-22)^2) = 680 > \epsilon_1$

Step 4 In the direction $- \bigtriangledown f(\mathbf{x}^{(0)})$ perform unidirection search

- Steepest descent direction vector is: $(14, 22)^T$
- Find α^0 such that $f(\mathbf{x}^1) = f(\mathbf{x}^0 - \alpha^0 \bigtriangledown f(\mathbf{x}^{(0)}))$ is minimum
- Let us compute: $\mathbf{x}^1 = \mathbf{x}^0 - \alpha^0 \bigtriangledown f(\mathbf{x}^{(0)})$

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} - \alpha^0 \times \begin{pmatrix} -14 \\ -22 \end{pmatrix} = \begin{pmatrix} 14\alpha^0 \\ 22\alpha^0 \end{pmatrix}$$

Gradient Descent

Example

Step 4 To find α^0, minimize the function $f\mathbf{x}^1$

- We have computed

$$\mathbf{x}^1 = \begin{pmatrix} 14\alpha^0 \\ 22\alpha^0 \end{pmatrix}$$

- Therefore

$$f(\mathbf{x}^1) = f\begin{pmatrix} 14\alpha^0 \\ 22\alpha^0 \end{pmatrix}$$

- Substituting in objective function $f(x_1, x_2) = (x_1^{+} x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2$ we have:
- $f(\mathbf{x}^1) =$ $((14\alpha^0)^2 + (22\alpha^0) - 11)^2 + ((14\alpha^0) + (22\alpha^0)^2 - 7)^2$
- Minimize $f(\mathbf{x}^1)$ to find best α^0

Gradient Descent

Example

Step 4 Using Golden section search or any other single variable optimization procedure we obtain $\alpha^0 = 0.127$. Compute $\mathbf{x}^1 = (\mathbf{x}^0 - \alpha^0 \bigtriangledown f(\mathbf{x}^0)) = (14\alpha^0, 22\alpha^0) = (1.788, 2.810)^T$

Step 4 Since the termination condition does not satisfy

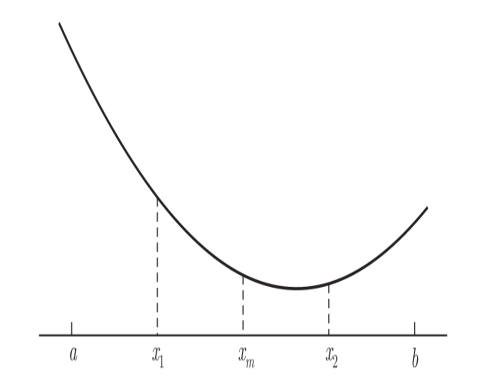
- Terminate when $\bigtriangledown f(\mathbf{x}^{(1)}). \bigtriangledown f(\mathbf{x}^{(0)}) \leq \epsilon_2$
- $\bigtriangledown f(\mathbf{x}^{(1)}) = (30.707, -18.803)^T$
- $\bigtriangledown f(\mathbf{x}^{(0)}) = (-14, -22)^T$
- $$(30.707, -18.803)\begin{pmatrix} -14 \\ -22 \end{pmatrix} \leq \epsilon_2?$$

Step 5 increment k = k + 1; that is k = 1; Repeat the algorithm until termination criteria is met

The optimization obtains $\mathbf{x}^*$ as $(3.008, 1.999)^T$

Single variable optimization

Interval halving method



Single variable optimization

Interval halving method

- Given interval (a, b)
- If $f(x_1) < f(x_m)$ then minimum cannot lie beyond x_m
  That is $f(x_1) < f(x_{m+1}) < \cdots < f(b)$
- The interval will reduce to (a, x_m)
- If $f(x_1) > f(x_m)$ then minimum cannot lie in (a, x_1)

Single variable optimization

Interval halving method - algorithm

Step 1 Given interval (a, b), choose ϵ. Let $x_m = \frac{(a+b)}{2}$; $L = (b - a)$

Step 2 Initialize $x_1 = a + \frac{L}{4}$; $x_2 = b - \frac{L}{4}$; Compute $f(x_1), f(x_2)$

Step 3 If $f(x_1) < f(x_m)$ then $b = x_m; x_m = x_1$; Go to step 5; else go to step 4

Step 4 If $f(x_2) < f(x_m)$ then $a = x_m; x_m = x_2$; Go to step 5; else $a = x_1, b = x_2$; go to step 5;

Step 5 Calculate $L = (b - a)$. If ($|L| < \epsilon$) terminate else go to step 2

Text books to read

Optimization

- Engineering Optimization - Theory and Practice Singiresu S Rao
- Chapter 1 of the above book, sections 6.8 and 6.9
- `mec.nit.ac.ir/file_part/master_doc/20149281833165301436305785.pdf`
- Optimization for Engineering Design Kalyanmoy Deb
- Section 3.4 of the above book.