Supervised Learning

data   labels

$$\left\{ \begin{array}{ll} x^1 & y^1 \\ x^2 & y^2 \\ \vdots & \vdots \\ x^m & y^m \end{array} \right. \qquad x^i \quad y^i \qquad X \quad Y$$

$y \in$ discrete space, finite → classification

$y \in$ continuous space, → regression

Regression

House dataset: { Area, # rooms } → Price ($)

Our data is 2-dimensional → $X \in \mathbb{R}^2$

Linear Regression

$X \to Y$

$h(X) = Y$

Assume h is a linear function.

x_1 : feature 1

x_2 : feature 2   Linear equation

$h(x^{(i)}) = \theta_1 x_1^i + \theta_2 x_2^i + \theta_0 1$

Assumption about the model

↳ Inductive bias

$y = mx + b$

$h(x^{(i)}) = \theta_0 + \theta_1 x_1^i + \theta_2 x_2^i$

$= \theta_0 + \sum_{i=1}^{2} \theta_i x_i$

$y = \theta_1 x_1 + \theta_0$

n data points

d dimensions

$x_0 = 1$

$h(x^{(i)}) = \sum_{i=0}^{d} \theta_i x_i$ } model

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \quad x^{(i)} = \begin{bmatrix} x_0^i \\ x_1^i \\ \vdots \\ x_d^i \end{bmatrix}$

$= \theta^T x^{(i)}$

$h(x^1) = \theta^T x^1$

$h(x^2) = \theta^T x^2$

$h(x^m) = \theta^T x^m$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \qquad X = \begin{bmatrix} x^1 & x^2 & \cdots & x^m \end{bmatrix} \; d$

$d \times 1 \qquad d \times m$

$h(X) = \theta^T X = 1 \times m$

$[\hat{y}^1 \; \hat{y}^2 \; \cdots]$

Single data point → $x^{(i)}$

$h(x)$ y

→ $(h(x) - y)^2$

n data points

$(h(x^1) - y^1)^2$

$(h(x^2) - y^2)^2$

$$\text{Loss} = \frac{1}{m} \sum_{i=1}^{m} \left(h(x^{(i)}) - y^{(i)}\right)^2$$

$$J = \frac{1}{m} \sum_{i=1}^{m} \left\{ \left(\sum_{j=0}^{d} \theta_j x_j \right) - y \right\}^2$$

$x^{(i)} \to y^{(i)}$

θ

If we find θ that minimizes $J(\theta)$ we have the best linear model for that dataset.

θ set θ : min $J(\theta)$

$x^{(i)} = [x_0 \; x_1 \; \cdots \; x_d]^T \quad d \times 1$

$\theta = [\theta_0 \; \cdots \; \theta_d]^T \quad d \times 1$

$h(x) = \theta^T x = \sum_{i=0}^{d} \theta_i x_i \qquad x_0 = 1$

$J(\theta) = (h(x) - y)^2$

$\theta^* \to \min J(\theta)$

$\theta \to$ random

Slope: How steep?

error

direction of max increase

Take the opposite route

$\{ \; \theta := \theta - \nabla_\theta J(\theta)$

$\theta - \nabla$

$f(x)$

$f'(x)$

$-10 \quad x = 10$

$x = 0 \quad y = x^2$

(1)

$J(\theta) = \left\{ \left(\sum_{i=0}^{d} \theta_i x_i \right) - y \right\}^2$

$[\quad \ldots$

$\frac{\partial}{\partial \theta_0} J(\theta) \quad \frac{\partial}{\partial \theta_1} J(\theta_1) \quad \frac{\partial}{\partial \theta_d} J(\theta)$

$f(x) = x^2$

$f'(x) = 2x$

$x = 10$

$f'(x) = 2x = 20$

$$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{\partial}{\partial \theta_i} \left\{ \left(\sum_{i=0}^{d} \theta_i x_i \right) - y \right\}^2$$

$$= 2\left\{ \left(\sum_{i=0}^{d} \theta_i x_i \right) - y \right\} \frac{\partial}{\partial \theta_i} \left\{ \left(\sum_{i=0}^{d} \theta_i x_i \right) - y \right\}$$

$$= \cdots \frac{\partial}{\partial \theta_i} \left(\theta_0 x_0 + \theta_1 x_1 + \cdots + \boxed{\theta_i x_i} + \cdots \theta_d x_d \right) - 0$$

$$= 2\left\{ \left(\sum_{i=0}^{d} \theta_i x_i \right) - y \right\} (x_i - 0)$$

$x = x - \frac{f'(x)}{10}$

$= 10 - 2$

$= 8$

$f'(x) = 2x = 4$

0.4

$\frac{\partial}{\partial \theta_i} \theta_0 x_0 = 0$

$\frac{\partial}{\partial \theta_i} \theta_1 x_1 = 0$

$\frac{\partial}{\partial \theta_i} \theta_i x_i = x_i$

$\frac{\partial}{\partial \theta_i} J(\theta) = 2(h(x) - y) x_i$

Equation (1)

$\theta_i := \theta_i - \frac{\partial}{\partial \theta_i} J(\theta)$

$\theta_i := \theta_i - 2(h(x) - y) x_i$

$J(\theta) = \frac{1}{2}(h(x) - y)^2$

$$\boxed{\theta_i := \theta_i - (h(x) - y) x_i}$$

Update rule

$\theta^T x$

Compute in parallel

$\theta_0 \to \theta_0'$

x^1

θ_1

$\vdots$

θ_d

$X = [\; x^0 \; x^1 \; x \; \cdots \; x^n \;]$

Update all θ_i's.

How many times?

1. Converge $\to \nabla J(\theta) = 0$

$\theta_i := \theta_i - \alpha (h(x) - y) x_i$

↳ Learning rate → 0.1, 0.001

Approach 1 : Batch gradient descent

Repeat for all data points

$\theta_i^0 := \ldots$

$\theta_i^1 := \ldots$

$\vdots$

$\theta_i^n = \ldots$

n items

1 update

average the updates

1 mill

1 update

$\theta_i := \theta_i -$ avg update

Approach 2 : Stochastic GD

$\theta_i^0 = \ldots$ ← Apply update value

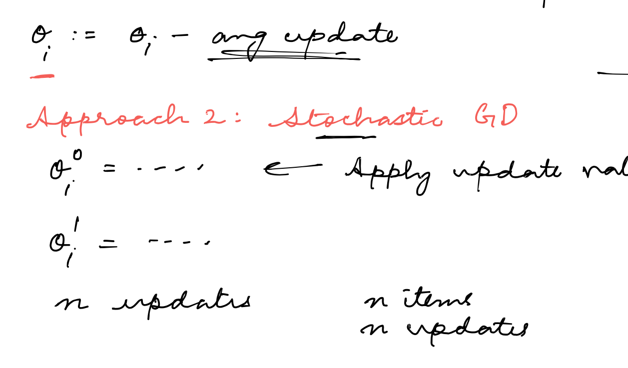
$\theta_i^1 = \ldots$

n updates

n items

n updates

Performance / Speed



Approach 3 : Mini-Batch gradient descent

$d \begin{bmatrix} \xrightarrow{n} \end{bmatrix} \qquad d \left[\overset{m}{[\;]} \; \overset{m}{[\;]} \; \overset{m}{[\;]} \right]$

$m << n$

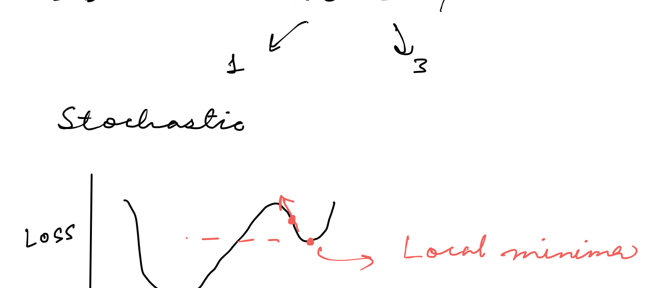
Pick one mini batch

Batch gradient descent

3 batches → n data points

1   3

Stochastic



LR → Guarantee → $J(\theta)$ → Convex function

$\left. \begin{array}{l} \text{\# Local minima} = 0 \\ \text{\# global minima} = 1 \end{array} \right\}$

$f(x)$ — x_1 x_2

$x^2 \to$

x_1 x_2

$\forall x \in [x_1, x_2]$

$f(x) <$

$e^x =$

$J(\theta)$ → Convex function

$f(x) = x^2 \to \forall x \to f''(x) \geq 0$

↓ Hessian

$J(\theta)$

↓ vector

$f''(\ldots) = [\quad]$

↓

Hessian → Positive semi-definite

$J(\theta) = \frac{1}{2}(\theta^T x - y)^2$

↳ 1 global minima

↳ GD → $\alpha \uparrow$

→ LR model → $\theta^T x$

→ Loss → $J(\theta)$

→ min $J(\theta)$ → Trained model

Generalization → Unseen data

dataset → Training

Training	Validation	Test set
80%	10%	10%

Performing good on val data

Test data

Industry   Research → Test

Comment → LR

$\theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_d x_d$

Polynomial regression

$x_0^2 \quad x_0^3$

$\theta_1 x_0^2 \qquad \theta_1 x_1^2$

$1 \qquad\quad 0$

$h(x) = x_0^2$

8:27

LR → $J(\theta)$ → Directly

$f(x)$

minima

$\left. \begin{array}{l} \nabla_\theta J(\theta) = 0 \\ \theta \end{array} \right\} (X^T X)$

1 million

$f'(x) = 0, \; f''(x) > 0$

x^*