Lecture 12 08.10.21

Admin

- Last class today. ~ GD for losses.
  HW coming
- No class next week.
- New topics & projects following week

SO regression / binary class

— analysis. $\nabla \hat{L}(w) = 0$

— optimization $w_{k+1} = w_k - \eta \nabla \hat{L}(w_k)$

to find w^* $h_w(x)$

$h_w(x)$ "best" model

Major topic

Multiclass: ✓

Generalization

$h_w(x)$ $\hat{L}(w) = 0$ fit all data perfectly

what about new data?

$\Rightarrow$ Statistical learning theory.

Defining other Important ML - Deep AI problems.

contrastive loss pairs

$s(x_1, x_2)$

$d(x_1, x_3)$

learn $f(x)$

s.t. $f(x) \cdot f(x') \geq f(x) \cdot f(x'')$

when $x \sim x'$

$x \nsim x''$

NLP >5 year

contrastive loss

< 2 year

ImageNet

current

$$pmi = \log\left(\frac{p_{ij}}{p_i p_j}\right)$$

count word pair freq p_{ij}

ind freq p_i, p_j

Solve $f(x_i) \cdot f(x_j) = \log\left(\frac{p_{ij}}{p_i p_j}\right)$

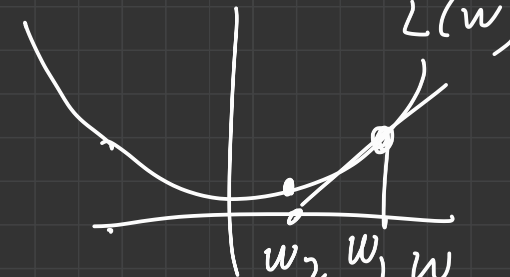
similarity loss

Chain Rule

$$h_w(x)$$

$$\ell(h, y)$$

$$\hat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(h_w(x_i), y_i)$$



Studied $\quad \nabla \hat{L}(w) = 0$

Now GD $\quad w_{k+1} = w_k - h \nabla \hat{L}(w_k)$

(with stepsize/ learning rate h)

Chain Rule $\quad \nabla_w \hat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\frac{\partial \ell}{\partial h}(h_w(x_i), y_i)}_{\text{loss}'} \underbrace{\nabla_w h_w(x_i)}_{\text{model gradient}}$

$h_w(x) = w \cdot x \quad \nabla_w h_w = x$

Abstract GD:

$\hat{L}(w)$ μ-convex & L-smooth

$H(w) = D^2 L(w)$

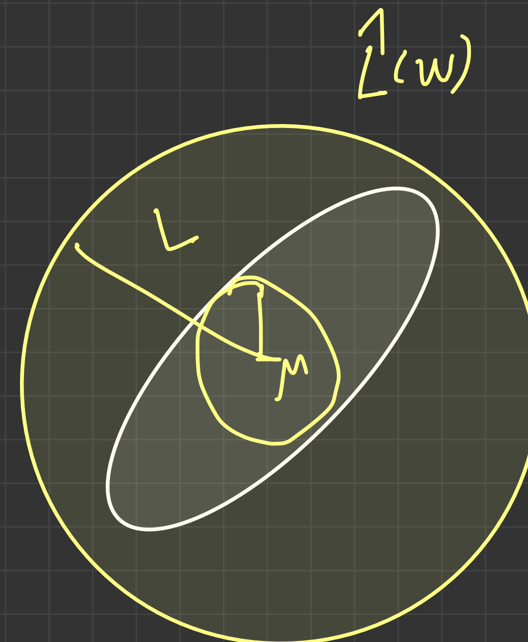
<u>2 Defn</u>

$$\mu I \preceq H(w) \preceq L I$$

M, N p.s.d. matrix

$M \preceq N \iff x^T M x \le x^T N x \ \forall x$

all eigenvalues of H are between μ and L



$$w_{k+1} = w_k - h \nabla L(w_k)$$

cond number of convex fn

$$C = \frac{L}{\mu}$$

convergence rate of GD

worst case analysis

$h = \frac{1}{L}$ rate. $e_k \le C^k e_0$

$f^* = \min_x f(x)$ $e_k = f(x_k) - f^*$

$\hat{L}^* = \min_w \hat{L}(w)$ $e_t = \hat{L}(w_t) - \hat{L}^*$

Line seach ...

$$w_{k+1} = w_k - h \nabla L(w_k)$$

Look for better.

Loss gradients.

① Threshold models

$S_m = \{y_1, \ldots, y_m\}$

$h_w = w$

$= \ell_2(\underbrace{h-y}_{e}) = (h-y)^2/2 = e^2/2$

$\ell_2'(e)$

$\hat{L}(w) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, y_i)$

$= \frac{1}{m}\sum_{i=1}^{m} (w-y_i)^2$

$w^* = \bar{y} = \frac{1}{m}\sum_{i=1}^{m}$

$\ell'(e) = e$

$\hat{L}'(w) = \frac{1}{m}\sum_{i=1}^{m} w - y_i$

$= w - \bar{y}$

<u>GD</u>

$w_{k+1} = w_k - h(w_k - \bar{y})$

$= (1-h)w_k + h\bar{y}$

$e_k = w_k - \bar{y}$

<u>case 1</u> $h = 1 \Rightarrow w_1 = \bar{y}$

<u>case 2</u> $h > 1$ diverges.

<u>case 3</u> $h = \frac{1}{2}$ $w_{k+1} - \bar{y} = \frac{1}{2}(w_k - \bar{y}) \Rightarrow \underline{e_k = \frac{1}{2^k} e_0}$

GD2

Linear Model

$x \in \mathbb{R}^d \quad w \in \mathbb{R}^d$

$h_w(x) = w \cdot x \quad \nabla_w h = x$

Chain Rule $\quad \nabla_w \hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\frac{\partial l}{\partial h}(h_w(x_i), y_i)}_{h(x_i) - y_i} \underbrace{\nabla_w h_w(x_i)}_{x_i}$

$l_2 \quad \frac{\partial l}{\partial h} = e = h(x_i) - y_i$

$\nabla_w \hat{L}(w) = \frac{1}{m} \sum_{i=1}^{m} (\underbrace{h(x_i)}_{w \cdot x_i} - \underbrace{y_i}_{w^* \cdot x_i}) x_i$

$y = w^* \cdot x$

$(w - w^*) \cdot x_i$

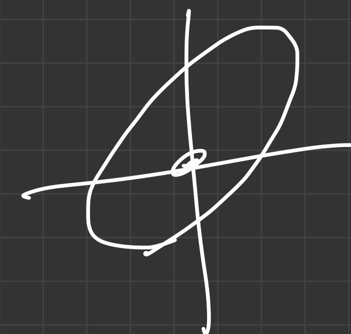
Simplify

...

$\underline{d=2}$ $\quad (x_1, x_2) = (s, 1)$

$$y = ms + b$$

$$(w_1, w_2) = (w_m, w_b)$$



<u>HW</u> $\quad \dfrac{\partial L}{\partial w_b} = (w_b - b) + (w_m - m)\bar{s}$

$$\frac{\partial L}{\partial w_m} = (w_b - b)\bar{s} + (w_m - m)\overline{s^2}$$

$\bar{s} = \frac{1}{m}\sum s_i \qquad \overline{s^2} = \frac{1}{m}\sum s_i^2$

$$\nabla_w L(w) = H(w - w^*)$$

$$= \begin{bmatrix} \overline{s^2} & \bar{s} \\ \bar{s} & 1 \end{bmatrix} [w - w^*]$$