bit.ly/3Pghxme

$$\min_{x \in \mathbb{R}^d} f(x)$$

ML (supervised)

notation: Data (a_i, y_i) $\quad a_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$

$\uparrow$ feature $\quad \uparrow$ label $\qquad \{-1, +1\}$

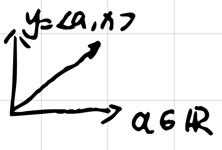
Prediction model : $y_i \approx g_x(a_i)$

x = params of model

Examples: linear $g_x(a) = \langle a, x \rangle$ $\quad a \in \mathbb{R}^d, \ x \in \mathbb{R}^d$

$$\langle a, x \rangle_{\mathbb{R}^d} \overset{\text{def.}}{\triangleq} \sum_i a[i]\, x[i]$$

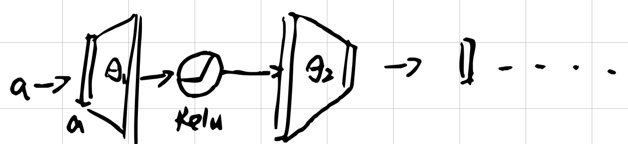
$d = 1$.



with bias : $g_x(a) = \langle a, x^{(1)} \rangle + x^{(2)}$ $\quad x^{(1)} \in \mathbb{R}^d, \ x^{(2)} \in \mathbb{R}$ $\quad x = (x^{(1)}, x^{(2)})$

Neural networks.

MLP



$x = (\theta_1, \theta_2, \theta_3)$

Data: $(a_i, y_i)_{i=1}^n$ $\qquad$ Model : g_x

ERM (Empirical Risk minima...)

$$\min_x \left[\frac{1}{n} \sum_{i=1}^n \ell(g_x(a_i), y_i) \triangleq f(x) \right]$$

param $\uparrow$ $\quad$ input $\uparrow$

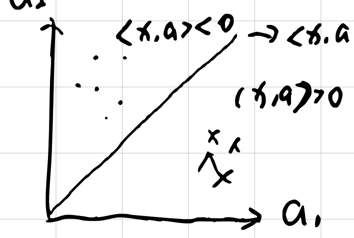
Loss : Regression : $y_i \in \mathbb{R}$

$$\ell(y, y') = (y - y')^2$$

MSE (mean squared error)

Classification loss : $\quad y \in \{-1, +1\}$ (binary classi...)

"Ideal" : $\ell_0(y, y') = \begin{cases} 0 & y = y' \\ 1 & \text{otherwise} \end{cases}$



$g_x(a) = \langle a, x \rangle$

Decision: $\operatorname{sign}(g_x(a)) = \operatorname{sign}(\langle a, x \rangle)$

$$\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{\langle a_i, x \rangle y_i \le 0\} = f(x) \quad \text{not continuous}$$

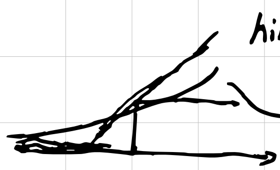
nice function : differentiable, convex ---

$\sigma(t) = \mathbb{1}_{t>0}$

$$\frac{1}{n} \sum_{i=1}^n \sigma(-y_i \langle a_i, x \rangle) = f(x)$$

SVM

logistic regression

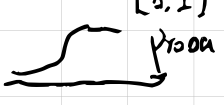


hinge $\sigma_{svm}(t) = \max(t+1, 0)$

logistic $\sigma_{log}(t) = \dfrac{\log(1+e^t)}{\log(2)}$

$\log(1+e^t) \approx t$

$\sigma'(t) = \dfrac{e^t}{1+e^t}$ sigmoid



Matrix notation: Design matrix

$$A = \underbrace{\begin{bmatrix} \quad \\ a_i \\ \quad \end{bmatrix}}_{d} \Big\} n$$

Linear model: $(\langle a_i, x \rangle)_{i=1}^{n} = Ax$

$A@x \longrightarrow$

MSE:

$$\min_x \frac{1}{n} \sum (\overbrace{\langle a_i, x \rangle}^{(Ax)_i} - y_i)^2 = f(x)$$

$$f(x) = \frac{1}{n} \| A \overset{@}{x} - y \|^2$$

$\|z\|^2 = \langle z, z \rangle_{\mathbb{R}^n} = \sum_{i=1}^{n} z_i^2$ np.lina... norm

Logistic: $\min_x \frac{1}{n} \sum_{i=1}^{n} \sigma_{\log}(-y_i \langle x, a_i \rangle)$

$$(y_i \cdot \langle x, a_i \rangle)_{i=1}^{n} = \mathrm{diag}(\overset{\{-1,+1\}^n}{\underset{n \times n}{y}}) \times \underset{n \times d}{A} \times \underset{d \times 1}{x}$$

$\boxed{A @ x * y}$ $y * (A @ x)$

Global loss: $L(z) = \frac{1}{n} \sum_{i=1}^{n} \sigma_{\log}(z_i)$

$$L(z) = \frac{1}{n} \cdots \cdots (\quad)$$

$$= L(\underbrace{-\mathrm{diag}(y) A}_{B} x)$$

$$= L(Bx)$$

$$B = \begin{bmatrix} \quad \\ -y_i a_i \\ \quad \end{bmatrix}$$

Summary:
- MSE $\min_x \frac{1}{n} \|Ax - y\|^2$
- log $\min_x L(Bx)$
- $B = A$ $L(z) = \frac{1}{n} \|z - y\|^2$

Gradient of $f: \mathbb{R}^d \to \mathbb{R}$

Def: $$\nabla f(x) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_d}(x) \end{pmatrix} \in \mathbb{R}^d$$

partial derivatives:

$$\frac{\partial f}{\partial x_k}(x) = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon \delta_k) - f(x)}{\varepsilon}$$

$$\delta_k = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow k$$

Def: f is differentiable at x iff

$\forall u$: $f(x + \varepsilon u) = f(x) + \varepsilon \langle u, \nabla f(x) \rangle + o(\varepsilon)$

$\Leftrightarrow \frac{f(x + \varepsilon u) - f(x)}{\varepsilon} \xrightarrow{\varepsilon \to 0} \langle u, \nabla f(x) \rangle$

⚠ f is diff $\underset{\not\Leftarrow}{\Rightarrow}$ f has a gradient

$$f(x_1, x_2) = \frac{2 x_1 x_2 (x_1 + x_2)}{x_1^2 + x_2^2} \qquad f(0) = 0?$$

f is continuous

$\nabla f(0, 0) = 0$

Prop°: f has a ∇f and ∇f is continuous $\Rightarrow f$ is diff...

Why ∇? Theory "x is a local minimizer $\Rightarrow \nabla f(x) = 0$"

Algorithm "$-\nabla f(x)$ is steepest descent direction"

GD $x_0 \leftarrow$ Init

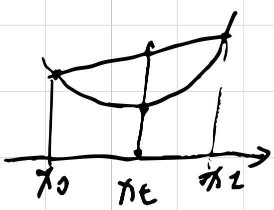
$x_{k+1} = x_k - \tau_k \nabla f(x_k)]$

↳ step size / learning rate

Thm: x^* is local minimizer of f

f is diff at x^* $\Downarrow$ $\nabla f(x^*) = 0$

Convexity:



$x_t = (1-t)x_0 + t x_1 \quad t \in [0,1]$

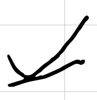
Def: f is convex iff $\forall (x_0, x_1)$, $\forall t \in [0,1]$

$f((1-t)x_0 + t x_1) \leq (1-t) f(x_0) + f(x_1)$

Thm: if f is cvx and diff

x is a minimizer $\Leftrightarrow$ $\nabla f(x^*) = 0$

x is a loc min $\Rightarrow$ x is a global min

- f cvx $\Leftrightarrow$ $f'' \geq 0$
- f is above its tangent 

Prop: f, g are convex, $\lambda, u \geq 0$ $\Rightarrow$ $\lambda f + u g$ cvx

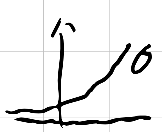
A a matrix, b a vector: f cvx $\Rightarrow$ $f(Ax+b)$ cvx

Example: MSE: $f(x) = \frac{1}{n} \|Ax - y\|^2$

$z \mapsto \|z\|^2 = \sum z_i^2$

Logistic $f(x) = L(Bx)$

$L(z) = \sum \sigma(z_i)$



$\sigma(s) = \log(1 + e^s)$

$\sigma'(s) = \frac{e^s}{1+e} \qquad \sigma''(s) \geq 0$

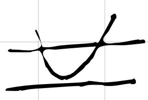
① $f(x) = e^{-x} \geq 0$ no minimizer

$f(x) = x$

Prop: f is continuous

f is " $f(x) \xrightarrow{\|x\| \to +\infty} +\infty$

Then there exists a minimizer

"Proof" 

lower semi-continuous (LSC)

$x_k \to x \qquad \lim f(x_k) \geq f(x)$

Gradient:

$$f(x+\varepsilon u) = f(x) + \varepsilon \langle \nabla f(x), u \rangle + o(\varepsilon)$$

MSE: $f(x) = \|Ax - y\|^2$

$f(x+\varepsilon u) = \|A(x+\varepsilon u) - y\|^2$

$= \|(Ax-y) + \varepsilon Au\|^2$ $\quad \underline{\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2}$

$= \|Ax-y\|^2 + 2\langle Ax-y, \varepsilon Au\rangle + \varepsilon^2 \|Au\|^2$

$= f(x) + \varepsilon \cdot 2\langle Ax-y, Au\rangle_{\mathbb{R}^n} + o(\varepsilon)$

$\overset{?}{=} \cdots \quad \langle \nabla f(x), u \rangle_{\mathbb{R}^d}$

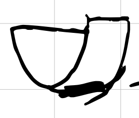
$\langle Aw, z\rangle_{\mathbb{R}^n} = \langle w, A^T z\rangle_{\mathbb{R}^d}$

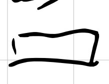
$$= f(x) + \varepsilon \langle \underbrace{2A^T(Ax-y)}_{\nabla f(x)}, u\rangle + o(\varepsilon)$$

Conclu: $f(x) = \|Ax-y\|^2 \quad \nabla f(x) = 2A^T(Ax-y)$

corolary: $\min_{x \in \mathbb{R}} \|Ax-y\|^2$

$\nabla f(x^*) = 0 = 2A^T(Ax-y)$

 may not be unique

$A =$  $\quad \ker(A) = \{z; Az = 0\}$

"under-determined" $\quad d$ big $\quad n$ small

x^* sol $\quad z \in \ker(A) \Rightarrow x^* + z$ sol

$\|Ax-y\|^2 = \|A(x^*+z) - y\|^2$

$[\underbrace{(A^TA)}x^* = A^Ty]$

$\triangleq C \in \mathbb{R}^{d\times d}$ "correlation"

Prop°: $A^TAx = y$ has a unique solution

$\Leftrightarrow \ker(A) = \{0\}$

Proof: $\Leftrightarrow \ker(A^TA) = \{0\}$

$\qquad \overset{\shortparallel}{\ker(A)}$

① $Az = 0 \Rightarrow A^TAz = 0$

② $A^TAz = 0 \quad \langle A^TAz, z\rangle = \langle 0, z\rangle = 0$

$\Downarrow$

$\langle Az, Az\rangle = 0 \Rightarrow \|Az\|^2 = 0 \Rightarrow Az = 0$

$C = A^TA$ symetric semi-definite matrix

$[C^T = C$ diagonizable in ortho-basis

real eigenvalue

if $C = A^TA \Leftrightarrow$ Eigenvalues are ≥ 0

Prop. if overdetermined, $\ker(A) = \{0\}$

$n \gg d$

$x^* = \underline{(A^TA)^{-1} \cdot A^T} y \qquad x^* = A^+ y$

Moore-... Pseudo Inverse

Ridge Penality // weight decay:

$$\min_{x \in \mathbb{R}^d} \|Ax-y\|^2_{\mathbb{R}^n} + \underset{\substack{\downarrow \\ \text{Ridge param}}}{\lambda} \|x\|^2_{\mathbb{R}^d}$$

Select λ: cross validation

$f(x) = \|Ax-y\|^2 + \lambda\|x\|^2$

$\nabla f(x) = 2A^T(Ax-y) + \lambda 2x$

$\nabla f(x^*) = 0 \Leftrightarrow \quad A^TAx + \lambda x = A^Ty$

$(A^TA + \lambda Id)x = A^Ty$

if u_i is eigenvalue of B

$u_i + \lambda \quad \cdots \quad B + \lambda Id$

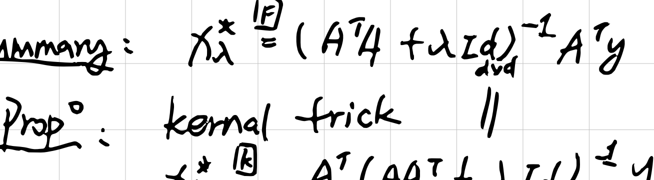
$Bz = u_i z \Rightarrow (B + \lambda Id) = (u_i + \lambda) z$

A^TA has eigenvalue ≥ 0

$A^TA + \lambda Id \qquad \geq \lambda > 0$

$\Rightarrow (A^TA + \lambda Id) z = A^Ty$

have a unique solution



$\|Ax-y\|^2 \quad + \lambda\|u\|^2$

Summary: $x^*_\lambda \overset{\boxed{F}}{=} (A^TA + \lambda Id_{d\times d})^{-1} A^Ty$

Prop°: kernal trick $\quad \|$

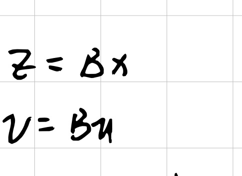
$x^*_\lambda \overset{\boxed{K}}{=} A^T(AA^T + \lambda Id_{n\times n})^{-1} y$

$\boxed{F}$ is better $\quad d < n$

$\boxed{K}$ is better $\quad d > n^6 \quad \boxed{d = +\infty}$

logistic: $f(x) = L(Bx)$ $\quad B \to \mathbb{R}^{n\times d}$

$L(z) = \sum_{i=1}^{n} \sigma(z_i)$ $\quad \uparrow$ diff



$f(x+\varepsilon u) = L(Bx + \varepsilon Bu)$

Taylor exp. of L at point $z = Bx$

$v = Bu$

$L(z+\varepsilon v) = L(z) + \varepsilon\langle \nabla L(z), v\rangle + o(\varepsilon)$

$f(x+\varepsilon u) = L(Bx) + \varepsilon\langle \nabla L(Bx), Bu\rangle + o(\varepsilon)$

$f(x+\varepsilon u) = f(x) + \varepsilon\langle \underbrace{B^T \nabla L(Bx)}_{\nabla f(x)}, u\rangle + o(\varepsilon)$

Prop: $f(x) = L(Bx) \qquad \nabla f(x) = B^T \nabla L(Bx)$

$\nabla(L \circ B) = B^T \circ \nabla L \circ B$

example logistic: $L(z) = \sum_{i=1}^{n} \sigma(z_i)$

Jacobian : $F : \mathbb{R}^d \to \mathbb{R}^p$

$$x \longmapsto f(x)$$

Def: $\partial F(x) \in \mathbb{R}^{p \times d}$

$$\partial F(x) : \mathbb{R}^d \to \mathbb{R}^p \quad \text{Linear}$$

$$u \to \partial F(x) \times u$$

$$F(x) = \begin{pmatrix} F_1(x) \\ F_2(x) \\ \vdots \\ F_p(x) \end{pmatrix} = (F_i(x))_{i=1}^{p}$$

Ex. $f(x) = \begin{pmatrix} x_1^2 + x_2^2 \\ x_1 x_2 \end{pmatrix}$ $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\partial f(x) = \begin{pmatrix} 2x_1 & 2x_2 \\ x_2 & x_1 \end{pmatrix}$$

$$\partial F(x) \triangleq \left(\frac{\partial F_i}{\partial x_j}(x) \right)_{\substack{i = 1 \dots p \\ j = 1 \dots d}} \in \mathbb{R}^{p \times d}$$

⚠ $p = 1$ $\quad F : \mathbb{R}^d \to \mathbb{R}$

$$\partial F(x) \in \mathbb{R}^{1 \times d}$$

$$\nabla F(x) \in \mathbb{R}^{d \times 1}$$

$$\nabla F(x) = (\partial F(x))^T$$

Prop/Def: F differentiable iff

$$F(x + \varepsilon) = F(x) + \varepsilon \underbrace{\partial F(x)}_{\mathbb{R}^{p \times d}} \times \underbrace{u}_{\mathbb{R}^d} + o(\varepsilon)$$

Chain rule

$F : \mathbb{R}^d \to \mathbb{R}^p$
$G : \mathbb{R}^p \to \mathbb{R}^q$ } diff

$$\text{"}\partial (G \circ F) = \partial G \circ \partial F\text{"}$$

$$\underbrace{\partial (G \circ F)(x)}_{q \times d} = \underbrace{\partial G(F(x))}_{q \times p} \times \underbrace{\partial F(x)}_{p \times d}$$

$$f(x) = L(B(x))$$

$$\partial f(x) = \partial L(B(x)) \times \partial B(x)$$

$$\nabla f(x) = (\partial f(x))^T$$

$$\nabla f(x) = \partial B(x)^T \times \nabla L(Bx)$$

$$F(x) = Bx$$

$$F(x + \varepsilon u) = Bx + \varepsilon \underbrace{Bu}_{\partial F(x) \times u} + 0$$

Conclu : $\nabla (L \circ B) = B^T \circ \nabla L \circ B$