Linear Regression

Derivatives, Gradients, Hessians

$f(x): \mathbb{R} \to \mathbb{R}$, $f'(x) = \frac{df}{dx} : \mathbb{R} \to \mathbb{R}$ $\quad f(x) = x^2,\ f'(x) = 2x,\ f''(x) = 2$

$f(w): \mathbb{R}^d \to \mathbb{R}$, $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$. $\nabla f = \left(\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_d}\right)$

$\nabla^2 f: \mathbb{R}^d \to \mathbb{R}^{d \times d}$

$$\begin{matrix} & & j & \\ i & \begin{bmatrix} \frac{\partial^2 f}{\partial w_1 \partial w_1} & \cdots & & \\ & & \frac{\partial^2 f}{\partial w_i \partial w_j} & \\ \vdots & & & \ddots \end{bmatrix} & & \end{matrix}$$

Statistical Learning

$(x_1, y_1), \ldots, (x_n, y_n)$ $\quad x_i \in \mathbb{R}^d,\ y_i \in \mathbb{R}$

$\ell_w(x, y)$: loss function

Goal: $\underset{w}{\operatorname{argmin}}\ \underset{(x,y)\sim D}{E}[\ell_w(x,y)]$, given $(x_i, y_i) \sim D$

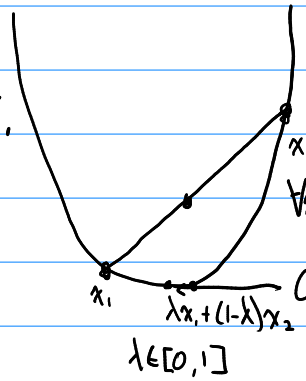
$\ell_w(x,y) = \begin{cases} 0 & \text{if } \operatorname{sign}(\langle w, x\rangle) = y \\ 1 & \text{else} \end{cases} \Rightarrow \underset{w}{\operatorname{argmin}}\ \underset{(x,y)\sim D}{Pr}[\operatorname{sign}(\langle w, x\rangle) \neq y]$

Empirical Risk Minimization (ERM)

$\underset{w}{\operatorname{argmin}}\ \frac{1}{n}\sum \ell_w(x_i, y_i) \to \underset{w}{\operatorname{argmin}}\ \underset{(x,y)\sim D}{E}[\ell_w(x,y)]$

Convexity and Optimization

Convex:



$f: \mathbb{R}^d \to \mathbb{R}$ is convex iff

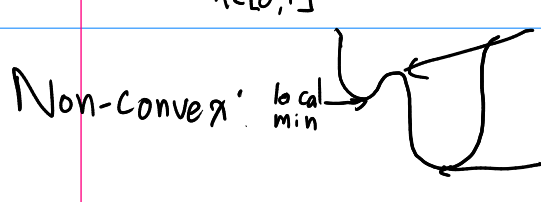
$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

$\forall x$: $f''(x) \geq 0$ (1D fn's) $\quad \nabla^2 f(x) \geq 0$

$M \in \mathbb{R}^{d \times d}$ is Positive Semidefinite (PSD) iff $v^T M v \geq 0 \ \forall v \in \mathbb{R}^d$

Global min only local min

Non-convex:

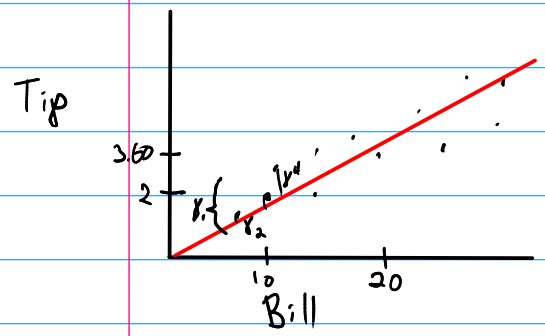


Fermat's condition

If x is an local extremum of f, then $\nabla f(x) = 0$.

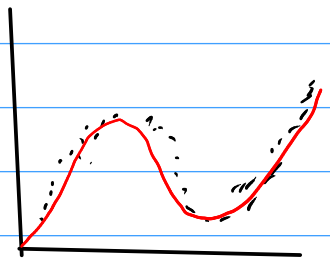
Further, if f convex, converse is true. $\nabla f(x) = 0 \Rightarrow$ global extremum

Linear Regression $[X, 1]\ [w, b]$



$$\ell_w(X,y) = (y - \langle w, X \rangle)^2$$

$$\text{Loss} = \sum \gamma_i^2$$



$$\sum (y_i - \langle w, x_i \rangle)^2 \qquad A = \begin{bmatrix} - & x_1 & - \\ & \vdots & \\ - & x_n & - \end{bmatrix} \in \mathbb{R}^{n \times d} \qquad z = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

$$\|Aw - z\|_2^2 = (Aw - z)^T (Aw - z) = (w^T A^T - z^T)(Aw - z)$$
$$= w^T A^T A w - z^T A w - w^T A^T z + z^T z$$
$$= w^T A^T A w - 2 z^T A w + z^T z$$

Claim: If $f(x) = x^T A x + x^T b + c$, $\nabla f(x) = (A + A^T)x + b$.

$$\nabla_w \|Aw - z\|_2^2 = (A^T A + (A^T A)^T) w - 2 A^T z$$
$$= 2 A^T A w - 2 A^T z.$$

$$\nabla_w^2 \|Aw - z\|_2^2 = 2 A^T A. \qquad 2(v^T A^T) A v \geq 0 \ \ \forall v$$
$$\succeq 0 \qquad = 2 \|Av\|_2^2 \geq 0$$

$\therefore$ convex

$$\nabla_w \|Aw - z\|_2^2 = 0 \iff 2 A^T A w - 2 A^T z = 0$$

$$A^T A w = A^T z$$

Optimize LS. by finding $\hat{w}$ s.t. $A^T A \hat{w} = A^T z$.

$$\hat{w} = (A^T A)^{-1} A^T z ?$$

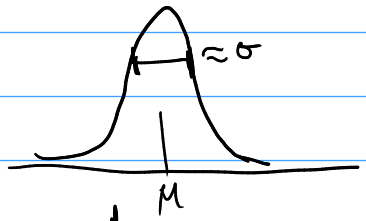
X

$\hat{y} = \langle \hat{w}, X \rangle$

Squared loss via Maximum Likelihood Estimation (MLE)

Gaussian $N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

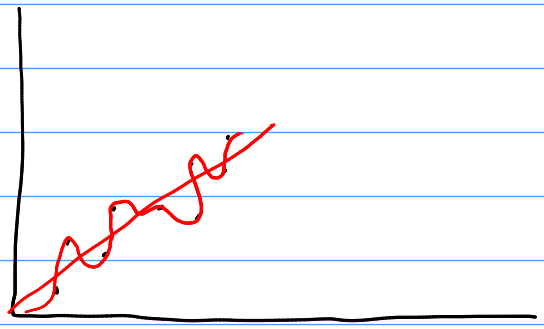


$$\underset{\text{params}}{\text{argmax}} \; \Pr[\text{observed data} \mid \text{model params}].$$

$$y_i = \langle w, x_i \rangle + z_i, \quad z \sim N(0, \sigma^2)$$

$$\begin{aligned}
\hat{w} &= \underset{w}{\text{argmax}} \; \Pr[(x_1, y_1) \ldots (x_n, y_n) \mid w] \\
&= \underset{w}{\text{argmax}} \prod_i \Pr[(x_i, y_i) \mid w] \\
&= \underset{w}{\text{argmax}} \prod_i \Pr[y_i \mid x_i, w] \Pr[x_i \mid w] \\
&= \underset{w}{\text{argmax}} \prod_i \Pr[y_i \mid x_i, w] \\
&= \underset{w}{\text{argmax}} \log\left(\prod \Pr[y_i \mid x_i, w]\right) \\
&= \underset{w}{\text{argmax}} \sum \log \Pr[y_i \mid x_i, w] \qquad y_i \mid x_i, w \sim N(\langle w, x_i \rangle, \sigma^2) \\
&= \underset{w}{\text{argmax}} \sum \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{(y_i - \langle w, x_i \rangle)^2}{2\sigma^2}\right)\right) \\
&= \underset{w}{\text{argmax}} \; -\frac{1}{2\sigma^2} \sum (y_i - \langle w, x_i \rangle)^2 \\
&= \underset{w}{\text{argmin}} \sum (y_i - \langle w, x_i \rangle)^2
\end{aligned}$$

Regularization



$X_1 = [0, 1]$ $y_1 = 1$

$X_2 = [\varepsilon, 0]$, $y_2 = -1$ $\quad \hat{w} = \begin{bmatrix} -2/\varepsilon \\ 1 \end{bmatrix}$

Tikhonov Reg. Ridge Regression

$$\min \|Aw - z\|_2^2 + \lambda \|w\|_2^2$$

λ: hyperparameter

Lasso

$$\min \|Aw - z\|_2^2 + \lambda \|w\|_1$$

sparse

not closed form solve

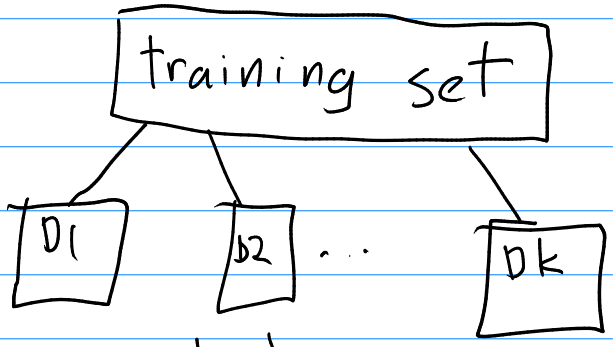
Hyperparameters, Cross Validation

Training dataset

~~Validation dataset~~

Test dataset

$\lambda \in \{0.01, 0.1, 0.5, 1\}$



$k \approx 5$? $k = 10$?

For each λ:

For $i = 1$ to k:

$$w_{\lambda,i} = \text{train}\left(\bigcup_{\substack{j=1 \\ j \neq i}}^{k} D_j\right)$$

$$\text{perf}_{\lambda,i} = \text{acc}_{w_{\lambda,i}}(D_i)$$

$$\text{perf}_\lambda = \sum_{i=1}^{k} \text{perf}_{\lambda,i}$$

Return $\arg\max_\lambda \text{perf}_\lambda$

$\|Aw - z\|_2^2 + \lambda \|w\|_2^2 \leftarrow$ train

$\|Aw - z\|_2^2 \leftarrow$ test, validation

Claim: If $f(x) = x^T A x + x^T b + c$, $\nabla f(x) = (A + A^T)x + b$.

$\nabla f(x) = \nabla x^T A x + \nabla x^T b + \nabla c$

$\nabla c = \vec{0}$

$\nabla b^T x$, $b^T x = \sum_{i=1}^{d} b_i x_i$, $\frac{\partial}{\partial x_i} b^T x = b_i \Rightarrow \nabla b^T x = b$

$\nabla x^T A x$, $x^T A x = \sum_{u,v} x_u x_v A_{uv}$

$\frac{\partial}{\partial x_i} x^T A x = 2 x_i A_{ii} + \sum_{v \neq i} x_v A_{iv} + \sum_{u \neq i} x_u A_{ui}$

$= \underbrace{\sum_{v} x_v A_{iv}}_{\langle A_{i\cdot}, x \rangle \; (Ax)_i} + \underbrace{\sum_{u} x_u A_{ui}}_{\sum_u x_u A^T_{iu} \; (A^T x)_i}$

$\nabla x^T A x = (Ax) + (A^T x) = (A + A^T)x$

$x = \begin{pmatrix} x_1 \\ x_n \end{pmatrix}$

$A = \begin{pmatrix} - A_{1\cdot} - \\ - A_{i\cdot} - \end{pmatrix}$.

$(Ax)_i = \langle A_{i\cdot}, x \rangle$