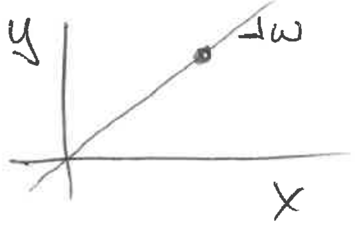
Linear Regression in 1D

Simplified model

$\hat{y}(x) = w \cdot x$ x is scalar

w is scalar

Goal: minimize mean squared error on train set



Given 1 data pair (x_1, y_1)

exactly one value of w "solves"!

$$w^* = \frac{y_1}{x_1}$$

exactly predicts with zero error

What should we do w/ many training examples?

need to fit w^* value that minimizes error

(unlikely / impossible to get zero error)

another term for aggregate error on training set = "loss"

$$\{x_n, y_n\}_{n=1}^{N} \rightarrow \boxed{\text{fit}(\ )} \rightarrow w^*$$

$$\min_{w \in \mathbb{R}} \boxed{\frac{1}{N}} \sum_n (y_n - w x_n)^2$$

equivalent to $\min_w \sum_n (y_n - w x_n)^2$

equiv. to $\min_w \boxed{\frac{1}{2}} \sum_n (y_n - w x_n)^2$

const in front only scales magnitude of loss

any choice of const will leave minimizer w^* unchanged

Linear Regression in 1D

Loss: $J(w) = \frac{1}{2} \sum_n (y_n - w x_n)^2$

First Deriv: $\frac{\partial}{\partial w} J(w) = \frac{\partial}{\partial w}\left[\frac{1}{2} \sum_n (y_n - w x_n)^2 \right]$

$$= \frac{1}{2} \sum_n \frac{\partial}{\partial w}\left[(y_n - w x_n)^2 \right]$$

$$= \cancel{\tfrac{1}{2}} \sum_n \cancel{2} (y_n - w x_n) \frac{\partial}{\partial w}\left[-w x_n \right]$$

$$= -\sum_n (y_n - w x_n) x_n$$

Second Deriv: $\frac{\partial}{\partial w}\frac{\partial}{\partial w} J(w) = -\sum_n \frac{\partial}{\partial w}\left[(y_n - w x_n) x_n \right]$

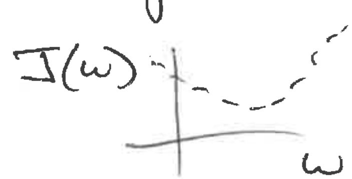
$$= -\sum_n \frac{\partial}{\partial w}\left[x_n y_n - w x_n^2 \right]$$

$$= -\sum_n \frac{\partial}{\partial w}\left[\cancelto{0}{x_n y_n} \right] - \frac{\partial}{\partial w}\left[w x_n^2 \right]$$

$$= \sum_n x_n^2$$

Observation: 2^{nd} deriv positive or zero always

thus, our objective is convex



Solving for optimal $w^* = \arg\min_w J(w)$

Can set first deriv. to zero and solve for $\underline{w}$

$$\frac{\partial}{\partial w} J(w) = 0$$

$$-\sum_n (y_n - w x_n) x_n = 0$$

$$-\left(\sum_n x_n y_n\right) + w \sum_n x_n^2 = 0$$

$$w \sum_n x_n^2 = \sum_n x_n y_n$$

$$w = \frac{\sum_n x_n y_n}{\sum_n x_n^2}$$

thus $\boxed{w^* = \frac{\sum_n x_n y_n}{\sum_n x_n^2}}$

sanity check:

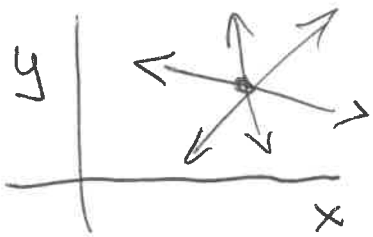
does it cover the $N=1$ case?

$$w^* = \frac{x_1 y_1}{x_1^2} = \frac{y_1}{x_1} \quad \checkmark$$

Linear Regression in 1D inputs with bias/intercept

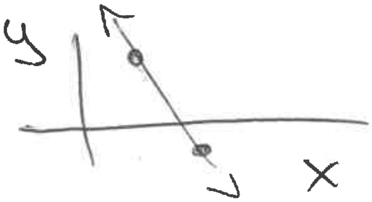
Prediction function:

$$\hat{y}(x) = wx + b$$



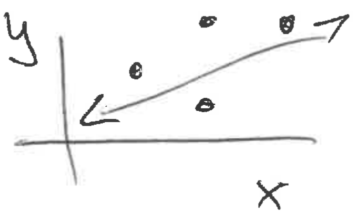
1 data pair (x_1, y_1)

infinitely many lines with error zero
"under-determined"



2 data pairs (x_1, y_1)
(x_2, y_2)

just one line with zero error



3+ distinct pairs
"over determined"
may not achieve zero error

Loss function

$$J(w,b) = \frac{1}{2} \sum_n \left(y_n - (w x_n + b)\right)^2$$

Partial derivatives

$$\frac{\partial}{\partial w} J = -\sum_n (y_n - w x_n - b)\, x_n$$

$$\frac{\partial}{\partial b} J = -\sum_n (y_n - w x_n - b)$$

How would we check convexity?

$$\text{Hessian } H = \begin{bmatrix} \frac{\partial^2}{\partial w^2} J & \frac{\partial^2}{\partial w \partial b} J \\ \frac{\partial^2}{\partial w \partial b} J & \frac{\partial^2}{\partial b^2} J \end{bmatrix}$$

is it positive semi definite?

then convex

Linear Regr w/ 1D inputs and bias/intercept

want: optimal values $w^*, b^* = \arg\min_{w,b} J(w,b)$

just set partial derivs to zero and solve.

$\boxed{\frac{\partial}{\partial w} J(w,b) = 0}$ (A)

$\boxed{\frac{\partial}{\partial b} J(w,b) = 0}$ (B)

two eqns (A) and (B)
two unknowns w^*, b^*

Example: solve for b^*

$$-\sum_n (y_n - w x_n - b) = 0$$

plug in $\frac{\partial}{\partial b} J$ from prev. page

$$\frac{1}{N} \sum_n (y_n - w x_n - b) = 0$$

multiply both sides by $-\frac{1}{N}$

$$\bar{y} - w\bar{x} - b = 0$$

define $\bar{y} = \frac{1}{N}\sum_n y_n$
$\bar{x} = \frac{1}{N}\sum_n x_n$

$$\boxed{b^* = \bar{y} - w\bar{x}}$$

now plug this into $\frac{\partial}{\partial w} J(w,b) = 0$ and solve for w

we'll skip details, but we get w^* shown on slides
(lots and lots of algebra)

Linear Regr. w/ vector inputs

Prediction model:

$$\hat{y}(x_i) = b + w_1 x_{i1} + w_2 x_{i2} + \cdots w_F x_{iF}$$

$$= b + w^T x_i \quad \text{\underline{inner product}}$$

Let's rewrite w/ more <u>compact</u> <u>notation</u>

$$\theta = [b \ \ w_1 \ \ w_2 \ \cdots \ w_F]$$

$$\tilde{x}_i = [1 \ \ x_{i1} \ \ x_{i2} \ \cdots \ x_{iF}]$$

$$\text{so} \quad \hat{y}(x_i) = \theta^T \tilde{x}_i$$

Write squared error loss function

$$J(\theta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \theta^T \tilde{x}_n)^2$$

sum of N scalars

Another way to write

$$\underset{N \times (F+1)}{\tilde{X}} = \begin{bmatrix} - \tilde{x}_1 - \\ \vdots \\ - \tilde{x}_N - \end{bmatrix} \qquad \underset{(N \times 1)}{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \qquad \underset{(F+1 \times 1)}{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_{F+1} \end{bmatrix}$$

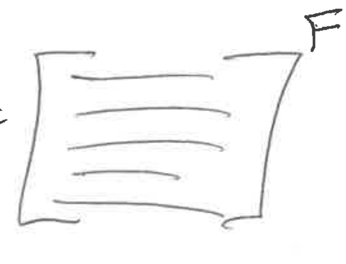
$$J(\theta) = \frac{1}{2} \sum_n (y_n - \tilde{x}_n^T \theta)(y_n - \tilde{x}_n^T \theta)$$

$$= \frac{1}{2} (y - \tilde{X}\theta)^T (y - \tilde{X}\theta) \quad \text{inner prod of len-}N\text{ vectors}$$

Rules for matrix-vector gradients

$\theta : F \times 1$ $\qquad$ $v : F \times 1$ $\qquad$ $Q : N \times F$

$$\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_F \end{bmatrix} \qquad v = \begin{bmatrix} v_1 \\ \vdots \\ v_F \end{bmatrix} \qquad Q =$$



expr	shape	∇_θ expr	shape of $\nabla_\theta[\text{expr}]$
scalar func of θ	1×1	col vector	$F \times 1$
(A) $v^T\theta$	1×1	v	$F \times 1$
$\theta^T v$	1×1	v	$F \times 1$
$\theta^T\theta$	1×1	2θ	$F \times 1$
(B) $\theta^T Q^T Q \theta$	1×1	$2Q^TQ\theta$	$F \times 1$

Want: $\theta^{*} = \underset{\theta}{\text{argmin}}\ J(\theta) = \frac{1}{2}\underbrace{(y-\tilde{X}\theta)^T(y-\tilde{X}\theta)}_{\text{inner product of length } N \text{ vectors}}$

Procedure: set gradient $\nabla_\theta J$ to zero and solve for θ^*

$$\nabla_\theta J(\theta) = \frac{1}{2}\nabla_\theta\Big[\underbrace{(y-\tilde{X}\theta)^T(y-\tilde{X}\theta)}_{\text{FOIL}}\Big]$$

$$= \frac{1}{2}\nabla_\theta\left[y^Ty - y^T\tilde{X}\theta - \theta^T\tilde{X}^Ty + \theta^T\tilde{X}^T\tilde{X}\theta\right]$$

realize $a = \theta$
$b = (y^T\tilde{X})^T$

$$-b^Ta - a^Tb$$
$$= -2b^Ta$$

$$= \frac{1}{2}\nabla_\theta\Big[\underbrace{\cancel{y^Ty}}_{\text{zero under gradient}} \underbrace{-2y^T\tilde{X}\theta}_{\text{grad rule A}} + \underbrace{\theta^T\tilde{X}^T\tilde{X}\theta}_{\text{grad rule B}}\Big]$$

$$= \frac{1}{2}\left[-2\tilde{X}^Ty + 2\tilde{X}^T\tilde{X}\theta\right]$$

$$\boxed{\nabla_\theta J(\theta) = \tilde{X}^T\tilde{X}\theta - \tilde{X}^Ty}$$

check dims

$\tilde{X}^T\tilde{X}\theta = F\times N\ N\times F\ F\times 1 = F\times 1$ ✓

$\tilde{X}^Ty = F\times N\ N\times 1 = F\times 1$ ✓

We have $\nabla_\theta J(\theta) = \tilde{X}^T \tilde{X} \theta - \tilde{X}^T \underset{\sim}{y}$

now set to zero vector and solve!

$$\tilde{X}^T \tilde{X} \theta - \tilde{X}^T \underset{\sim}{y} = \vec{0}$$

$$\tilde{X}^T \tilde{X} \theta = \tilde{X}^T \underset{\sim}{y}$$

multiply both sides by $(\tilde{X}^T \tilde{X})^{-1}$

$$\boxed{\theta^* = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \underset{\sim}{y}}$$

only works if $\tilde{X}^T \tilde{X}$ is an invertible matrix

---

special case check:

suppose $\tilde{X}$ is square and invertible and symmetric

simplifies to $(\tilde{X})^{-1} (\tilde{X})^{-1} \tilde{X} \underset{\sim}{y} = \tilde{X}^{-1} \underset{\sim}{y}$

in 1D, this becomes $\frac{y}{x}$ ✓ which matches result from page 1