EECS 245 Fall 2025

Math for ML

Lecture 15: Projections; Regression via Lin. Alg.

→ Read: Ch 2.10 (tons of new content),

Ch 3.1 (in progress)

Agenda

→ Projecting onto span of multiple vectors

  → The "normal equation"

→ Simple linear regression using linear algebra :o

Announcements:

① Apply to be an IA! (dept form due Thurs, 245 form + video due Sat)

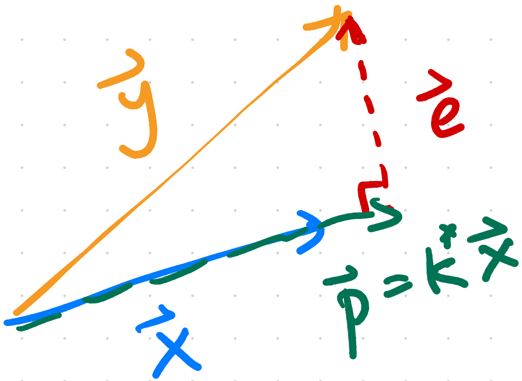
② HW 7 due Friday

  → Lab tomorrow: HW 7 work session

$$X_{n \times d} = \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \cdots & \vec{x}^{(d)} \\ | & | & & | \end{bmatrix} \qquad \vec{y} \in \mathbb{R}^n$$

Issue: $\vec{y}$ is not (necessarily) in $\text{colsp}(X)$

Question: Among all vectors in $\text{colsp}(X)$, which is "closest" to $\vec{y}$?



k^* chosen such that

$\vec{e}$ orthogonal to $\vec{x}$;

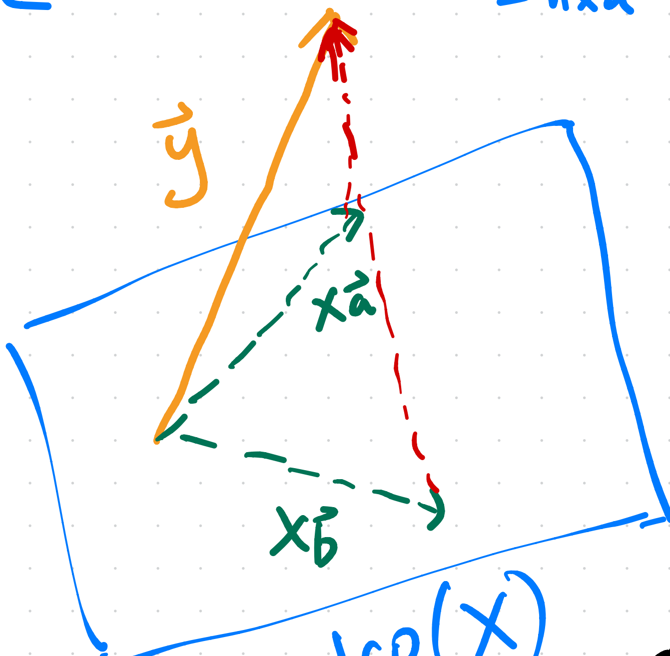
that k^* minimized

$$f(k) = \|\vec{y} - k\vec{x}\|^2 = \|\vec{e}\|^2$$

$$k^* = \frac{\vec{x}\cdot\vec{y}}{\vec{x}\cdot\vec{x}}$$

$$X = \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \cdots & \vec{x}^{(d)} \\ | & | & & | \end{bmatrix}_{n \times d}, \quad \vec{y} \in \mathbb{R}^n$$

$$X = \begin{bmatrix} 1 & 2 \\ 0 & -1 \\ 3 & -1 \end{bmatrix}_{3 \times 2}$$



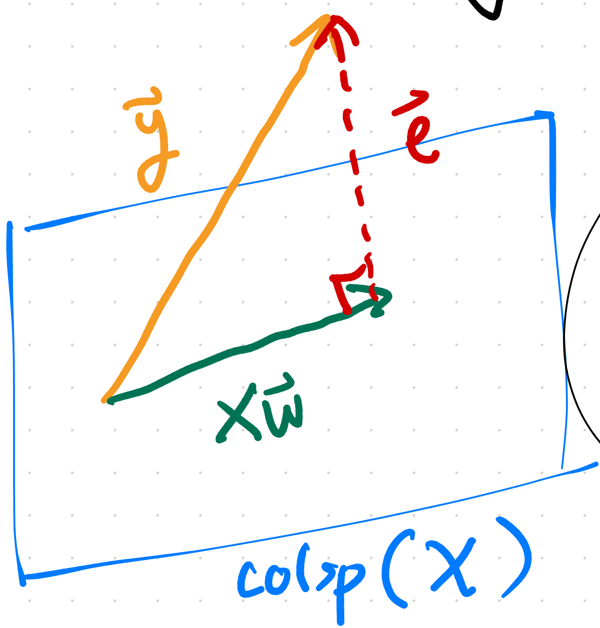
$X w_d$

$\text{colsp}(X)$ = set of all lin combs of X's cols = set of all outputs of $X\vec{w}$, $\vec{w} \in \mathbb{R}^d$

key idea: to minimize

$$\|\vec{e}\|^2 = \|\underbrace{\vec{y} - X\vec{w}}_{\text{function of } \vec{w} \text{ only}}\|^2,$$

pick $\vec{w}$ so that $\vec{y} - X\vec{w}$ orthogonal to every vector in $\text{colsp}(X)$



← most important diagram of the semester well.... which $\vec{w}$ satisfies our goals?

Goal: Need $\vec{y} - X\vec{w}$ to be orthogonal to every col of X

$$X = \begin{bmatrix} | & | & & | \\ \vec{x}^{(1)} & \vec{x}^{(2)} & \cdots & \vec{x}^{(d)} \\ | & | & & | \end{bmatrix}_{n \times d}$$

$$\vec{x}^{(1)} \cdot (\vec{y} - X\vec{w}) = 0$$
$$\vec{x}^{(2)} \cdot (\vec{y} - X\vec{w}) = 0$$
$$\vdots$$

$$X^T = \begin{bmatrix} \text{---} & \vec{x}^{(1)T} & \text{---} \\ \text{---} & \vec{x}^{(2)T} & \text{---} \\ & \vdots & \\ \text{---} & \vec{x}^{(d)T} & \text{---} \end{bmatrix}_{d \times n}$$

$$X^T(\vec{y} - X\vec{w}) = \begin{bmatrix} \vec{x}^{(1)} \cdot (\vec{y} - X\vec{w}) \\ \vec{x}^{(2)} \cdot (\vec{y} - X\vec{w}) \\ \vdots \\ \vec{x}^{(d)} \cdot (\vec{y} - X\vec{w}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}_d$$

remember: we're looking for the "best" $\vec{w}$

$\Rightarrow$ equivalent to finding $\vec{w}$ that satisfies

$$X^T(\vec{y} - X\vec{w}) = \vec{0}$$

$$X^T\vec{y} - X^T X \vec{w} = \vec{0}$$

$$X^T X \vec{w} = X^T \vec{y}$$

the normal equations

system of d equations, d unknowns

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \end{bmatrix}$$

X^TX
dot products of all pairs of cols

$$\underbrace{X^TX}_{d \times d}\vec{w} = X^T\vec{y}$$

Case ①: X^TX invertible

which happens if and only if X's cols are linearly independent

(remember, $\text{rank}(X) = \text{rank}(X^TX)$, so X^TX is invertible when $\text{rank}(X^TX) = d = \text{rank}(X)$)

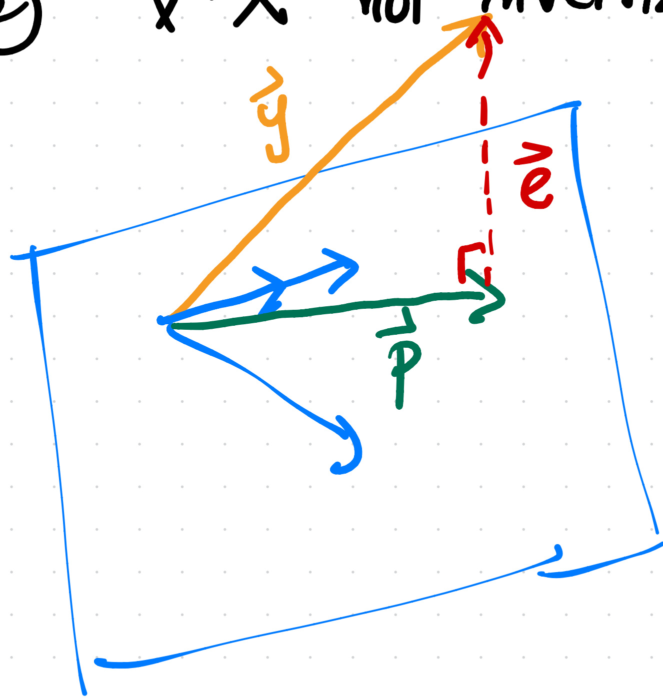
Then, unique solution:

$$\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$$

most important equation

$$X^T X \vec{w} = X^T \vec{y}$$

Case ② $X^T X$ not invertible → then, there are infinitely many $\vec{w}^*$ that satisfy the normal equations, but all of them correspond to the same projection, $\vec{p}$



Example

$$X = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 1 & 1 \\ 0 & -1 \end{bmatrix}, \quad \vec{y} = \begin{bmatrix} 1 \\ 0 \\ 4 \\ 5 \end{bmatrix}$$

Q: which lin comb of X's cols is closest to $\vec{y}$?

A: $X\vec{w}^*$, where

$$\vec{w}^* = (X^TX)^{-1}X^T\vec{y}$$

$$= \begin{bmatrix} 2 \\ -8/3 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 6 & 3 \\ 3 & 3 \end{bmatrix}$$

$$(X^TX)^{-1} = \frac{1}{9}\begin{bmatrix} 3 & -3 \\ -3 & 6 \end{bmatrix}$$

if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$,

then

$$A^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

one question on the midterm
will be this exactly

→ practice it!

of all vectors of the form

$$w_1 \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} + w_2 \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix},$$

the one that is closest to $\vec{y} = \begin{bmatrix} 1 \\ 6 \\ 4 \\ 5 \end{bmatrix}$ is

$$X = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$2 \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix} - 8/3 \begin{bmatrix} 0 \\ 1 \\ 1 \\ -1 \end{bmatrix}$$

orthogonality

$$\vec{p} = 2\begin{bmatrix}1\\2\\1\\0\end{bmatrix} - 8/3\begin{bmatrix}0\\1\\1\\-1\end{bmatrix} = X\vec{w}^*$$

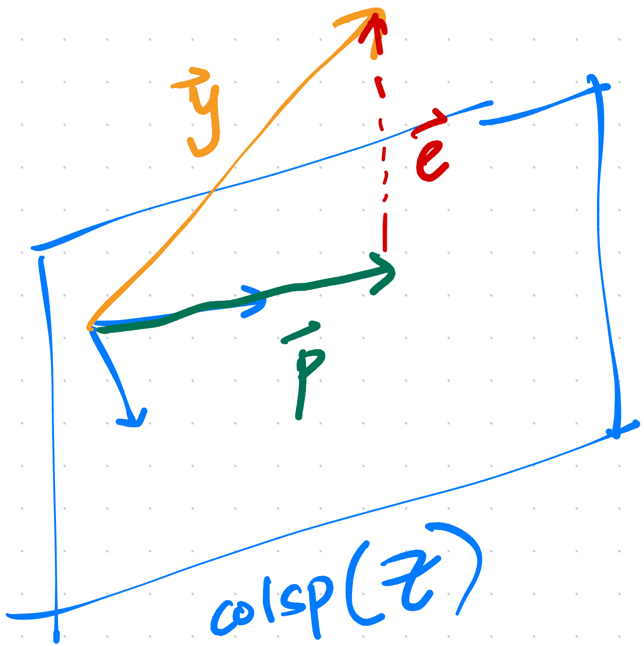
$\vec{e} = \vec{y} - \vec{p}$ is orthogonal to both cols of X

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 2 \\ 0 & 2 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \end{bmatrix} \qquad \vec{y} \in \mathbb{R}^3$$

let $\vec{p}_X$ be the orthogonal proj of $\vec{y}$ onto $\operatorname{colsp}(X)$

$\vec{p}_Z$ . . . - - - — $\operatorname{colsp}(Z)$

<u>Why</u> is it the case that the error vector for $\vec{p}_X$ sums to 0, but not for $\vec{p}_Z$?


$\vec{y}$
$\vec{e}$
$\vec{p}$
colsp(Z)

$$X = \begin{bmatrix} 1 & 2 \\ 3 & 2 \\ 0 & 2 \end{bmatrix}$$

$$\vec{e}_x = \vec{y} - \vec{P_x} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

Fact: $\vec{e}_x$ orthogonal to all cols of X

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} \cdot \left(\text{any lin comb of } X\text{'s cols} \right) = 0$$

$$\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = 0 \rightarrow 2e_1 + 2e_2 + 2e_3 = 0$$

$$e_1 + e_2 + e_3 = 0$$

orthogonal proj of $\vec{y}$ onto colsp(X)

$$\vec{p} = X\vec{w}^* = \underbrace{X(X^TX)^{-1}X^T}_{P = X(X^TX)^{-1}X^T}\vec{y}$$

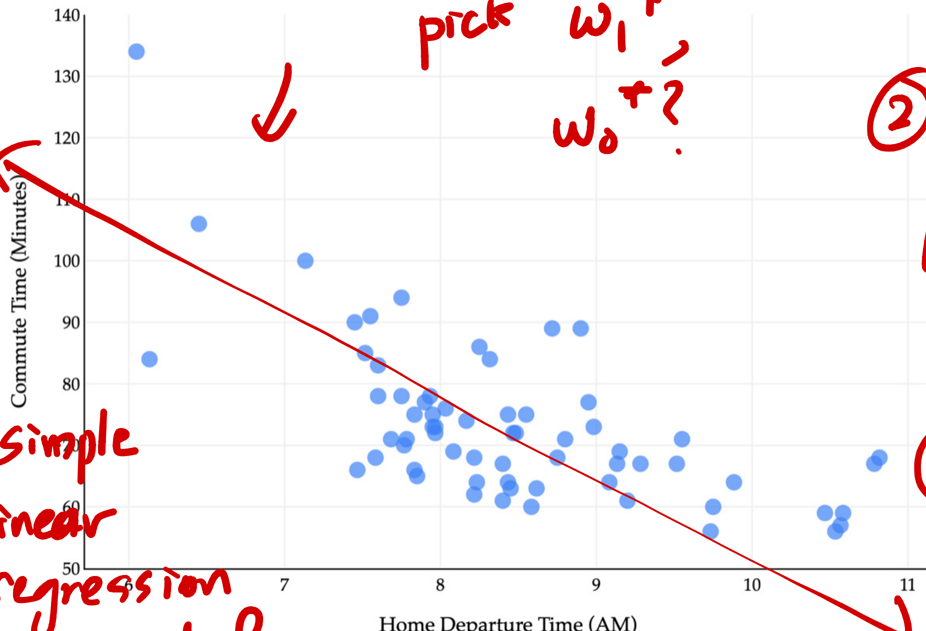
P "projection matrix" "hat matrix"

linear transformation of $\vec{y}$

read notes Ch. 2.10

Linear regression!

how did we pick w_1^*, w_0^*?



simple linear regression model

3-step modeling process

① choose a model

$$h(x_i) = w_0 + w_1 x_i$$

② choose loss function

$$L_{sq}(y_i, h(x_i)) = (y_i - h(x_i))^2$$

③ minimize average loss

mean squared error
average squared loss
empirical risk

$$R_{sq}(w_0, w_1) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big(\overbrace{y_i - \underbrace{(w_0 + w_1 x_i)}_{h(x_i)}}^{\text{actual - pred}}\Big)^2}_{\text{this sum is the (squared) norm of error vector}}$$

definitions (Ch. 3.1)

given dataset

$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_n$$

"observation vector"

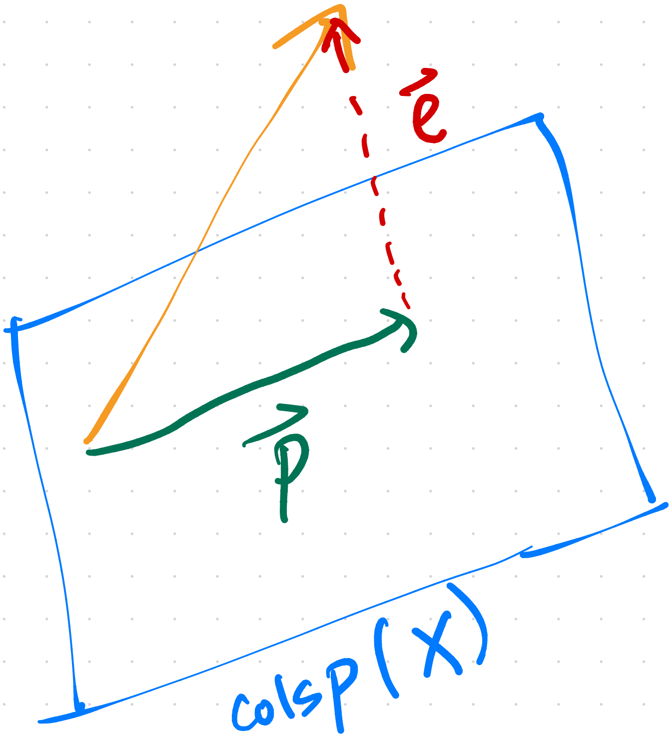
actual commute times

"parameter vector" $\vec{w}$

$$\vec{p} = \begin{bmatrix} w_0 + w_1 x_1 \\ w_0 + w_1 x_2 \\ \vdots \\ w_0 + w_1 x_n \end{bmatrix}_n = w_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + w_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}_{X} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

"prediction vector"

"design matrix"



$$\vec{e} = \vec{y} - \vec{p} = \begin{bmatrix} y_1 - (w_0 + w_1 x_1) \\ y_2 - (w_0 + w_1 x_2) \\ \vdots \\ y_n - (w_0 + w_1 x_n) \end{bmatrix}$$

Goal: minimize $\frac{1}{n} \|\vec{e}\|^2$

$$= \frac{1}{n}\Big(\big(y_1 - (w_0 + w_1 x_1)\big)^2 + \dots + \big(y_n - (w_0 + w_1 x_n)\big)^2 \Big)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \big(y_i - (w_0 + w_1 x_i)\big)^2$$

if you define

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

"design matrix"

then

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y} = \overbrace{\begin{bmatrix} \bar{y} - r\frac{\sigma_y}{\sigma_x}\bar{x} \\ r\frac{\sigma_y}{\sigma_x} \end{bmatrix}}^{\text{from before!}}$$