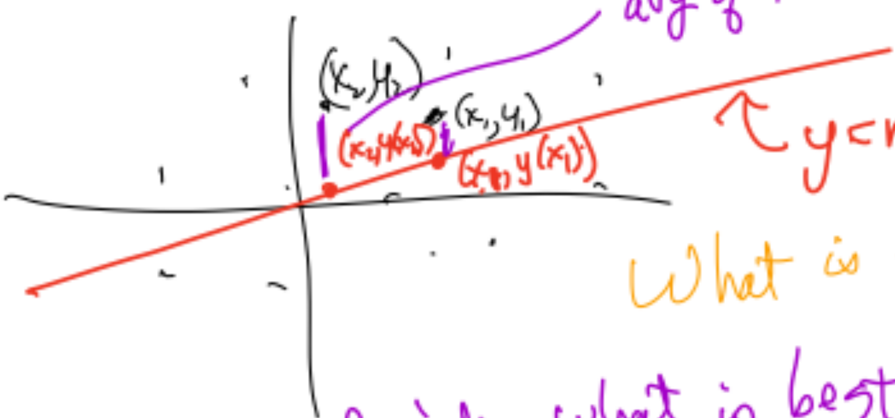
Regression

<u>Problem:</u> Given N data point

$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$.

<u>Want:</u> Find the line that is closest to the data as possible.

avg of these $(\text{distance})^2 = \text{MSE}$



$y = mx + b$

What is the best?

How we decide what is best?

We minimize the <u>cost function</u>.

Let $y(x) = mx + b$.

Cost funct = MSE mean square error

Predicted y-values: $y(x_1), y(x_2), \dots, y(x_N)$

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^{N} (y(x_k) - y_k)^2 = \frac{1}{N}\left((y(x_1) - y_1)^2 + (y(x_2) - y_2)^2 + \dots\right)$$

$$\begin{pmatrix}\text{RMSE} \\ \text{root mean square} \\ \text{error}\end{pmatrix} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(y(x_k)-y_k\right)^2}$$

a function of m & b

$$RMSE(b,m) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\Big(\underbrace{mx_k+b}_{y(x_k)}-y_k\Big)^2}$$

everything in here is a constant except m & b.

Example: Data pts $(\underset{x_1}{-4},\underset{y_1}{2}),(3,7),(3,8),(4,6)$

$$RMSE = F(m,b) = \sqrt{\frac{1}{4}\Big[\big(\underbrace{m(-4)+b}_{y(x_1)=\hat{y}_1}-2\big)^2+\big(m(3)+b-7\big)^2 + \big(m(3)+b-8\big)^2+\big(m(4)+b-6\big)^2}$$

$\hat{y}_1 = y(x_1) =$ predicted value of $y_1 = m(-4)+b$

$y_1 = 2 =$ actual value of y_1

want to find the global min of $F(m,b)$ for all possible choices of m & b.

Ans: $m =$ ____ ; $b =$ ____ . $y(x) =$ ___ $x +$ ___

Why choose RMSE as the function?

- If RMSE is at a minimum, the graph is pretty close to the points.
- It turns out that we always get a unique solution for m & b with this function.

An alternative: Minimize

$$\sum \left| \underbrace{y(x_k) - y_k}_{m x_k + b - y_k} \right|$$

(sum of distances)

- If this is minimum (wrt to choice of m, b)
- But it turns out we often don't get a unique solution!!!

How do we solve this.

① Simplify function — Instead of RMSE, let's use $G(b,m) = \sum_{k=1}^{N} (y(x_k) - y_k)^2$

(Don't divide by N, don't take square root.)

This is OK, because if we find the b,m such that $G(b,m)$ is minimum, Also $RMSE(b,m)$ will also be minimum

② We want to find the minimum of
$$G(b,m) = \sum_{k=1}^{N} (mx_k + b - y_k)^2 ?$$

Ans#1: $\frac{\partial G}{\partial b} = 0$, $\frac{\partial G}{\partial m} = 0$.

$$\begin{cases} \frac{\partial G}{\partial b} = \sum_{k=1}^{N} 2(mx_k + b - y_k) \cdot 1 = 0 \\ \frac{\partial G}{\partial m} = \sum_{k=1}^{N} 2(mx_k + b - y_k) \cdot x_k = 0 \end{cases}$$

Simplify. divide by 2

$$\begin{cases} \sum_{k=1}^{N} (mx_k + b - y_k) = 0 \\ \sum_{k=1}^{N} (mx_k^2 + bx_k - x_k y_k) = 0 \end{cases}$$

(Distr. property $\Longrightarrow$ & comm & assoc prop.)

$$\begin{cases} m\left(\sum_{k=1}^{N} x_k\right) + b\sum_{k=1}^{N} 1 - \sum_{k=1}^{N} y_k = 0 \\ m\left(\sum_{k=1}^{N} x_k^2\right) + b\sum_{k=1}^{N} x_k - \sum_{k=1}^{N} x_k y_k = 0 \end{cases}$$

N

$$\begin{cases} N b + \left(\sum_{k=1}^{N} x_k\right) m = \sum_{k=1}^{N} y_k \\ \left(\sum_{k=1}^{N} x_k\right) b + \left(\sum_{k=1}^{N} x_k^2\right) m = \sum_{k=1}^{N} x_k y_k \end{cases}$$

deleting subscripts to make it look wiser.

$$\begin{pmatrix} N & \Sigma x \\ \Sigma x & \Sigma x^2 \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} \Sigma y \\ \Sigma xy \end{pmatrix}$$

$$\begin{pmatrix} b \\ m \end{pmatrix} = \begin{pmatrix} N & \Sigma x \\ \Sigma x & \Sigma x^2 \end{pmatrix}^{-1} \begin{pmatrix} \Sigma y \\ \Sigma xy \end{pmatrix}$$

matrix   variable vector   vector

$$Ax = C$$

$$x = A^{-1} C$$

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \frac{1}{\det} \begin{pmatrix} D & -B \\ -C & A \end{pmatrix}$$

$$\begin{pmatrix} b \\ m \end{pmatrix} = \frac{1}{N\Sigma x^2 - (\Sigma x)^2} \begin{pmatrix} \Sigma x^2 & -\Sigma x \\ -\Sigma x & N \end{pmatrix} \begin{pmatrix} \Sigma y \\ \Sigma xy \end{pmatrix}$$

$$\begin{pmatrix} b \\ m \end{pmatrix} = \frac{1}{N(\Sigma x^2) - (\Sigma x)^2} \cdot \begin{pmatrix} (\Sigma x^2)(\Sigma y) - (\Sigma x)(\Sigma xy) \\ -(\Sigma x)(\Sigma y) + N(\Sigma xy) \end{pmatrix}$$

Answer: $b = \dfrac{(\sum x^2)(\sum y) - (\sum x)(\sum xy)}{N(\sum x^2) - (\sum x)^2}$

$$m = \left(\frac{-(\sum x)(\sum y) + N(\sum xy)}{N(\sum x^2) - (\sum x)^2} \right)$$

We got only one solution.

It's a global minimum of the function

$$G(b,m) = \sum_{k=1}^{N} (m x_k + b - y_k)^2 \geq 0$$

$\to \infty$ as m, b goes to ∞ in any direction.

2nd deriv matrix:

$$\nabla G(b,m) = \left(\sum 2((m x_k + b) - y_k), \sum 2(m x_k^2 + b x_k - x_k y_k) \right)$$

2nd deriv $D_2 G(b,m) = \begin{pmatrix} \sum 2 & \sum 2 x_k \\ \sum 2 x_k & \sum 2 x_k^2 \end{pmatrix}$

$$= 2 \begin{pmatrix} N & \sum x \\ \sum x & \sum x^2 \end{pmatrix}$$

all positive eigenvalues $\Rightarrow G(b,m)$ concave up

A much better solution:

We are trying to minimize

$$G(b,m) = \sum_{k=1}^{N} \left((m x_k + b) - y_k \right)^2$$

with respect to all possible choices of b & m.

$$G(b,m) = (mx_1+b-y_1)^2 + (mx_2+b-y_2)^2 + \dots + (mx_N+b-y_N)^2$$

$$G(b,m) = \left\| (mx_1+b-y_1, \dots, mx_N+b-y_N) \right\|^2$$

$$G\begin{pmatrix} b \\ m \end{pmatrix} = \left\| \begin{pmatrix} mx_1+b-y_1 \\ mx_2+b-y_2 \\ \vdots \\ mx_N+b-y_N \end{pmatrix} \right\|^2$$

$$= \left\| \begin{pmatrix} mx_1+b \\ \vdots \\ mx_N+b \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \right\|^2$$

$$= \left\| \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \right\|^2$$

$$= \Bigg\| \, b\underset{\mathbb{1}}{\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}} + m\underset{X}{\begin{pmatrix}x_1\\ \vdots\\ x_N\end{pmatrix}} - \underset{Y}{\begin{pmatrix}y_1\\ \vdots\\ y_N\end{pmatrix}} \Bigg\|^2$$

plane containing X & $\mathbb{1}$
(contains all possible linear combinations of X & $\mathbb{1}$).

The best b & m (to make $G(b,m)$ smallest).

$$\underset{\text{dot product}}{(b\mathbb{1}+mX)\cdot\big(Y-(b\mathbb{1}+mX)\big)} = 0$$

ie $$\Big[\underbrace{(\mathbb{1}|X)}_{A}\underbrace{\begin{pmatrix}b\\ m\end{pmatrix}}_{\vec{P} \leftarrow \text{parameters.}}\Big]\cdot\Big(Y - \underbrace{(\mathbb{1}|X)\begin{pmatrix}b\\ m\end{pmatrix}}_{A\vec{P}}\Big) = 0$$

$$A\vec{p} \cdot (Y - A\vec{p}) = 0$$

Dot product $V \cdot W = V^T W$

$$\begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix} \cdot \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = (v_1 \cdots v_N) \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix}$$

$$= \sum v_k w_k$$

$$\Rightarrow (Ap)^T (Y - Ap) = 0$$

Solve for p.

Recall $(AB)^T = B^T A^T$

$$p^T \underbrace{A^T (Y - Ap)}_{\text{make this } = 0.} = 0$$

$$-A^T Y + A^T A p = 0$$

$$A^T A p = A^T Y$$

$$\Rightarrow p = (A^T A)^{-1} A^T y.$$

Recall $A = (\mathbb{1} | X)$ $N \times 2$

A^T $2 \times N$ $A^T A = 2 \times 2$

$$\boxed{P = \begin{pmatrix} b \\ m \end{pmatrix} = (A^T A)^{-1} A^T \underset{\uparrow}{Y}} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

The least squares solution to minimize the distance^2 from $b\mathbb{1} + mX$ to Y.

This solution is easy to generalize.

<u>example:</u> polynomial regression:

Find a_0, a_1, a_2, a_3 so that $y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$ fits the data best.

$$\Bigg\| \underbrace{\left(\mathbb{1} \; \begin{vmatrix} x_1 \\ x_2 \\ \vdots \end{vmatrix} \; \begin{vmatrix} x_1^2 \\ x_2^2 \\ \vdots \end{vmatrix} \; \begin{vmatrix} x_1^3 \\ x_2^3 \\ \vdots \end{vmatrix} \right)}_{A} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix}}_{P} - \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}}_{Y} \Bigg\|^2$$

<u>Solution:</u> $\boxed{P = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_3 \end{pmatrix} = (A^T A)^{-1} A^T Y}$

$A_{N\times 4}$ $A^T A$ 4×4.

example: $(x_1, w_1, z_1, y_1), (x_2, w_2, z_2, y_2), \dots$

(x_N, w_N, z_N, y_N)

find the function

$$y = ax + bw + cz + d$$

that fits the data best.

minimize $$\left\| \underbrace{\left(X \,\middle|\, W \,\middle|\, Z \,\middle|\, \mathbb{1} \right)}_{A} \underbrace{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}}_{p} - \left(Y \right) \right\|^2$$

best $$\boxed{\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = p = \left(A^T A\right)^{-1} A^T Y}$$

Cost function $$\| Ap - Y \|^2 = G(p)$$

parameters

SSE

Sum of Squares of errors in model.

$$MSE = \frac{\|Ap - y\|^2}{N}$$

$$RMSE = \sqrt{\frac{\|Ap - y\|^2}{N}}$$

Other statistics things we can do with linear algebra:

① Centering data in a vector.
(so the mean is 0)

$X = (x_1, \dots, x_N)$ original

$\tilde{X} = (x_1 - \bar{x}, \dots, x_N - \bar{x})$

mean of x_j's $\left(\frac{1}{N}\sum_{j=1}^{N} x_j\right) = \bar{x}$.

$$\Rightarrow \tilde{X} = (x_1, \dots, x_N) - \bar{x}(1, 1, \dots, 1)$$

$$\tilde{x} = X - \bar{x}\mathbb{1}$$

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j = \frac{\sum_{j=1}^{N} x_j \cdot 1}{\sum_{j=1}^{N} 1 \cdot 1}$$

$$\bar{x} = \frac{X \cdot \mathbb{1}}{\mathbb{1} \cdot \mathbb{1}}$$

$$\text{Recentered } X = \tilde{X} = X - \left(\frac{X \cdot \mathbb{1}}{\mathbb{1} \cdot \mathbb{1}}\right)\mathbb{1}$$

projection of vector x onto the vector $\mathbb{1}$.

Now $\tilde{X} \cdot \mathbb{1} = 0$.

After we make the mean 0, we often standardize the data by making the standard deviation = 1.

$$\text{New } \tilde{\tilde{X}} = \frac{\tilde{X}}{\left(\text{stand dev of } \tilde{X}\right)} \qquad \sigma(t) = \sqrt{\frac{\sum (t_j - \bar{t})^2}{N}}$$

$$\tilde{\tilde{X}} = \frac{\tilde{X}}{\sqrt{N}\|\tilde{X}\|} \longleftarrow$$

$\tilde{\tilde{X}}$ will have mean 0 & standard deviation 1.

$$\tilde{\tilde{X}} = \frac{\tilde{X}}{\sigma} = \sqrt{\frac{1}{N}\sum \left(\frac{x_j}{\sigma}\right)^2}$$

$$\sigma_{\tilde{\tilde{X}}} = \sqrt{\frac{1}{N}\sum \frac{\tilde{x}_j^2}{\|\tilde{X}\|^2} \cdot N} = 1.$$

Correlation Coefficient = r

tells us how closely the data matches the model.

$\Longleftrightarrow$ Does Y vector find inside the $X, \mathbb{1}$ plane exactly?

$\| \|^2 = 0.$

$X, Y \to \tilde{X}, \tilde{Y}$

standardize so mean $= 0$, std $= 1$.

$\boxed{r = \cos\theta}$

θ = angle between normalized X & normalized Y.