LINEAR REGRESSION

- CLASSIFICATION $(x, f(x))$

  $f(x) \in \{0, 1\}$

  - HALFSPACES
  - DECISION TREES

- REAL-VALUED LABELS (X, Y) $Y \in \mathbb{R}$

X, Y two RANDOM VARIABLES

WE WANT TO PREDICT THE VALUE/LABEL

WE GET TO SEE X.

- WE WANT TO PREDICT Y; WE DON'T SEE X

$\underline{(X, Y)} \underline{\sim \mathcal{D}}$ OPTIMAL GUESS FOR Y IS $\mathbb{E}[Y]$

- MEASURE OUR LOSS USING SQUARE-LOSS: $(\text{PREDICTION} - Y)^2$

- WE OBSERVE X WE WANT TO PREDICT Y

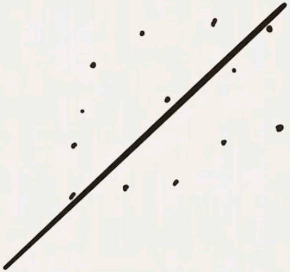
OPTIMAL PREDICTION $\mathbb{E}[Y \mid X] = f(X)$

REGRESSION FUNCTION

OBSTACLE: $f(X)$ COULD BE UNKNOWN OR VERY HARD TO COMPUTE!

LINEAR REGRESSION ASKS THE FOLLOWING QUESTION:

GIVEN X WHAT LINEAR FUNCTION OF X SHOULD WE USE TO PREDICT Y?



- WE WANT TO LEARN COEFFICIENTS $\underline{\beta_0}$ AND $\underline{\beta_1}$

TO MINIMIZE

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\left(y - (\beta_0 + \beta_1 x)\right)^2\right]$$

DRAW A TRAINING SET OF SIZE m

$$(x^1, y^1), \ldots\ldots, (x^m, y^m)$$

"SIMPLE LINEAR REGRESSION"

$$\min_{\beta_0, \beta_1} \underbrace{\frac{1}{m} \sum_{j=1}^{m} \left(y^j - (\beta_0 + \beta_1 x^j)\right)^2}_{\ell}$$

· TAKE DERIVATIVE W.R.T. β_0, β_1 SET THEM EQUAL TO 0.

$$\frac{\partial \ell}{\partial \beta_0} = \sum_{j=1}^{m} \left(y^j - \beta_0 - \beta_1 x^j\right)(-2) = 0$$

$$\frac{\partial \ell}{\partial \beta_1} = \sum_{j=1}^{m} \left(y^j - \beta_0 - \beta_1 x^j\right)(-2x^j) = 0$$

$$\frac{1}{m}\sum_{j=1}^{m}\left(y^j - \beta_0 - \beta_1 x^j\right) = 0$$

SOLVE FOR β_0

$$\underbrace{\frac{1}{m}\sum_{j=1}^{m}\left(y^j - \beta_0 - \beta_1 x^j\right)\left(x^j\right)}_{} = 0$$

β_0 IN TERMS OF β_1, $\bar{y}$, $\bar{x}$;

AVG y VAL

AVG X VALUE

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

β_1 WILL NOT INVOLVE β_0

$$\boxed{\beta_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2}}$$

$$\to \overline{xy} - \beta_0 \bar{x} - \beta_1 \overline{x^2}$$

$$\overline{xy} - \bar{x}\,\bar{y} + \beta_1 (\bar{x})^2 - \beta_1 \overline{x^2}$$

$$\beta_1 = \frac{COV(X,Y)}{VAR(X)}$$

$$COV(X,Y) = \mathbb{E}[X\cdot Y] - E[X]\cdot E[Y]$$

REGRESSION WITH MULTIPLE VARIABLES.

$x \in \mathbb{R}^n$ $y \in \mathbb{R}$ FITTING A LINE TO n-dimensional DATA.

X ← MATRIX X IS GOING TO BE AN $\underline{m \times n}$ MATRIX

- m ROWS ← EACH ROW IS EQUAL TO X^i DRAWN FROM $\mathcal{D}$
- n columns ← EACH point IS IN $\mathbb{R}^n$

$y \in \mathbb{R}^m$ ← LABELS FOR THESE m POINTS.

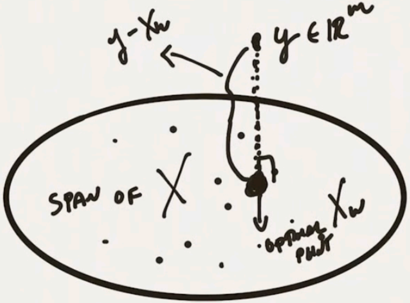
GOAL: FIND A VECTOR $w \in \mathbb{R}^n$ $\|X \cdot w - y\|_2^2$

$x^1 = x_1^1, \ldots, x_n^1$ y^1

$$\left(y - (x_1^1 w_1 + \cdots + x_n^1 w_n)\right)^2$$

$$\min_{w} \|Xw - y\|_2^2$$

Xw IS A VECTOR IN THE SPAN OF THE COLUMNS OF X.



VECTOR $y - Xw$ IS ORTHOGONAL TO X

$$X^T \cdot (y - Xw) = 0$$

$$X^Ty - X^TXw = 0$$

1st ISSUE: WHAT IF (X^TX) IS NOT INVERTIBLE?

$$X^Ty = X^TXw$$

NORMAL EQUATIONS

$$(X^TX)^{-1}X^Ty = w$$

"PSEUDO-INVERSE"

WHAT IS THE RUNNING TIME FOR COMPUTING w?

$$O(n^3 + m \cdot n^2)$$

MAXIMUM LIKELIHOOD

ASSUMPTION "SIMPLE LINEAR REGRESSION CASE"

X ; ASSUME $Y = \underline{\beta_0} + \underline{\beta_1} x + \epsilon$ ← RANDOM NOISE VARIABLE

$\epsilon \sim N(0, \sigma^2)$

DRAWN $x^1, \ldots, x^m$ AND $y^1, \ldots y^m$

WE WANT TO UNDERSTAND: FOR A FIXED CHOICE OF β_0 AND β_1 (σ^2 IS KNOWN)

WHAT IS THE PROBABILITY THAT WE SEE $(x^1, y^1) \ldots (x^m, y^m)$

LET'S WRITE DOWN LIKELIHOOD FUNCTION

PROBABILITY OF SEEING TRAINING SET GIVEN A CHOICE β_0 AND β_1 OF OUR PARAMETERS

$$\prod_{i=1}^{m} p(y^i \mid x^i; \beta_0, \beta_1) =$$

$$\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y^i - (\beta_0 + \beta_1 x^i))^2}{2\sigma^2}}$$

LIKELIHOOD OF OUR TRAINING SET.

CHOOSE β_0 AND β_1 THAT MAXIMIZES THIS LIKELIHOOD

$L(b_0, b_1)$

INSTEAD OF DIRECTLY MAXIMIZING LIKELIHOOD

WE WILL MAXIMIZE LOG-LIKELIHOOD

$$\log\left(L(\beta_0,\beta_1)\right) = \log \prod_{i=1}^{m} p(y^i \mid x^i; \beta_0, \beta_1)$$

$$= \sum_{i=1}^{m} \log\left(p(y^i \mid x^i; \beta_0, \beta_1)\right)$$

$$= \underline{-\frac{m}{2}\log 2\pi} \; \underline{- m\log\sigma} \; \underline{- \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y^i - (\beta_0 + \beta_1 x^i)\right)^2}$$

LEAST-SQUARES ESTIMATE
FOR SIMPLE LINEAR REGRESSION

TWO INTERPRETATIONS FOR COEFFICIENTS IN LINEAR REGRESSION:

- GEOMETRIC; COEFFICIENTS OF THE LINE THAT MINIMIZES SQUARED DISTANCE FROM LINE TO OUR LABELS

- STATISTICAL: COEFFICIENTS GIVE YOU THE MAXIMUM LIKELIHOOD ESTIMATOR FOR A TRAINING SET GENERATED $y \sim N(\beta_0 + \beta_1 x, \epsilon)$