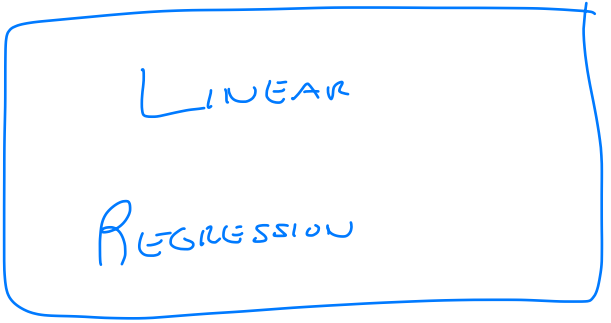
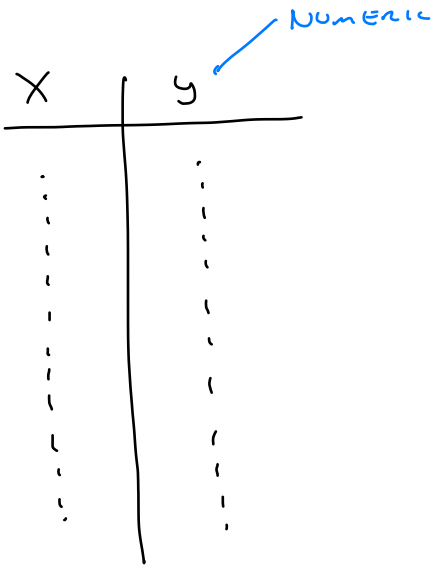
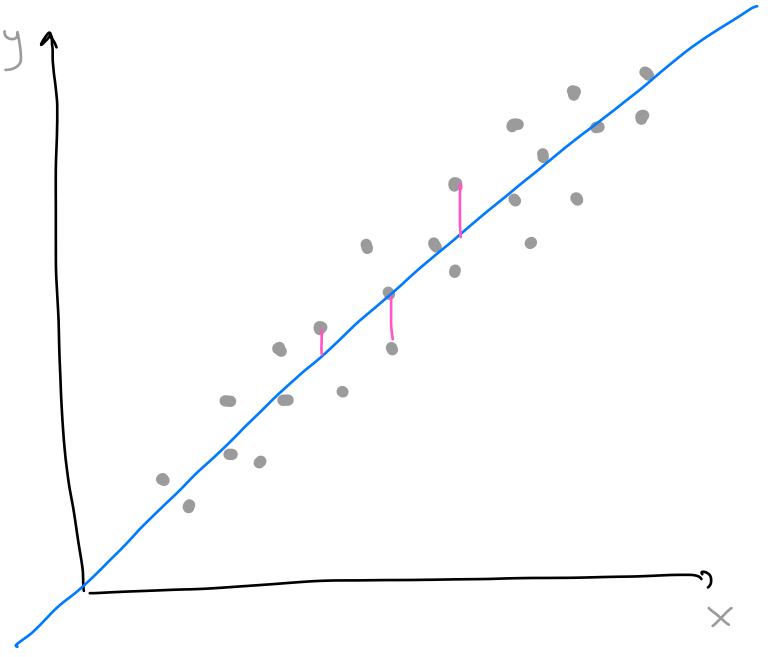
CS 307

Spring 2024

Dalpiaz




Numeric
X
y


y
x

Goal of Regression, for ML:

Make good predictions about

a numeric target variable

for new input feature data

STATISTICAL GOAL OF REGRESSION?

TO LEARN THE CONDITIONAL DISTRIBUTION

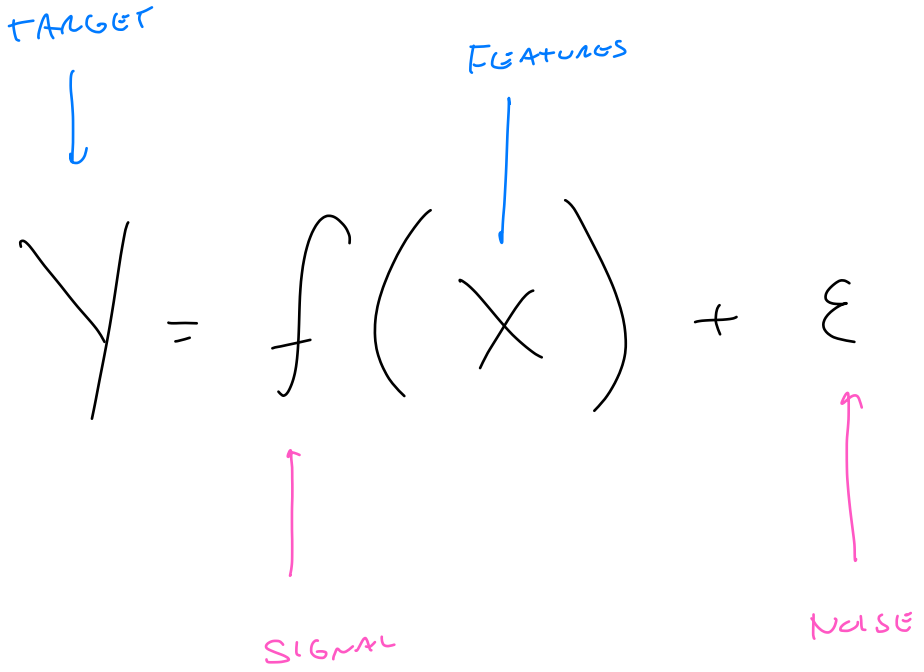
OF Y GIVEN X

SETTING FOR LEARNING

$$\mathbb{E}[Y \mid X = x]$$

CONDITIONAL MEAN OF Y GIVEN X

MEDIAN


TARGET
FEATURES
$Y = f(X) + \varepsilon$
SIGNAL
NOISE

NONPARAMETRIC METHODS

LEVERAGE "CLOSENESS" OF DATA

PARAMETRIC METHOD

MAKE STRONG ASSUMPTIONS ABOUT SHAPE OF f

LINEAR MODEL ASSUMPTIONS

LINEARITY

INDEPENDENCE

~~NORMAL~~

~~EQUAL~~

$H_0: \beta_1 = 0$

CI

PI

$$Y = f(X) + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$Y \mid X = x \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

$$\mathbb{E}[Y \mid X = x] = \mu(x) = \beta_0 + \beta_1 x$$

$$Y = f(X) + \varepsilon$$

WANT f SUCH THAT

$f(X)$ IS CLOSE TO Y

$$\min \left(Y - f(X)\right)^2$$

$$f(x) = \mathbb{E}[Y \mid X = x] = m(x)$$

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

x	y
.	.
.	.
.	.
.	.
.	.
.	.
.	.
.	.
.	.
.	.

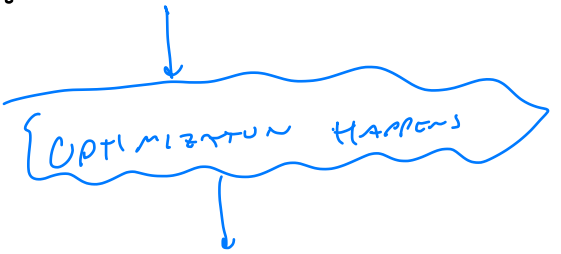
$(x_i, y_i) \quad i = 1, \ldots, n$

$$\min_{m} \sum_{i=1}^{n} \left(y_i - m(x_i)\right)^2$$

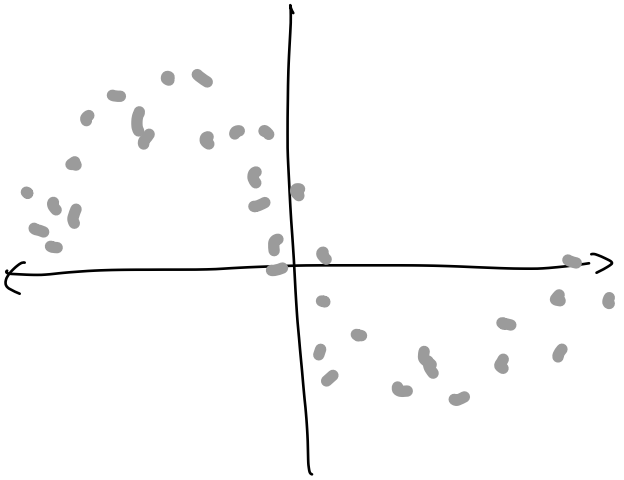
$$m(x) = \beta_0 + \beta_1 x$$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

"LEAST SQUARES"



$\tilde{\beta}_0, \tilde{\beta}_1$



X	y
⋮	⋮

FEATURE ENGINEERING →

sin(x)	y
⋮	⋮

Ex: $Y = \beta_0 + \beta_1 x + \varepsilon \dots$

$\sin(x)$

Multiple Linear Regression

x_1	x_2	x_3	y
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

$$\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

"LINEAR COMBINATION"

NESTED MODELS

$$\mu(x) = \beta_0$$

$$\mu(x) = \beta_0 + \beta_1 x_1, \quad \mu(x) = \beta_0 + \beta_2 x_2, \quad \mu(x) = \beta_0 + \beta_3 x_3$$

$$\mu(x) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\text{FIRST ORDER}} + \underbrace{\beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2}_{\text{QUADRATIC}} + \underbrace{\beta_7 x_1 x_2 + \beta_8 x_2 x_3 + \beta_9 x_1 x_3}_{\text{INTERACTION}}$$

SECOND ORDER

PARAMETRIC

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

MODEL PARAMETERS

LEARNED FROM DATA

NON PARAMETRIC

KNN, $K = 3$

TUNING PARAMETER

DEFINE HOW TO LEARN