# Generalized Linear Regression and Bias-Variance Tradeoff

HW0 : Due today 11:59 PM

HW1 : Release today

Due 4/21 11:59PM

# Process

$n$: # data

$x_i$ : feature $\in \mathbb{R}^d$

$y_i$ : label

Collect a **data set** $\{(x_i, y_i)\}_{i=1}^{n}$

Decide on a **model** function $f(x) \approx y$, $f(x) = x^T w$

Find the function which fits the data best

- **Choose a loss function** quadratic loss $(f(x) - y)^2$
- **Pick the function which minimizes loss on data** find $f$

Use function to make prediction on new examples

$x_{new}$ $f(x_{new}) \approx y_{new}$

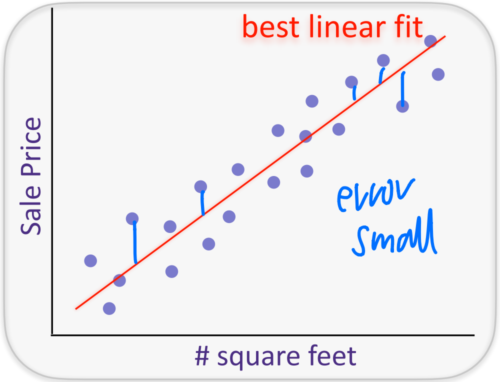
# The regression problem

$y$ is continuous / real number

Given past sales data on zillow.com, predict:

$y$ = **House sale price**

$x$ = **{# sq. ft., zip code, date of sale, etc.}**




Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis: linear

$y_i \approx x_i^T w$ for $w \in \mathbb{R}^d$

Loss: quadratic / least square

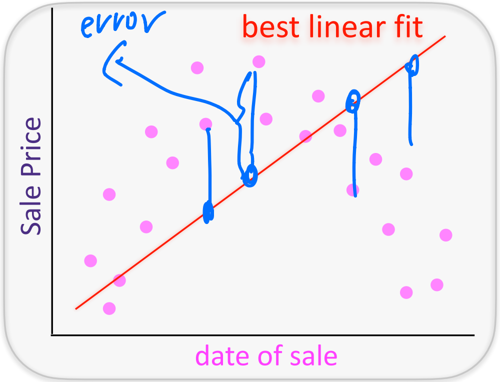
$$\min_w \sum_{i=1}^n (y_i - x_i^T w)^2$$

# The regression problem

Given past sales data on zillow.com, predict:

$y$ = **House sale price** *from*

$x$ = **{# sq. ft., zip code, date of sale, etc.}**




change hypothesis

poor fit

1) not a linear relationship

2) feature is not informative

change feature

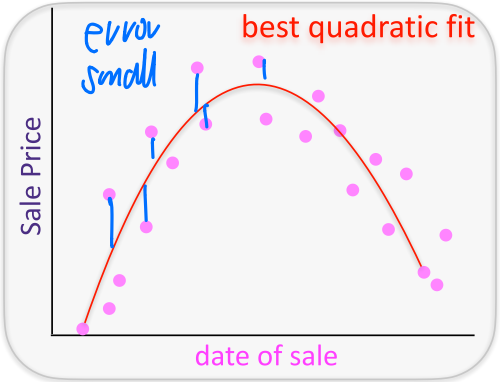
- - - - - -

# Quadratic Regression

Given past sales data on zillow.com, predict:

$y$ = **House sale price**

$x$ = **{# sq. ft., zip code, date of sale, etc.}**




Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis: quadratic

$$y_i \approx \sum_{j=1}^{d} \Big( \underbrace{x_{ij} \cdot w_{j,1}}_{\text{linear term}} + \underbrace{x_{i,j}^2 \cdot w_{j2}}_{\text{quadratic term}} \Big)$$

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix}$$

# Polynomial regression

Given past sales data on zillow.com, predict:

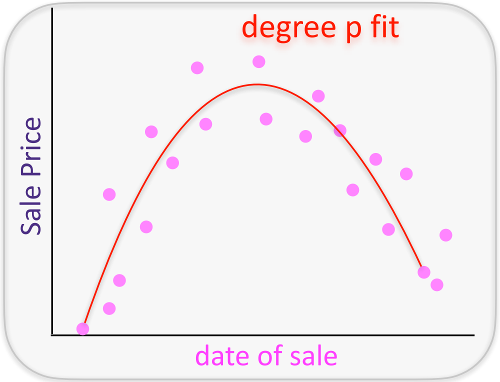
$y$ = **House sale price**

$x$ = **{# sq. ft., zip code, date of sale, etc.}**

dim of $W$ $\quad j: 1 \cdots d \quad l: 1 \cdots p \quad (d \cdot p)$

$d \quad \boxed{w_{j,l}} \quad W$

$p$




Training Data: $\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d \quad y_i \in \mathbb{R}$

Hypothesis: degree-$p$ polynomial

$$y_i \approx \sum_{j=1}^{d} \sum_{l=1}^{p} x_{i,j}^{l} \cdot w_{j,l}$$

if $p=1$ linear

$p=2$ quadratic

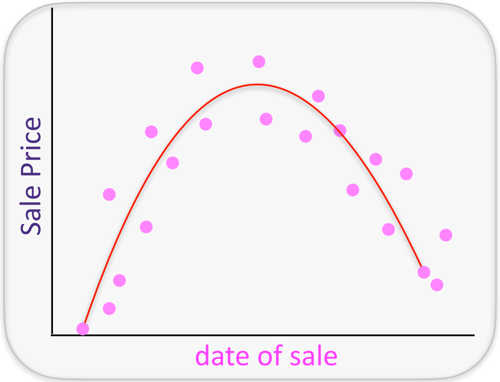
$$x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,d} \end{pmatrix}$$

# Generalized linear regression

Given past sales data on zillow.com, predict:

$y$ = **House sale price**

$x$ = **{# sq. ft., zip code, date of sale, etc.}**




Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Hypothesis: generalized linear regression

choose function $h: \mathbb{R}^d \to \mathbb{R}^q$

$x \to h(x)$

original feature → rich feature (high-dim)

$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^q$

linear in $h(x_i)$

# Generalized Linear Regression

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{id} \end{pmatrix}$$

Training Data: $\{(x_i, y_i)\}_{i=1}^n$ $\quad x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

Transformed data: $h(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_q(x) \end{pmatrix}$

Hypothesis: $y_i \approx h(x_i)^T w$

degree-p poly: $y_i = \sum_{j=1}^{d} \sum_{l=1}^{p} x_{i,j}^{l} \cdot w_{j,l}$

example 1: degree-p polynomial

$$h(x_i) = \begin{pmatrix} x_{i,1} \\ x_{i,2} \\ \vdots \\ x_{i,d} \\ x_{i,1}^2 \\ \vdots \\ x_{i,d}^2 \\ \vdots \\ x_{i,1}^p \\ \vdots \\ x_{i,d}^p \end{pmatrix} \in \mathbb{R}^{dp}$$

linear term

quadratic term

degree-p term

$w \in \mathbb{R}^{dp}$

Example 2:

generate vectors $\{u_j\}_{j=1}^q \subset \mathbb{R}^d$

$$h_j(x) = g(u_j^T x) = \begin{cases} (u_j^T x)^2 \\ 1/\exp(u_j^T x) \\ \cos(u_j^T x) \end{cases}$$

non-linear

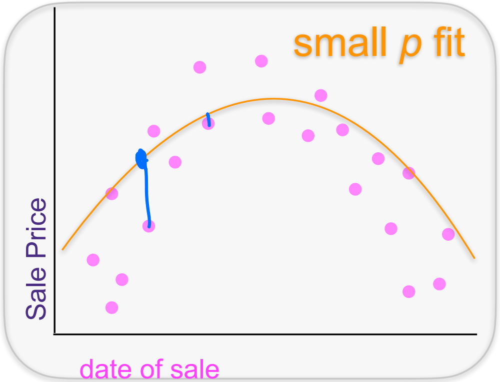
Loss: $\min_w \sum_{i=1}^n (y_i - h(x_i)^T w)^2$

# The regression problem

Training Data:

$$\{(x_i, y_i)\}_{i=1}^n \qquad x_i \in \mathbb{R}^d, \quad y_i \in \mathbb{R}$$

p=2, simple model




Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

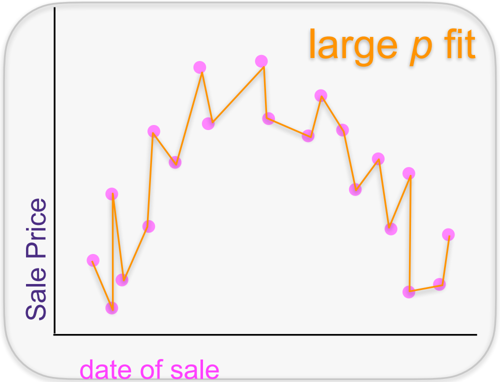
larger p → higher degree of freedom
in general p >= n, can fit all data

# The regression problem

Training Data: $\{(x_i, y_i)\}_{i=1}^n$

$x_i \in \mathbb{R}^d$
$y_i \in \mathbb{R}$

p >= 10, complex model




fit all data: 0 training error

Transformed data:

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \\ \vdots \\ h_p(x) \end{bmatrix}$$

Hypothesis: linear in h

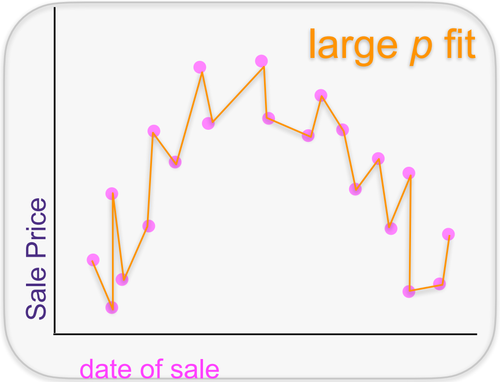
$$y_i \approx h(x_i)^T w \quad w \in \mathbb{R}^p$$

Loss: least squares

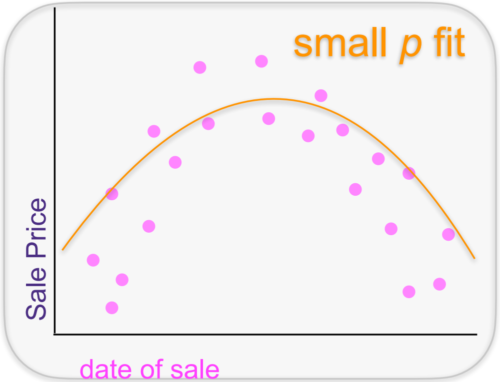
$$\min_w \sum_{i=1}^n \left(y_i - h(x_i)^T w\right)^2$$

# Which is better?

A: large p




B: small p



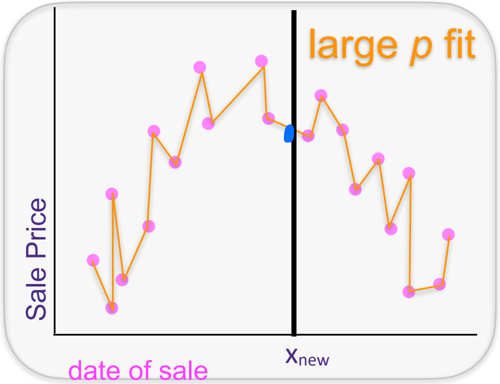

0 training error
non-smooth

>0 training error
smooth

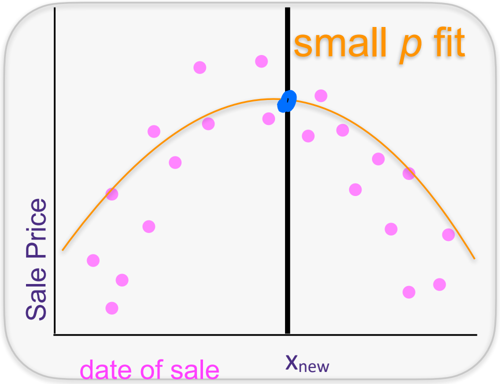
Q: how to measure the performance of a predictor?

# Predicting sale price for a new house: A vs B

A: large p




B: small p

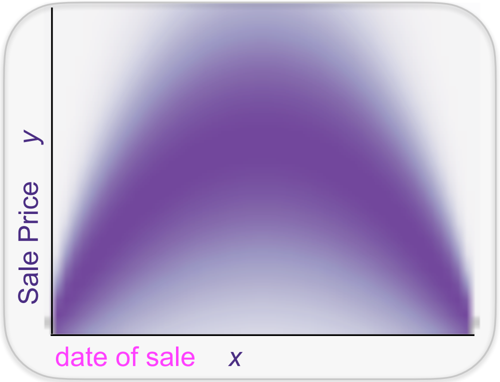



Our goal is to predict prices for new houses

# Average Accuracy

joint distribution of $(X, Y)$

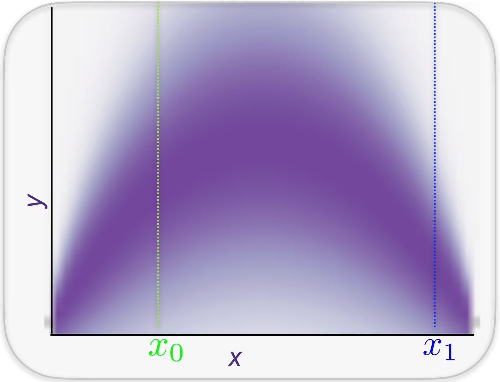
$$P_{XY}(X = x, Y = y)$$




On *average* over a house
drawn from this distribution,
we want to make a good prediction.

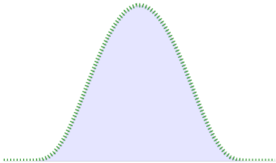
# Goal: predict future sale prices

$P_{XY}(X = x, Y = y)$



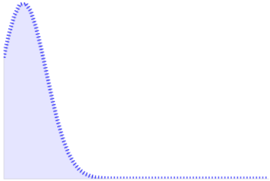

conditional distribution

$P_{XY}(Y = y|X = x_0)$



$f(x_0)$ small error on $P_{XY}(Y=y|X=x_0)$

$P_{XY}(Y = y|X = x_1)$



$f(x_1)$

# Statistical Learning

$$P_{XY}(X = x, Y = y)$$

**Goal: Predict Y given X**

**Find a function η that minimizes**

$$\mathbb{E}_{XY}[(Y - \eta(X))^2] = \mathbb{E}_X\left[\mathbb{E}_{Y|X}[(Y - \eta(x))^2 | X = x]\right]$$

$$\eta(x) = \arg\min_c \mathbb{E}_{Y|X}[(Y - c)^2 | X = x] = \mathbb{E}_{Y|X}[Y | X = x]$$

Under LS loss, optimal predictor: $\eta(x) = \mathbb{E}_{Y|X}[Y | X = x]$