2024-09-14

Machine Learning

Lecture 02: Linear Regression and Basic ML Issues

Linear regression

x_0	x_1	y
1	1	2
1	4	3
1	6	4

Dummy Variable x_0

$$y = w_0 + w_1 x_1 = \bar{w}^T \bar{x}$$

$$\bar{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \quad \bar{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

observed Predicted

$$J(\bar{w}) = \frac{1}{3}\left[(2 - (w_0 + w_1))^2 + (3 - (w_0 + 4w_1))^2 + (4 - (w_0 + 6w_1))^2\right]$$

Mean Squared error (MSE)

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial w_0} \\ \frac{\partial J}{\partial w_1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$

Matrix Notation

$$J(w) = \frac{1}{3}(\bar{y} - \bar{X}\bar{w})^T(\bar{y} - \bar{X}\bar{w})$$

3×1 ; 3×2 2×1 $\rightarrow$ 3×1

x_0	x_1	y
1	1	2
1	4	3
1	6	4

$\bar{X}$ 3×2 ; $\bar{y}$ 3×1

Design matrix

$$\nabla J = \bar{0} \quad (3\times 1)$$

$$\Rightarrow \bar{w} = (\bar{X}^T\bar{X})^{-1}\bar{X}^T y$$

2×3 3×2 $\rightarrow$ 2×2 ; 2×3 3×1 $\rightarrow$ 2×3 ; 2×1

ordinary least square (OLS) solution

closed-form / analytic solution

Probabilistic model for linear regression

So far, $y = \bar{w}^T \bar{x}$, y completely determined by $\bar{x}$

More reasonable model

$$y = \bar{w}^T \bar{x} + \text{impact of other minor factors}$$

y: random variable

$$= \bar{w}^T \bar{x} + \varepsilon$$

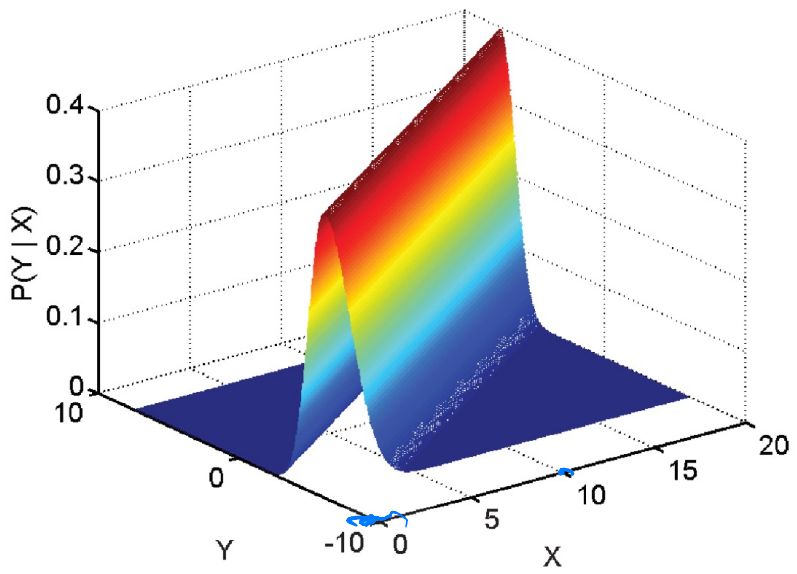
ε: noise, random variable

$\varepsilon \sim N(0, \sigma^2)$ (Central Limit Theorem)

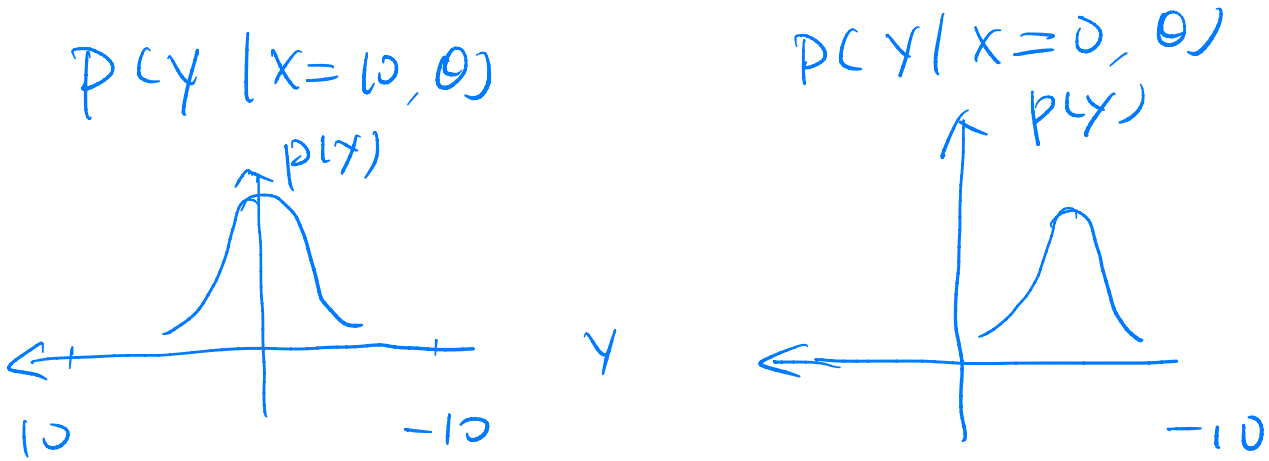
$$p(y \mid \bar{x}, \theta) = N(\bar{w}^T \bar{x}, \sigma^2)$$

$\bar{w}^T \bar{x}$: mean of y

$\theta = (\bar{w}, \sigma)$



$$p(y|x, \boldsymbol{\theta}) = \mathcal{N}(y|w_0 + w_1 x, \sigma^2)$$



$$D = \{\bar{x}_i, y_i\}_{i=1}^{N}$$

cross entropy

$$\min_{\theta} \; -\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(y_i \mid \bar{x}_i)$$

$$= -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - \bar{w}^T \bar{x}_i)^2}{2\sigma^2}} \right]$$

$$= -\frac{1}{N} \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2} \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{w}^T \bar{x}_i)^2$$

Assume $\sigma = 1$

MSE

Polynomial regression

x_0	x_1	y
1	1	2
1	4	3
1	6	4

$y = w_0 + w_1 x_1$ linear

More complex model

$$y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3$$

model

linear regression

⇓ feature transformation

Data

x_0	x_1	x_1^2	x_1^3	y
1	1	1	1	2
1	4	16	64	3
1	6	36	216	4

$$\bar{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \Rightarrow \phi(x) \begin{bmatrix} x_0 \\ x_1 \\ x_1^2 \\ x_1^3 \end{bmatrix}$$

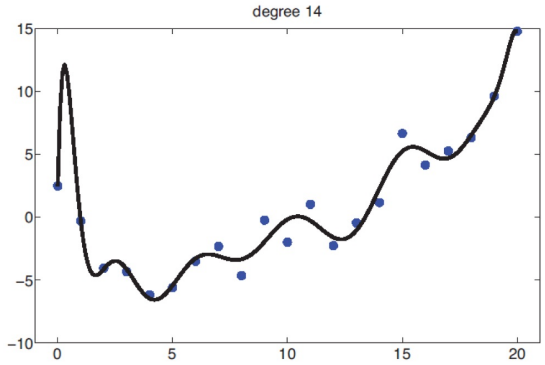
what d to use ?

Model selection

hyperparameter

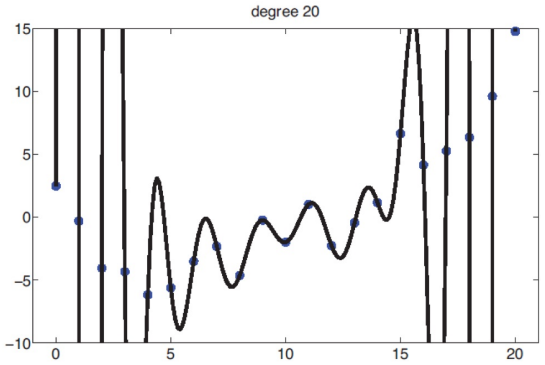
more complex

- Two examples with $d = 14$ and 20 and one feature $\mathbf{x} = (x)$.

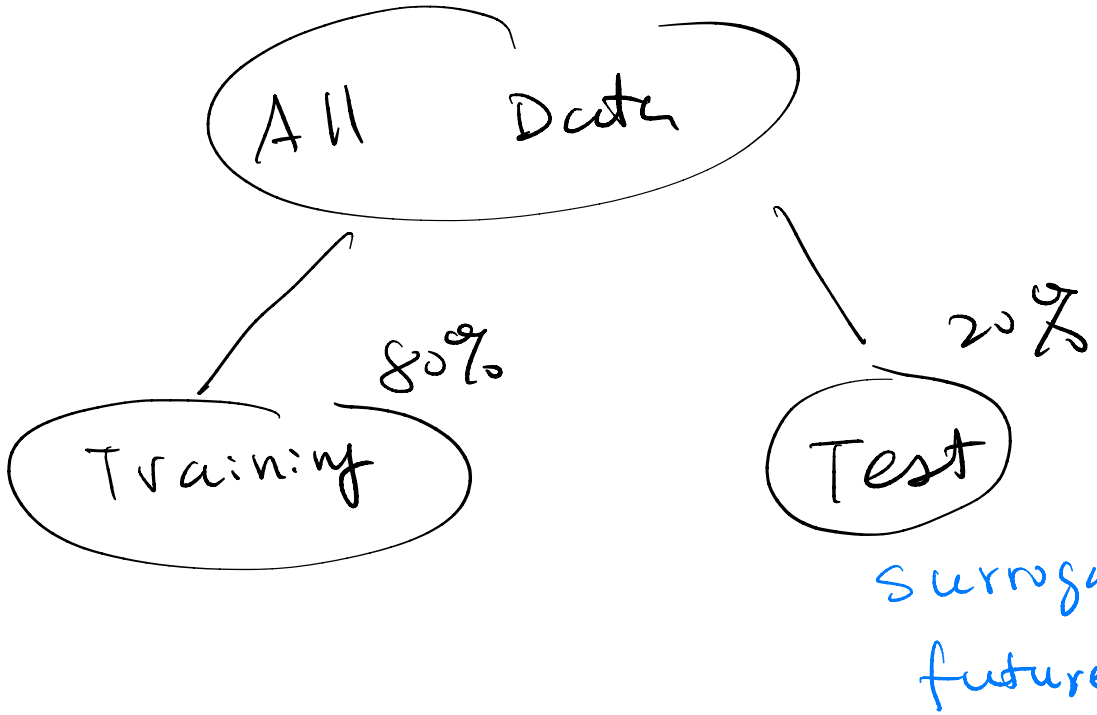


MSE small but not 0

✓ why?



MSE = 0



- We use a the error on a **test set** to measure how well a model generalize:

$$J^{(test)}(\mathbf{w}) = \frac{1}{N^{(test)}} ||\mathbf{y}^{(test)} - \mathbf{X}^{(test)}\mathbf{w}||_2^2$$

Want to make small

This is called the **test error** or the **generalization error**

Minimize

- In contrast, here is the **training error** we have been talking about so far:

$$J^{(train)}(\mathbf{w}) = \frac{1}{N^{(train)}} ||\mathbf{y}^{(train)} - \mathbf{X}^{(train)}\mathbf{w}||_2^2$$

Hypothesis Space

$Y = w_0 + w_1 x_1$

$S_1 = \{ Y = w_0 + w_1 x_1 \mid w_0, w_1 \in \mathbb{R}^1 \}$

$= \{ Y = 0.1 + 0.2 x_1,$

$Y = 0.5 - 3 x_1,$

$Y = -2 + 12 x_1, \}$

Task: Determine w_0, w_1

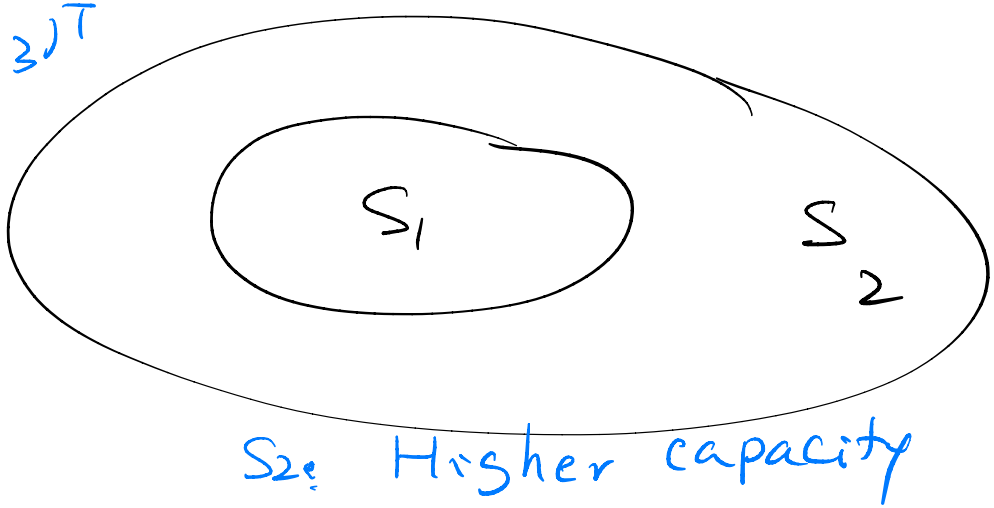
$\bar{w}_1 = (w_0, w_1)^T$

Task: pick one model from S_1

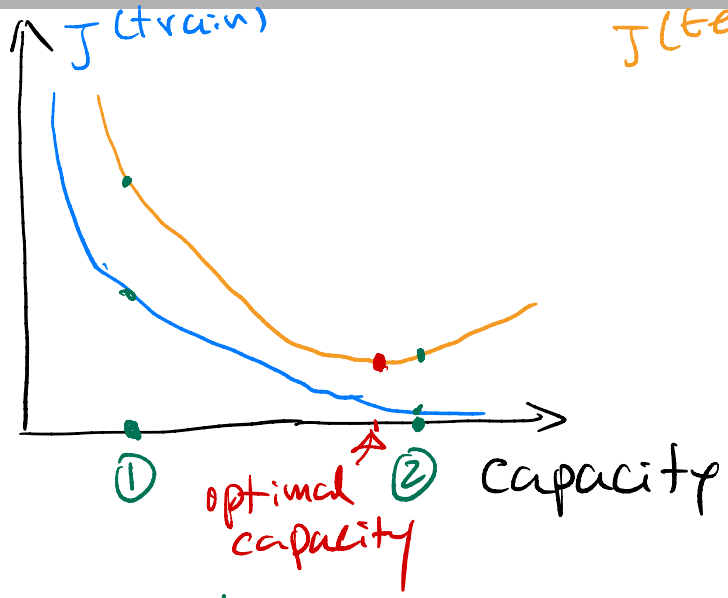
$S_2 = \{ Y = w_0 + w_1 x_1 + w_2 x_1^2 + w_3 x_1^3 \mid w_0, w_1, w_2, w_3 \in \mathbb{R}^1 \}$

$\bar{w}_2 = (w_0, w_1, w_2, w_3)^T$

$\min_{\bar{w}_1} J(\bar{w}_1) \geq \min_{\bar{w}_2} J(\bar{w}_2)$



S_2: Higher capacity



① Model too simple

* cannot fit pattern

* $J^{(train)}$, $J^{(test)}$ large

underfitting

Training data : pattern + noise 1

Test data : pattern + noise 2

② Model too complex

* fit pattern & noise 1

low $J^{(train)}$

* noise 2 ≠ noise 1

$J^{(test)}$ relatively large

Model Selection

pick the model with optimal capacity

* Validation

* Regularization

Validation

All data

- training
  - Training set : Determine model parameters
  - Validation : Determine hyper parameter
- Test
  - Test set : Testing

held out

d	Training Set 70%	Validation 30%
1	1.5	1.6
5	0.3	0.35
10	0.1	0.40