

Machine Learning

Prepared by Dr. Yasser Abdelhamid

DS 620
Machine Learning

Lecture 8





OUTLINE

- ◉ Logistic regression
- ◉ Bias
- ◉ Variance
- ◉ Overfitting
- ◉ Under-fitting
- ◉ Regularization
 - ◉ Lasso Regularization
 - ◉ Didge Regularization
- ◉ Lasso regression
- ◉ Ridge regression

BINARY CLASSIFICATION

- ❖ Binary classification is a type of classification where the target variable is a variable having (usually) the **0** value to represent negative result and **1** for positive result.
- ❖ target variable $y \in \{0,1\}$

MULTICLASS CLASSIFICATION

❖ Is a type of classification where the target variable may have **more than two categories**.

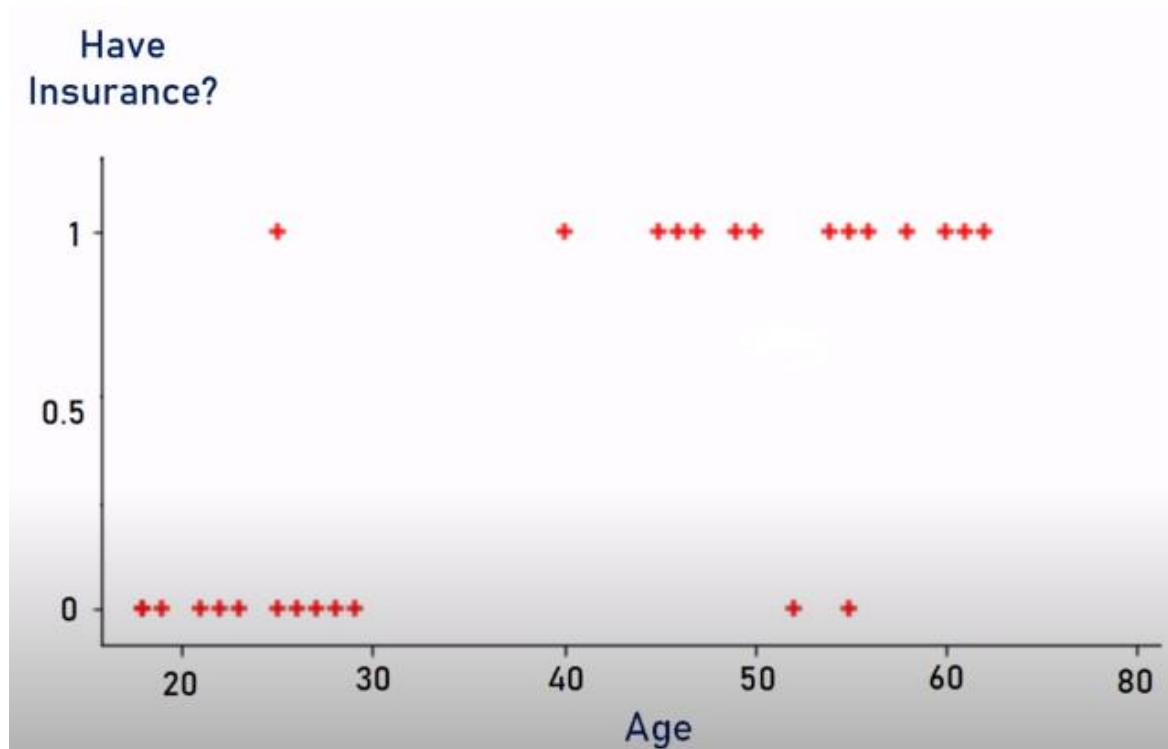
❖ Example:

- Election :
 - Democratic.
 - Republican.
 - Independent.

EXAMPLE: INSURANCE PREDICTION

Age	Bought insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1

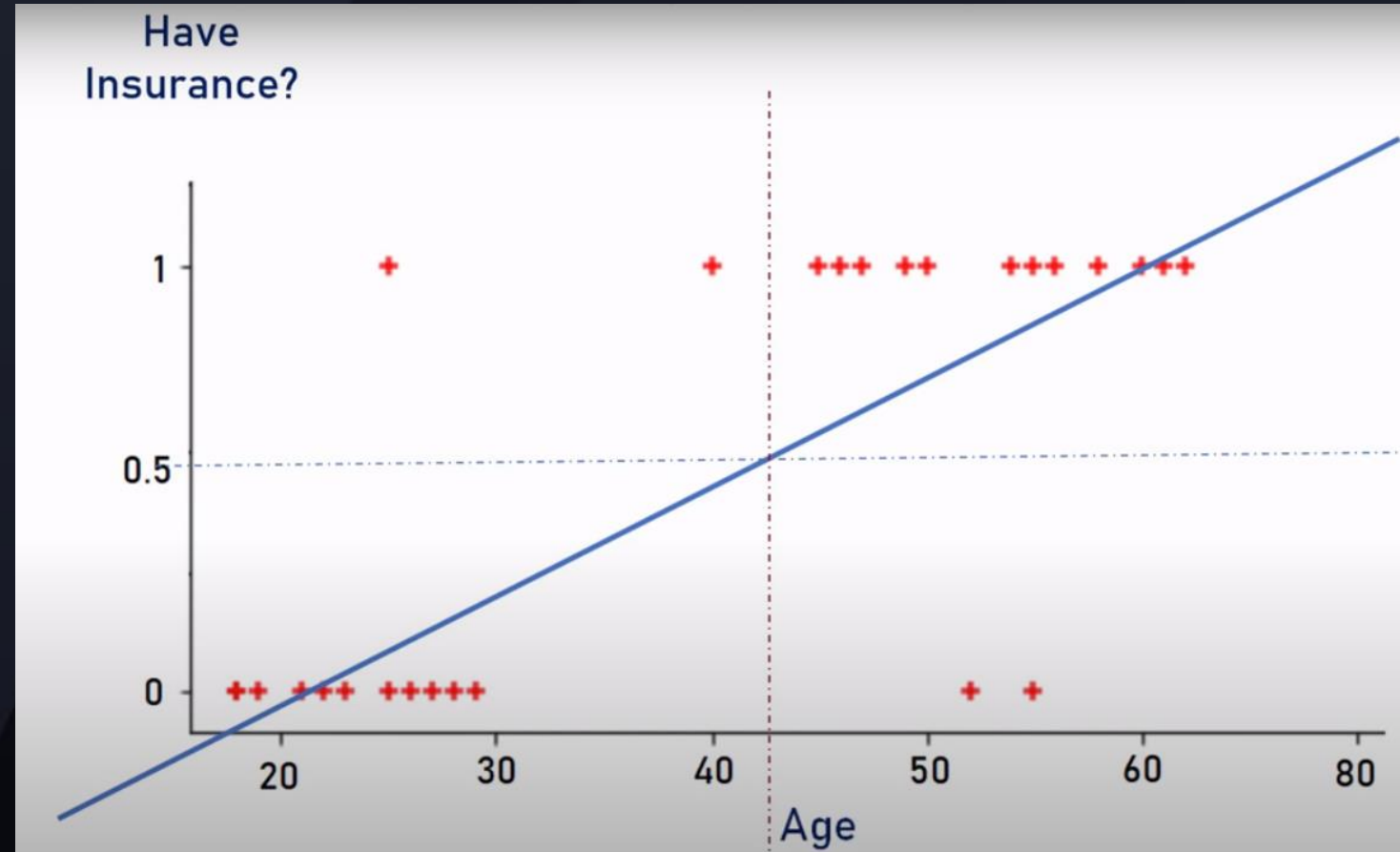
EXAMPLE: INSURANCE PREDICTION



Age	Bought insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1

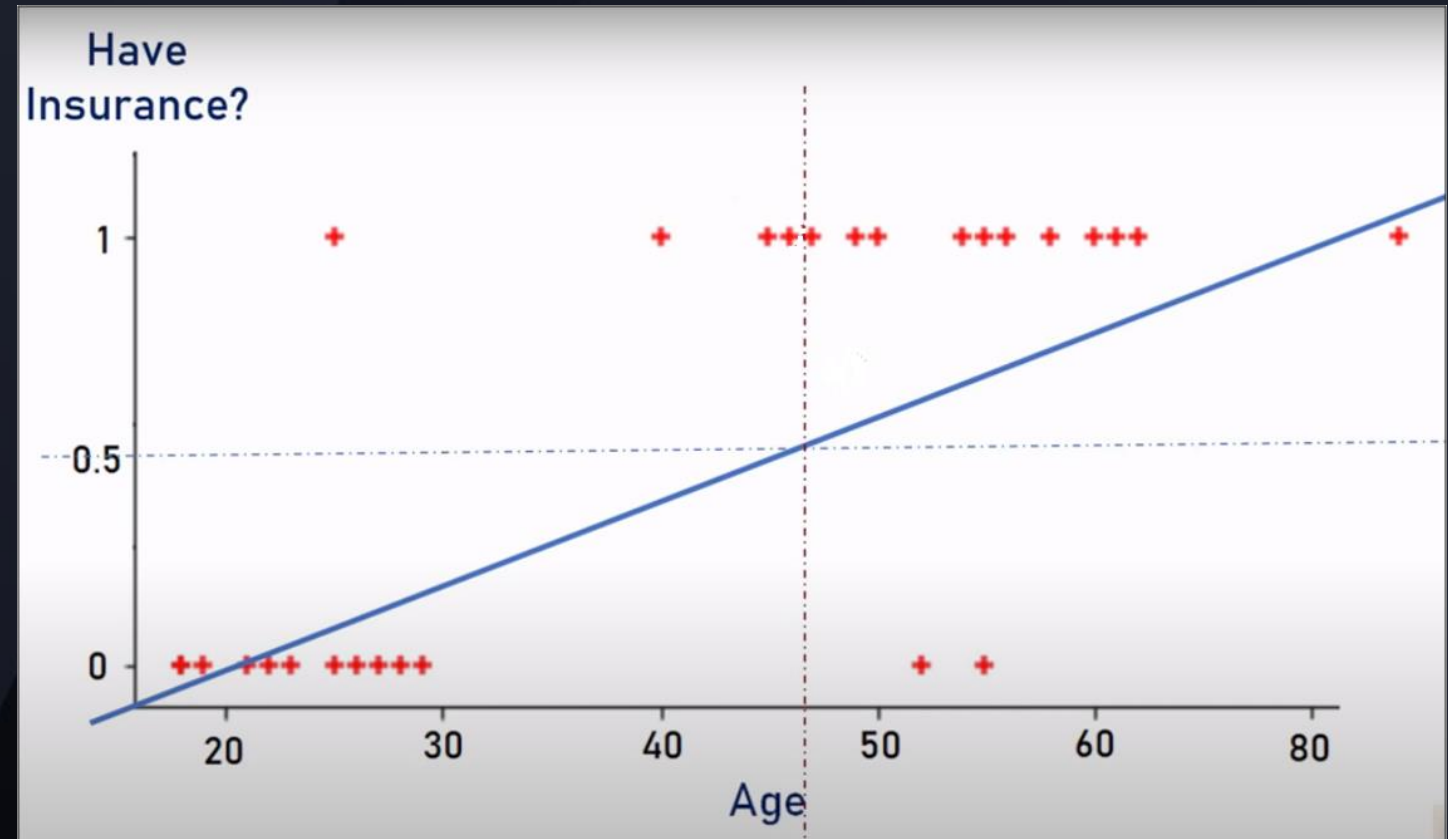
USING LINEAR REGRESSION

It looks that people with age ≥ 42 are likely to buy the insurance, while people with age < 42 will not buy the insurance.



USING LINEAR REGRESSION

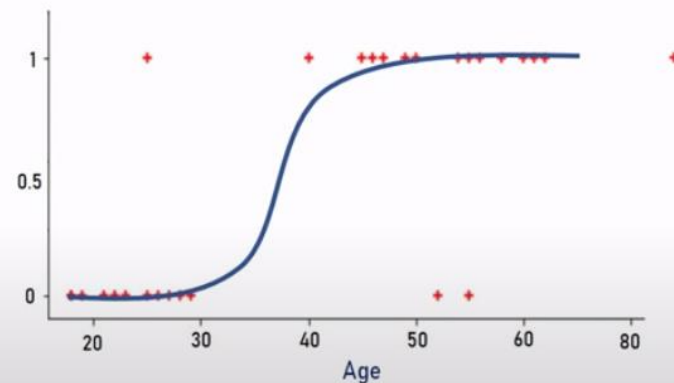
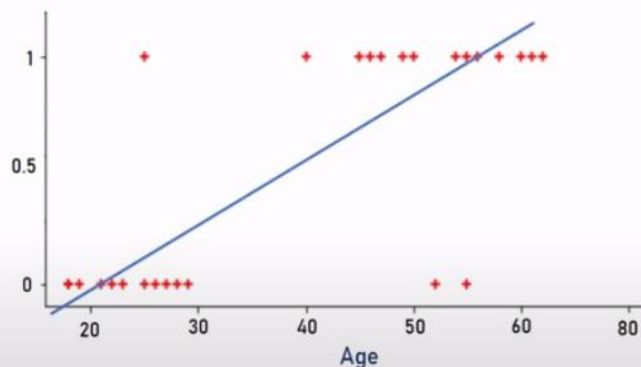
- By adding a new instance, and running linear regression, the threshold changes, giving worse hypothesis, and misclassifies more instances that were correctly classified without adding that new instance.
- y is either 0 or 1, but $h(x)$ may be <0 or >1
- we need $0 \leq h(x) \leq 1$



SIGMOID FUNCTION

- ❖ Sigmoid function converts its input to a range between 0 and 1.

$$\text{sigmoid}(z) = \frac{1}{1 + e^{-z}}$$



LOGISTIC REGRESSION MODEL

$$\diamond h_{\theta}(x) = \theta^T x$$

$$\diamond g(z) = \frac{1}{1+e^{-z}}$$

$$\diamond g(h_{\theta}(x)) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$

SIGMOID FUNCTION

$$\diamond p(y=1 | x; \theta) = \frac{1}{1 + e^{-\theta^T x}}$$

Prediction model

$$\diamond p(y_i=1 | x_i; \theta) = \begin{cases} 1, & \text{if } \frac{1}{1 + e^{-\theta^T x_i}} \geq 0.5 \\ 0, & \text{otherwise} \end{cases}$$

LOGISTIC REGRESSION MODEL

$$p(y_i=1 | x_i; \theta) = \frac{1}{1 + e^{-\theta^T x_i}}$$

$$p(y_i=0 | x_i; \theta) = 1 - \frac{1}{1 + e^{-\theta^T x_i}}$$

THE GENERAL FORM OF THE MODEL

$$p(y_i | x_i; \theta) = \left(\frac{1}{1 + e^{-\theta^T x_i}} \right)^{y_i} \cdot \left(\frac{1}{1 + e^{-\theta^T x_i}} \right)^{1 - y_i}$$

EXAMPLE: DEFAULT ON LOAN

A bank is building a prediction model for default customers depending on customer's credit score.

(a customer defaults on his loan when he fails to pay back a debt according to the initial arrangement).

EXAMPLE: DEFAULT ON LOAN

The bank is using the regression model:

$$h_{\vartheta}(x) = \vartheta_0 + \vartheta_1 x$$

using the regression model parameters:

$$\vartheta_0 = 0.73257, \quad \vartheta_1 = -9.9238 \times 10^{-5}$$

EXAMPLE: DEFAULT ON LOAN

To estimate the probability of default for a customer who has a credit score = 8000

$$h_{\vartheta}(x) = \vartheta_0 + \vartheta_1 x$$

With $\vartheta_0 = 0.73257$, and $\vartheta_1 = -9.9238 \times 10^{-5}$

$$h_{\vartheta}(8000) = 0.73257 + (-9.9238 \times 10^{-5} * 8000) = -0.06134$$

EXAMPLE: DEFAULT ON LOAN

$$p(\text{default}=1 | x; \theta) = \frac{1}{1 + e^{-h_{\theta}(x)}}$$

$$p(\text{default}=1 | 8000; [\vartheta_0=0.73257, \theta_1=-9.9238 \cdot 10^{-5}])$$

$$= \frac{1}{1 + e^{-(-0.06134)}}$$

EXAMPLE: DEFAULT ON LOAN

$$\diamond p(\text{default}=1 | x; \theta) = \frac{1}{1 + e^{(0.06134)}} = 0.4846$$

$$\diamond p(\text{default}=1) = 48.46\%$$

❖ This means that the customer **will not default on loan**, since $p(\text{default}=1) < 0.5$

LINEAR VS LOGISTIC REGRESSION

Linear regression

- Data is modeled using a linear function.
- Predicts a continuous value of y
- Output is interpreted as the value of predicted variable.

Logistic regression

- Maps the continuous values of input to the categorical values (classes) of y .
- Predicts the class of y .
- Output is interpreted as the probability of occurrence of event.

BIAS, VARIANCE, OVERFITTING, AND UNDERFITTING

BIAS

- ❖ Bias refers to the error introduced by approximating a real-world problem with a **simplified model**.
- ❖ It represents the model's tendency to consistently make certain assumptions or errors when predicting outcomes.

EXAMPLE

- ❖ A high bias model **oversimplifies** the problem and makes strong assumptions, leading to **Underfitting**.
- ❖ It may struggle to capture complex patterns and generalize well to unseen data.



HOW TO DETECT BIAS

- ❖ Common characteristics of a **high bias** model include:
 - High training error and
 - High testing error.
- ❖ It may exhibit systematic errors and have **difficulty learning from the training data**.

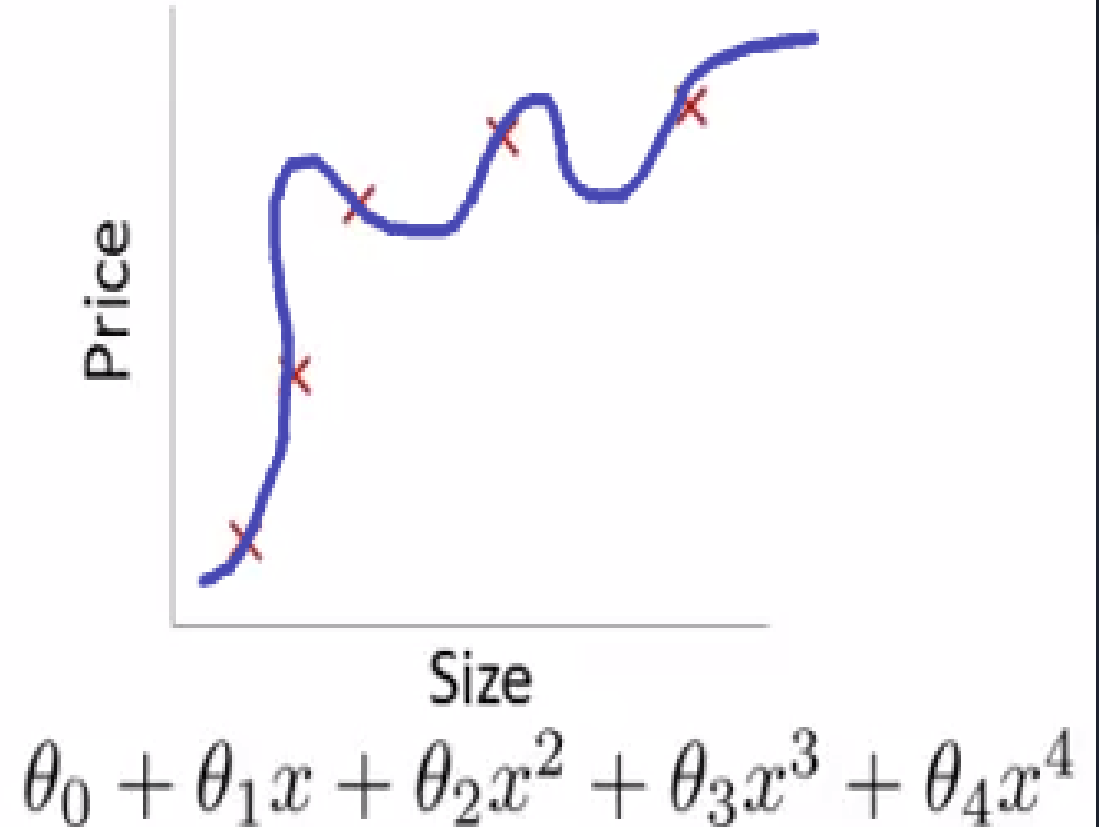


VARIANCE

- ❖ Variance refers to the variability or sensitivity of a model's predictions to fluctuations in the training data.
- ❖ It measures how much the model's predictions vary across different training sets.

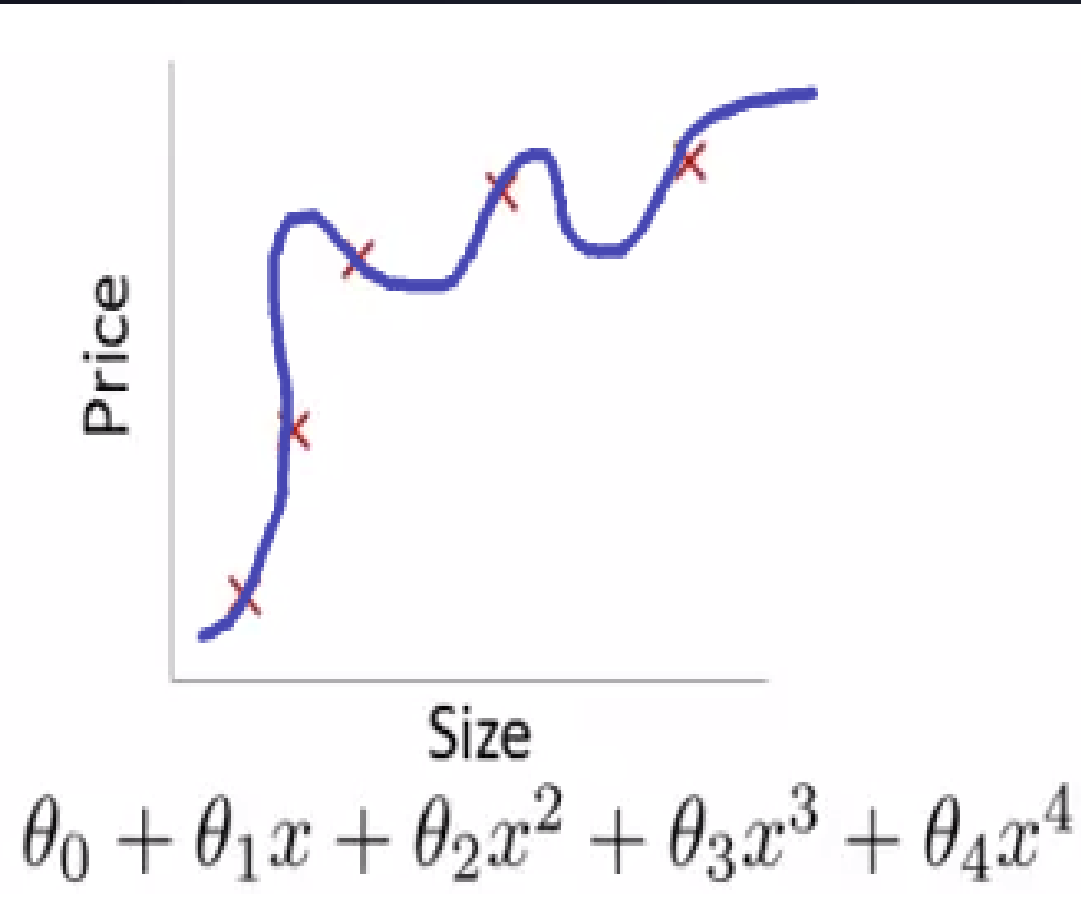
EXAMPLE

- ❖ A high variance model is **overly complex** and **captures noise** or random fluctuations in the training data.
- ❖ It tends to **memorize the training examples** and performs poorly on unseen data, **leading to overfitting**.



HOW TO DETECT VARIANCE

- ❖ Common characteristics of a high variance model include:
 - low training error
 - high testing error
- ❖ It may exhibit high sensitivity to the training data and struggle to generalize to new instances.



BIAS-VARIANCE TRADEOFF

- ❖ The bias-variance tradeoff is a fundamental concept in machine learning.
- ❖ It refers to the delicate balance between a model's ability to capture the underlying patterns (low bias) and its ability to generalize to new data (low variance).

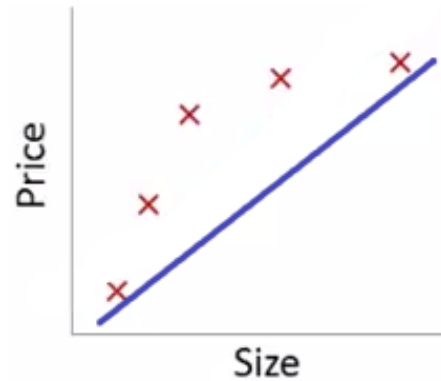
BIAS-VARIANCE TRADEOFF

- ❖ Ideally, we aim to find a model that minimizes both bias and variance, striking the right balance.
- ❖ However, reducing one often increases the other, making it a tradeoff.

HOW TO GET THE RIGHT BALANCE

- ❖ There are many techniques like:
 - Cross-validation,
 - Regularization, and
 - Model selection
- ❖ These techniques help manage the bias-variance tradeoff.

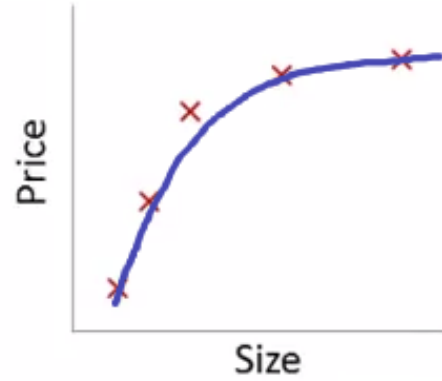
HOW TO GET THE RIGHT BALANCE



$$\theta_0 + \theta_1 x$$

Polynomial degree=1

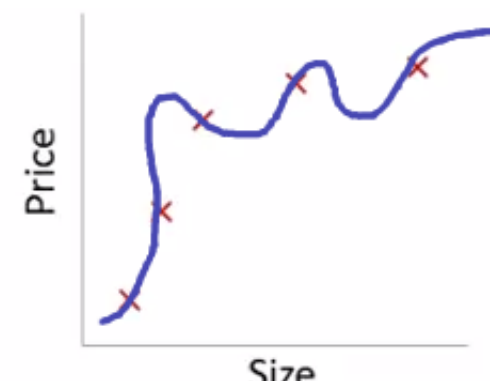
Under-fitting
High bias
High variance
Training accuracy low
Test accuracy low



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

Polynomial degree=2

Most generalizing
low bias
low variance
Training accuracy high
Test accuracy high



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

Polynomial degree=4

Over-fitting
low bias
High variance
Training accuracy high
Test accuracy low

REGULARIZATION

- ❖ Regularization is a technique used in machine learning to prevent overfitting and improve the generalization performance of models.
- ❖ It involves adding a **penalty term** to the loss function during training, which encourages the model to **avoid overly complex or extreme parameter values**.

REGULARIZATION

- ❖ The main idea behind regularization is to find a balance between fitting the training data well and avoiding excessive complexity.
- ❖ Overfitting occurs when a model becomes too complex and captures noise or random fluctuations in the training data, leading to poor performance on unseen data.
- ❖ Regularization helps in addressing this issue by imposing constraints on the model's parameters.

LASSO REGULARIZATION

- ❖ L1 Regularization (Lasso): L1 regularization adds a penalty term proportional to the absolute values of the model's parameters.
- ❖ It encourages sparsity by driving some of the parameters to zero, effectively performing feature selection.
- ❖ L1 regularization can be useful when dealing with datasets with many irrelevant or redundant features.

RIDGE REGULARIZATION

- ❖ L2 Regularization (Ridge): L2 regularization adds a penalty term proportional to the squared magnitudes of the model's parameters.
- ❖ It discourages large parameter values and encourages the model to distribute the importance of features more evenly.
- ❖ .

EXAMPLE

- ❖ Assume we have a dataset with two features, X_1 and X_2 , and we want to perform linear regression model.
- ❖ Assume the model equation
- ❖ $h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$
- ❖ The objective function for Lasso regression can be represented as:
Loss = Sum of squared errors + $\lambda * (|\theta_1| + |\theta_2|)$
- ❖ The regularization parameter, λ , controls the strength of the regularization. A higher λ value will result in more aggressive shrinking of the coefficients towards zero.

EXAMPLE

- ❖ During the training process, the model will adjust the coefficients (Θ_1 and Θ_2) to minimize this loss function.
- ❖ However, Lasso regularization **encourages sparsity in the coefficients**, meaning it tends to drive some of the coefficients to exactly zero.

EXAMPLE

- ❖ For example, let's say the Lasso regression model finds the following coefficients:
 - $\Theta_0 = 1.0$
 - $\Theta_1 = 0.9$
 - $\Theta_2 = 0.0$
- ❖ In this case, the coefficient of X_2 (Θ_2) becomes exactly zero, indicating that the feature X_2 is not contributing to the prediction.
- ❖ Lasso regression performs **feature selection** by automatically identifying and **eliminating less relevant features**.
- ❖ The resulting model equation would be:
- ❖ **$h_{\theta}(x) = 1.0 + 0.9 * X_1$**

LASSO REGRESSION COST FUNCTION

$$mse = \frac{1}{n} \sum_{i=1}^n (y_i - h_{\theta}(x_i))^2 + \lambda \sum_{i=1}^n |\theta_i|$$

RIDGE REGRESSION COST FUNCTION

$$ms\varepsilon = \frac{1}{n} \sum_{i=1}^n (y_i - h_{\theta}(x_i))^2 + \lambda \sum_{i=1}^n \theta_i^2$$



SUMMARY

- ◉ Logistic regression
- ◉ Bias
- ◉ Variance
- ◉ Overfitting
- ◉ Under-fitting
- ◉ Regularization
 - ◉ Lasso Regularization
 - ◉ Didge Regularization
- ◉ Lasso regression
- ◉ Ridge regression



Thank You

midocean.edu.km

