CS 60050
Machine Learning

An Introduction to

Bias-Variance Tradeoff

Some slides taken from course materials of Andrew Ng
and various sources available online

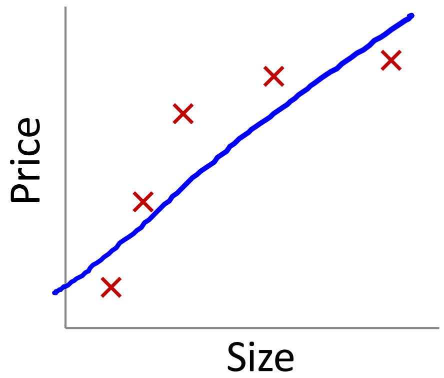
Overfitting

- How complex a hypothesis should we try to learn?
  - Complexity of a ML model ~ number of parameters

- Too simple hypothesis: the complexities in the data may not be captured → UNDERFITTING

- Too complex hypothesis: the learned hypothesis may fit the training set very well, but fail to generalize to new examples → OVERFITTING

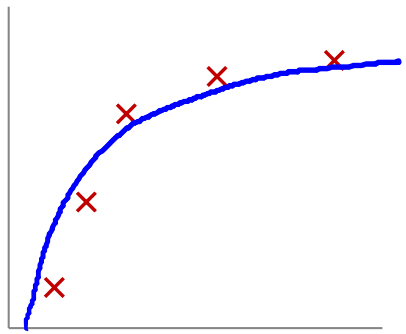
Bias vs. variance in linear regression



High bias
(underfitting)

d=1

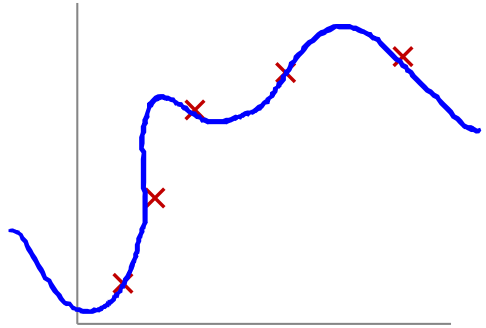
$\theta_0 + \theta_1 x$



"Just right"

d=2

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



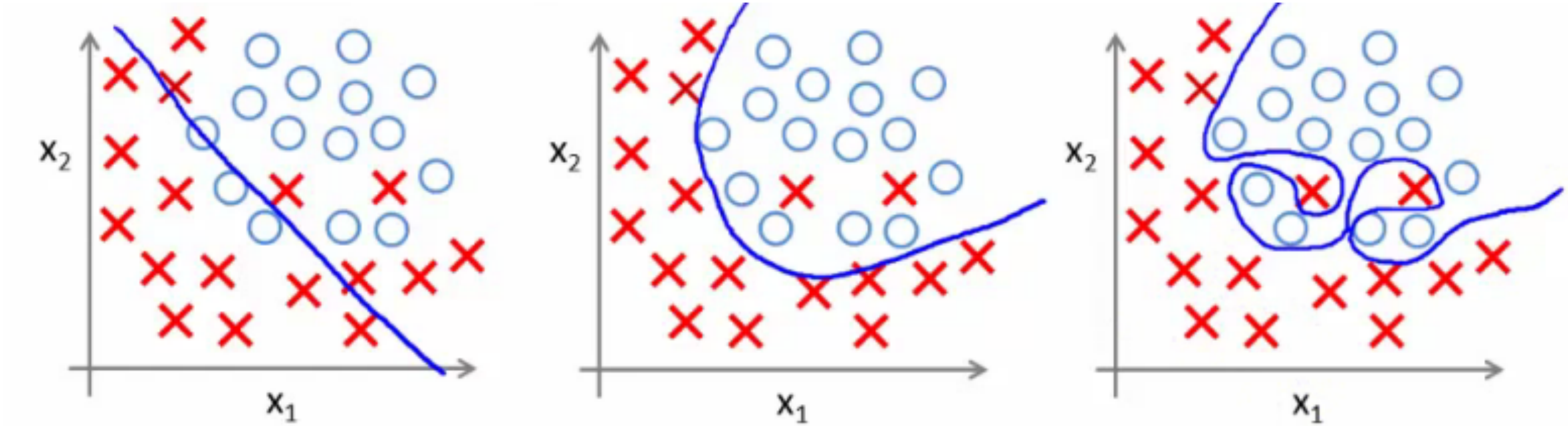
High variance
(overfitting)

d= 4

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

Bias vs. variance in logistic regression

Example: Logistic regression



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)

UNDERFITTING (high bias)

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2$
$+\theta_3 x_1^2 + \theta_4 x_2^2$
$+\theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2$
$+\theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2$
$+\theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots$

OVERFITTING (high variance)

Effect of noise

- Assume we have independent variables *x* that affect the value of a dependent variable *y*
- Let's denote the true dependence of *y* on *x* via function *f*
- Value of y can be affected by non-deterministic noise (e.g., due to measurement errors)

$$y = f(x) + \epsilon$$

- Noise is modelled by the random variable ε (assumed zero mean)

$$\mathbb{E}[\epsilon] = 0, \mathrm{var}(\epsilon) = \mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$$

Effect of the training data

- We will attempt to learn a function $\hat{f}$ that is as close to f as possible
- Mean squared error (MSE) is the average squared difference of a prediction $\hat{f}(x)$ from its true value y

$$\mathrm{MSE} = \mathbb{E}[(y - \hat{f}(x))^2]$$

- The function $\hat{f}$ that is learned depends upon the training data given
  - $\hat{f}$ can be different for different training data, and $\hat{f}(x)$ can change even though x is *fixed*
  - $\hat{f}$ is a random variable affected by the randomness in which we obtain training data

Bias and Variance

- Since $\hat{f}(x)$ is a random variable, we can talk of its expectation (over different realizations of training data)

- Bias*:* difference of the average prediction (*over different realizations of training data*) to the true function *f(x),* for a given unseen (test) point *x*

$$\text{bias}[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - f(x)$$

- Variance*:* mean squared deviation of $\hat{f}(x)$ from its expected value $\mathbb{E}[\hat{f}(x)]$ *over different realizations of training data*

$$\text{var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

Bias-Variance decomposition

- The formula that connects test MSE to bias, variance and irreducible error:

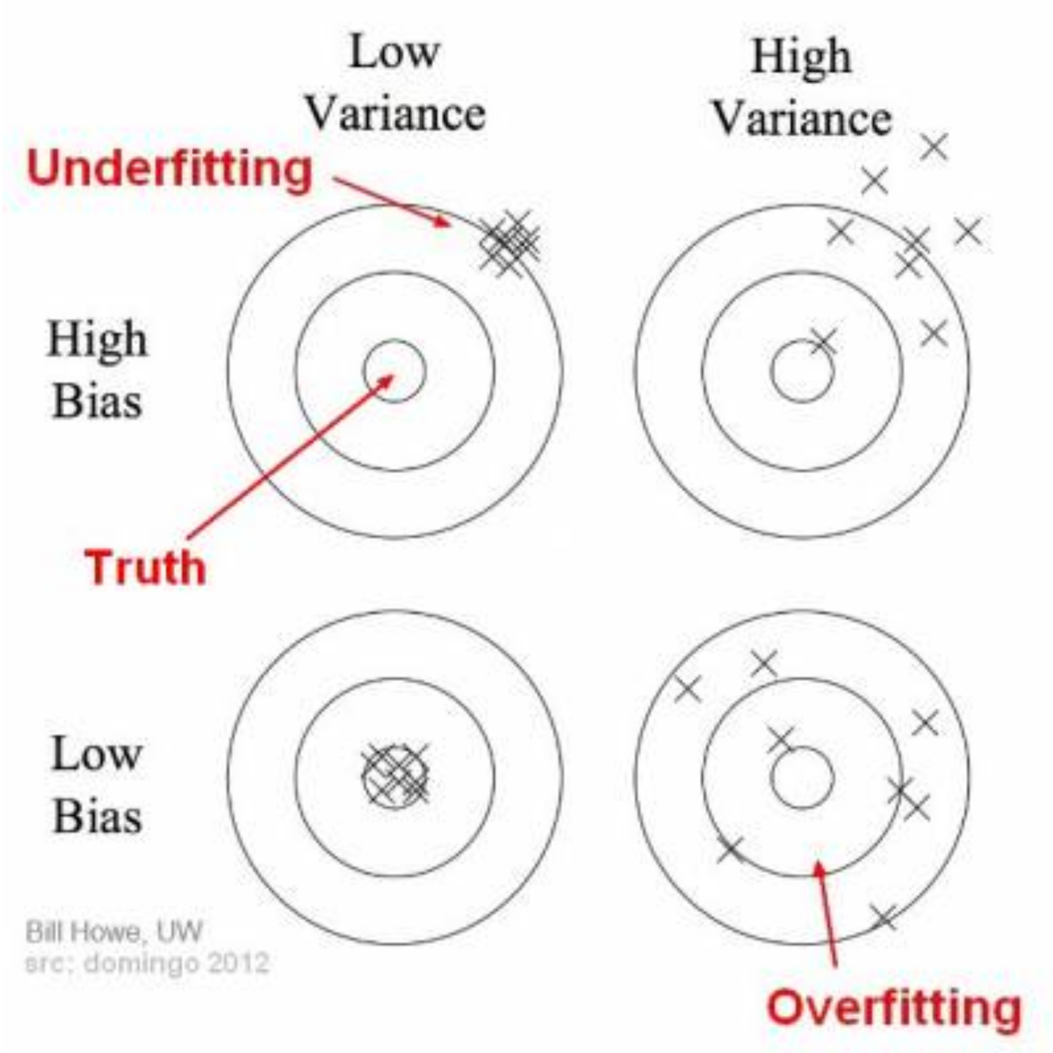
$$\mathbb{E}_x[\mathbb{E}_{\hat{f}}[(y-\hat{f}(x))^2]] = \mathbb{E}_x[\text{bias}[\hat{f}(x)]^2] + \mathbb{E}_x[\text{var}(\hat{f}(x))] + \sigma_\epsilon^2$$

  - First expectation is over the distribution of unseen (test) points x
  - Second expectation is over the distribution of the training data, or over $\hat{f}$, since $\hat{f}$ depends of the training data
- Total error = Bias^2 + Variance + Irreducible error
- See https://towardsdatascience.com/the-bias-variance-tradeoff-8818f41e39e9 for proof

Bias Variance Trade-off

- Model with high bias (underfitting)
  - Usually oversimplified, with too few parameters
  - Pays very little attention to the training data
  - Leads to high error on both training and test data

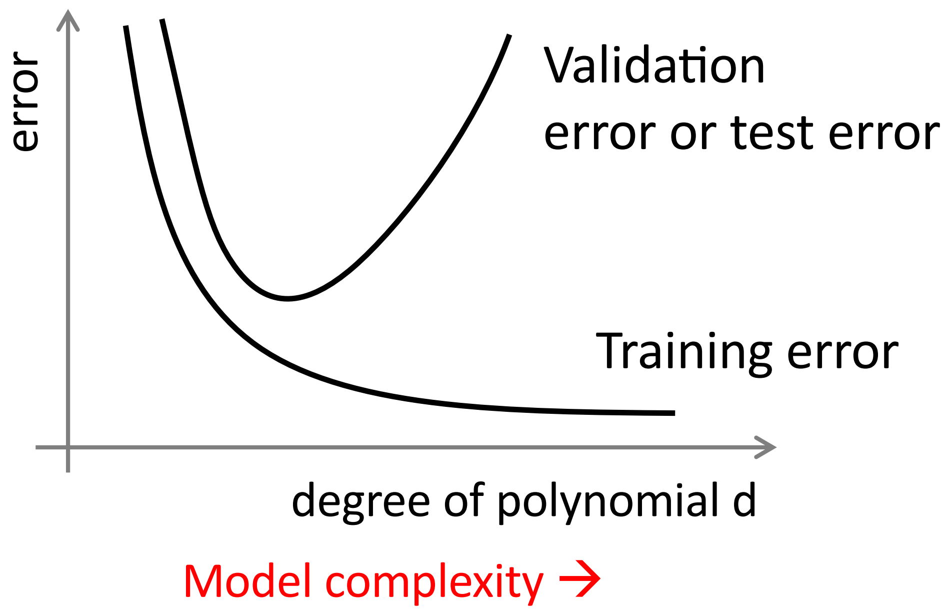
- Model with high variance (overfitting)
  - Usually too complex, with too many parameters
  - Pays a lot of attention to training data, but does not generalize well on unseen data
  - Can vary largely if different training data given



Crosses represent the function learned over different realizations of the training data

Practical implications

- Suppose your model is not performing as well as expected. Is it a bias problem or a variance problem?



High bias (underfit):
Both training error and validation / test error are high

High variance (overfit):
Low training error
High validation / test error

Will more training data help?

- A learnt model is not performing as well as expected. Will having more training data help?

- Note that there can be substantial cost for getting more training data.

Will more training data help?

- A learnt model is not performing as well as expected. Will having more training data help?

- Note that there can be substantial cost for getting more training data.

- If model is suffering from high bias, getting more training data will not (by itself) help much.
- If model is suffering from high variance, getting more training data is likely to help

A small note

- Overfitting of ML models may not be always bad
- Many modern Deep Learning models have millions of parameters, and it is actually desired that these models overfit the training data
  - Some technique like Regularization is used to ensure good generalization of the model
  - Details out of scope of this course

THANK YOU

Questions can be mailed to Dr. S. Ghosh (saptarshi@cse.iitkgp.ac.in)