Machine Learning

Advice for applying machine learning

Deciding what to try next

Debugging a learning algorithm:

Suppose you have implemented regularized linear regression to predict housing prices.

$$J(\theta) = \frac{1}{2m} \left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{m} \theta_j^2 \right]$$

However, when you test your hypothesis on a new set of houses, you find that it makes unacceptably large errors in its predictions. What should you try next?

- Get more training examples
- Try smaller sets of features
- Try getting additional features
- Try adding polynomial features $(x_1^2, x_2^2, x_1 x_2, \text{etc.})$
- Try decreasing λ
- Try increasing λ

$x_1, x_2, x_3, \ldots, x_{100}$

Machine learning diagnostic:

Diagnostic: A test that you can run to gain insight what is/isn't working with a learning algorithm, and gain guidance as to how best to improve its performance.

Diagnostics can take time to implement, but doing so can be a very good use of your time.

Advice for applying
machine learning

Evaluating a
hypothesis

Machine Learning

Evaluating your hypothesis

price

size

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$
$$+ \theta_3 x^3 + \theta_4 x^4$$

Fails to generalize to new examples not in training set.

$x_1 =$ size of house
$x_2 =$ no. of bedrooms
$x_3 =$ no. of floors
$x_4 =$ age of house
$x_5 =$ average income in neighborhood
$x_6 =$ kitchen size
$\vdots$
x_{100}

Andrew Ng

Evaluating your hypothesis

Dataset:

Size	Price
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

70%

30%

Training set

Test Set

$$(x^{(1)}, y^{(1)})$$
$$(x^{(2)}, y^{(2)})$$
$$\vdots$$
$$(x^{(m)}, y^{(m)})$$

$$(x_{test}^{(1)}, y_{test}^{(1)})$$
$$(x_{test}^{(2)}, y_{test}^{(2)})$$
$$\vdots$$
$$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$$

m_{test} = no. of test example

$(x_{test}^{(i)}, y_{test}^{(i)})$

Andrew Ng

Training/testing procedure for linear regression

→ - Learn parameter θ from training data (minimizing training error $J(\theta)$)

70 %

- Compute test set error:

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} \left(h_\theta(x_{test}^{(i)}) - y_{test}^{(i)} \right)^2$$

Training/testing procedure for logistic regression

- Learn parameter θ from training data
- Compute test set error:

$$J_{test}(\theta) = -\frac{1}{m_{test}} \sum_{i=1}^{m_{test}} y_{test}^{(i)} \log h_\theta(x_{test}^{(i)}) + (1 - y_{test}^{(i)}) \log h_\theta(x_{test}^{(i)})$$

- Misclassification error (0/1 misclassification error):

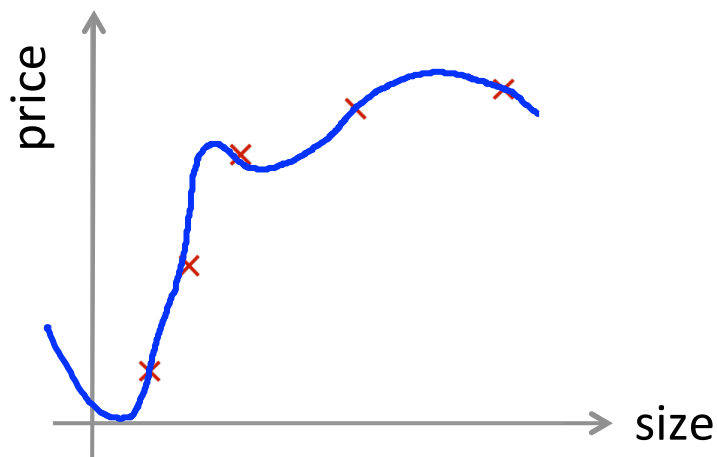Machine Learning

Advice for applying machine learning

Model selection and training/validation/test sets

Overfitting example

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$
$$+ \theta_3 x^3 + \theta_4 x^4$$

Once parameters $\theta_0, \theta_1, \ldots, \theta_4$ were fit to some set of data (training set), the error of the parameters as measured on that data (the training error $J(\theta)$) is likely to be lower than the actual generalization error.

Model selection

$d=1$ 1. $\rightarrow h_\theta(x) = \theta_0 + \theta_1 x \longrightarrow \Theta^{(1)} \longrightarrow J_{test}(\Theta^{(1)})$

$d=2$ 2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \longrightarrow \Theta^{(2)} \longrightarrow J_{test}(\Theta^{(2)})$

$d=3$ 3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3 \longrightarrow \Theta^{(3)} \rightarrow J_{test}(\Theta^{(3)})$

     $\vdots$ $\vdots$ $\vdots$

$d=10$ 10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10} \rightarrow \Theta^{(10)} \rightarrow J_{test}(\Theta^{(10)})$

$\rightarrow d = $ degree of polynomial

Choose $\boxed{\theta_0 + \ldots \theta_5 x^5} \leftarrow$

How well does the model generalize? Report test set error $\underline{J_{test}(\theta^{(5)})}$. $\boxed{\Theta_0, \Theta_1 \ldots}$

$\Theta^{(5)}$

Problem: $J_{test}(\theta^{(5)})$ is likely to be an optimistic estimate of generalization error. I.e. our extra parameter ($\underline{d}$ = degree of polynomial) is fit to test set.

Evaluating your hypothesis

Dataset:

Size	Price	
2104	400	
1600	330	
2400	369	60% Training set
1416	232	
3000	540	
1985	300	
1534	315	20% Cross validation set (CV)
1427	199	
1380	212	20% test set
1494	243	

$$(x^{(1)}, y^{(1)})$$
$$(x^{(2)}, y^{(2)})$$
$$\vdots$$
$$(x^{(m)}, y^{(m)})$$

$$(x_{cv}^{(1)}, y_{cv}^{(1)})$$
$$(x_{cv}^{(2)}, y_{cv}^{(2)})$$
$$\vdots$$
$$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$$

m_{cv} = no. of cv examples $(x_{cv}^{(i)}, y_{cv}^{(i)})$

$$(x_{test}^{(1)}, y_{test}^{(1)})$$
$$(x_{test}^{(2)}, y_{test}^{(2)})$$
$$\vdots$$
$$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$$

m_{test}

Andrew Ng

Train/validation/test error

Training error:

$$\Rightarrow \quad J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \qquad J(\theta)$$

Cross Validation error:

$$\Rightarrow \quad J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

Test error:

$$\Rightarrow \quad J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_\theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

Model selection

$d=1$ 1. $h_\theta(x) = \theta_0 + \theta_1 x$ $\longrightarrow$ $\min_\theta J(\theta) \rightarrow \theta^{(1)} \longrightarrow J_{cv}(\theta^{(1)})$

$d=2$ 2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$ $\longrightarrow$ $\theta^{(2)} \longrightarrow J_{cv}(\theta^{(2)})$

$d=3$ 3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$ $\longrightarrow$ $\theta^{(3)}$

       $\vdots$ $J_{cv}(\theta^{(4)})$

$d=10$ 10. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$ $\longrightarrow$ $\theta^{(10)} \longrightarrow J_{cv}(\theta^{(10)})$

$$d = 4$$

Pick $\theta_0 + \theta_1 x_1 + \cdots + \theta_4 x^4$ $\longleftarrow$

Estimate generalization error   for test set $\underline{J_{test}(\theta^{(4)})}$ $\longleftarrow$
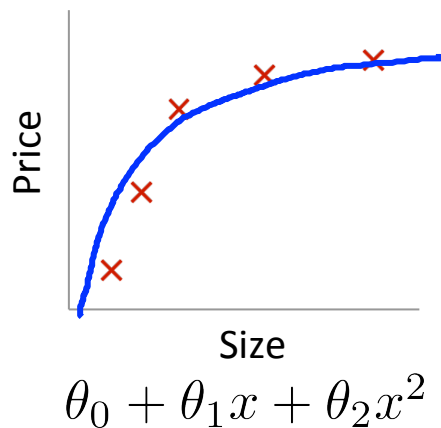
Machine Learning

Advice for applying machine learning

Diagnosing bias vs. variance

Bias/variance

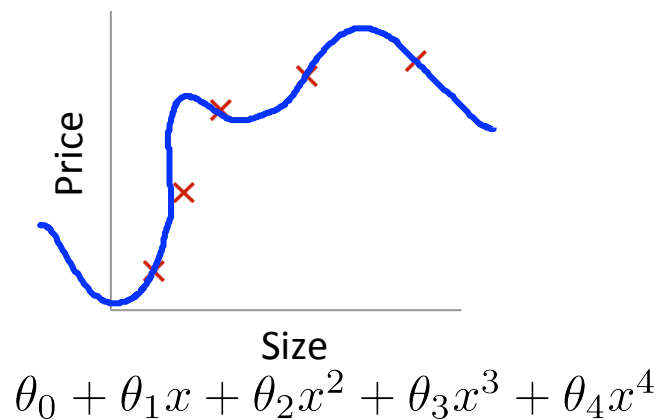$$\theta_0 + \theta_1 x$$

High bias
(underfit)

$d=1$

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

"Just right"

$d=2$

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

High variance
(overfit)

$d=4$

Andrew Ng

Bias/variance

Training error: $J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$

Cross validation error: $J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$ $\quad \left(\text{or } J_{test}(\theta)\right)$

$\leftarrow J_{cv}(\theta) \quad \left(\text{or } J_{test}(\theta)\right)$

$J_{train}(\theta)$

degree of polynomial d

$d=1$

$d=2$

Diagnosing bias vs. variance

Suppose your learning algorithm is performing less well than you were hoping. ($J_{cv}(\theta)$ or $J_{test}(\theta)$ is high.) Is it a bias problem or a variance problem?

Bias (underfit):

$\rightarrow J_{train}(\theta)$ will be high

$\left. J_{cv}(\theta) \approx J_{train}(\theta) \right\}$

Variance (overfit):

$\rightarrow J_{train}(\theta)$ will be low

$\left. J_{cv}(\theta) \gg J_{train}(\theta) \right\}$

$\gg$

Machine Learning

Advice for applying
machine learning

Regularization and
bias/variance

Linear regression with regularization

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ ←

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^{m} \theta_j^2$$ ←

Large λ ←
High bias (underfit)
$\lambda = 10000.\ \theta_1 \approx 0, \theta_2 \approx 0, \ldots$
$h_\theta(x) \approx \theta_0$

Intermediate λ ←
"Just right"

Small λ
High variance (overfit)
$\lambda = 0$

Andrew Ng

Choosing the regularization parameter λ

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \quad \leftarrow$$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{i=1}^{m} \theta_j^2 \quad \leftarrow$$

$$\rightarrow J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_\theta(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

$J(\theta)$

J_{train}
J_{cv}
J_{test}

Choosing the regularization parameter λ

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^{m} \theta_j^2$$
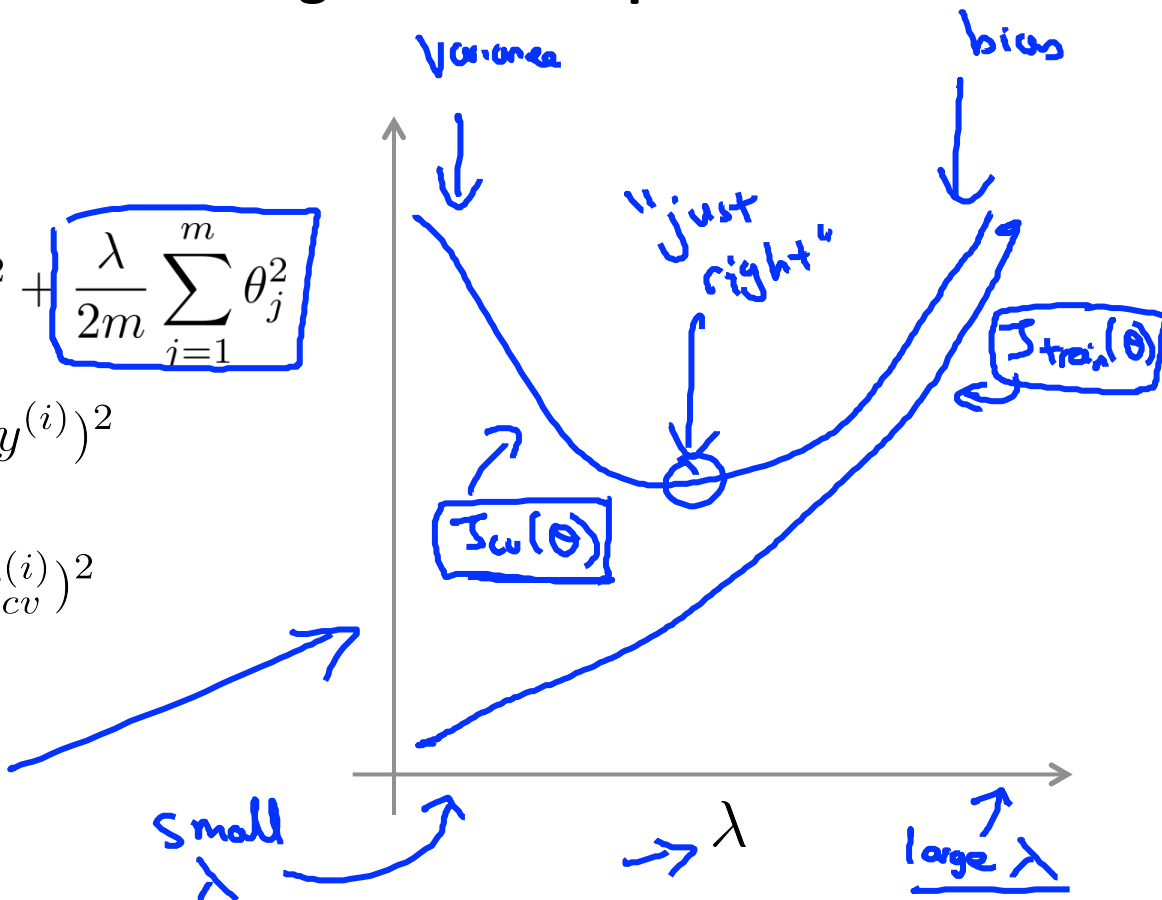
1. Try $\lambda = 0$ $\longrightarrow$ $\min_\theta J(\theta) \rightarrow \Theta^{(1)} \rightarrow J_{cv}(\Theta^{(1)})$

2. Try $\lambda = 0.01$ $\longrightarrow$ $\min_\theta J(\theta) \rightarrow \Theta^{(1)} \rightarrow J_{cv}(\Theta^{(2)})$

3. Try $\lambda = 0.02$ $\longrightarrow \Theta^{(3)} \rightarrow J_{cv}(\Theta^{(3)})$

4. Try $\lambda = 0.04$

5. Try $\lambda = 0.08$ $\longrightarrow \Theta^{(5)}$ $J_{cv}(\Theta^{(5)})$

$\vdots$

12. Try $\lambda = 10$ $\longrightarrow \Theta^{(12)} \rightarrow J_{cv}(\Theta^{(12)})$

     $\uparrow$ $\overline{10.24}$

Pick (say) $\theta^{(5)}$. Test error: $J_{test}(\Theta^{(5)})$

Bias/variance as a function of the regularization parameter λ

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \boxed{\frac{\lambda}{2m} \sum_{i=1}^{m} \theta_j^2}$$

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$\boxed{J_{cv}(\theta)} = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

Variance

bias

"just right"

$J_{train}(\theta)$

$J_{cv}(\theta)$

Small λ

λ

large λ

Advice for applying
machine learning

Learning curves

Machine Learning

Learning curves

$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_\theta(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$

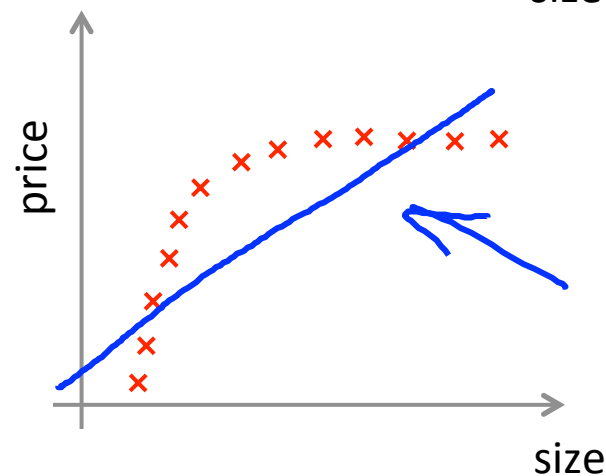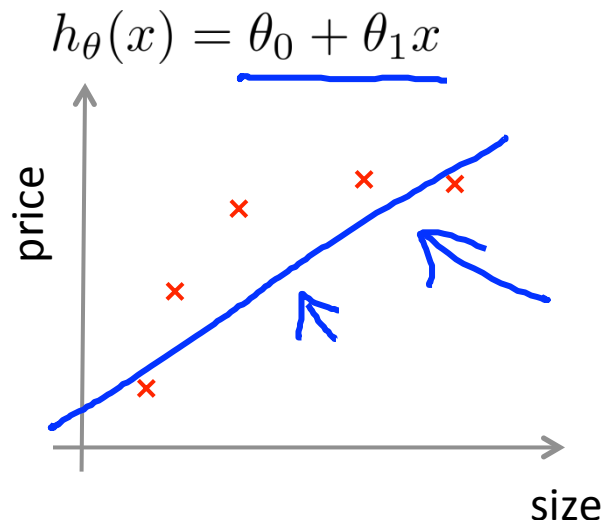$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

error

$J_{cv}(\theta)$

$J_{train}(\theta)$

m (training set size)

m=1

m=2

m=3

m=4

High bias

$$h_\theta(x) = \theta_0 + \theta_1 x$$

error

high error

$J_{cv}(\theta)$

$J_{train}(\theta)$

m (training set size)

price

size

If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.

price

size

High variance

m (training set size)
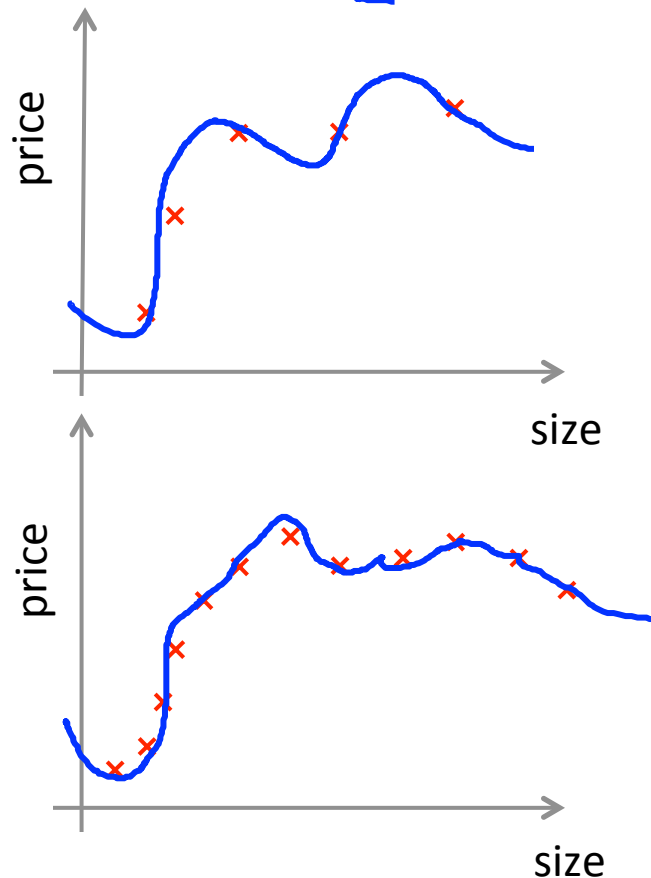
$J_{cv}(\theta)$ (or $J_{test}(\theta)$)

gap

$J_{train}(\theta)$

error

If a learning algorithm is suffering from high variance, getting more training data is likely to help. ←

$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$$
(and small λ)

price

size

price

size

Machine Learning

Advice for applying machine learning

Deciding what to try next (revisited)

Debugging a learning algorithm:

Suppose you have implemented regularized linear regression to predict housing prices. However, when you test your hypothesis in a new set of houses, you find that it makes unacceptably large errors in its prediction. What should you try next?

- Get more training examples $\rightarrow$ fixes high variance
- Try smaller sets of features $\rightarrow$ fixes high variance
- Try getting additional features $\rightarrow$ fixes high bias
- Try adding polynomial features$(x_1^2, x_2^2, x_1 x_2, \text{etc})$ $\rightarrow$ fixes high bias.
- Try decreasingλ $\rightarrow$ fixes high bias
- Try increasingλ $\rightarrow$ fixes high variance

Neural networks and overfitting

"Small" neural network
(fewer parameters; more
prone to underfitting)

"Large" neural network
(more parameters; more prone
to overfitting)
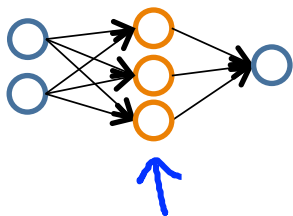
Computationally cheaper

Computationally more expensive.

Use regularization (λ) to address overfitting.

$J_{c_0}(\vec{\omega})$