



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_1x_2 , etc.)
- Try decreasing λ
- Try increasing λ

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.

Which of the following statements about diagnostics are true? Check all that apply.

- ☐ It's hard to tell what will work to improve a learning algorithm, so the best approach is to go with gut feeling and just see what works.

Well done!

- ☒ Diagnostics can give guidance as to what might be more fruitful things to try to improve a learning algorithm.

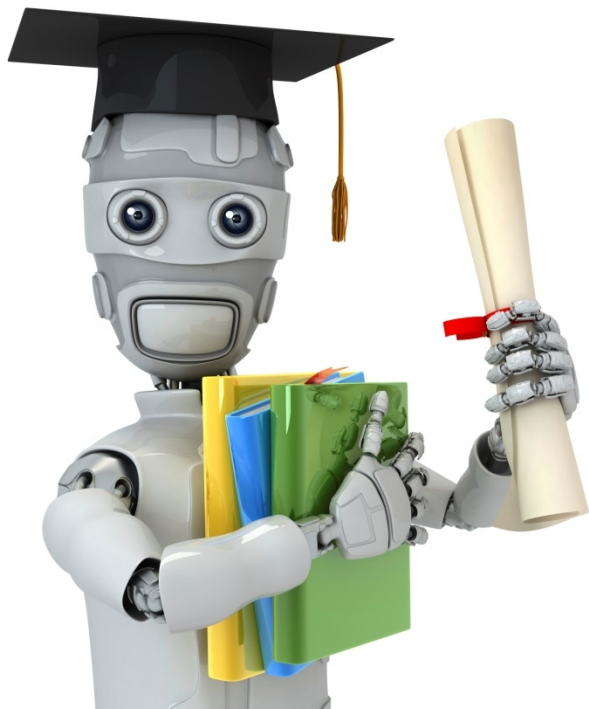
Well done!

- ☒ Diagnostics can be time-consuming to implement and try, but they can still be a very good use of your time.

Well done!

- ☒ A diagnostic can sometimes rule out certain courses of action (changes to your learning algorithm) as being unlikely to improve its performance significantly.

Well done!



Machine Learning

Advice for
applying
machine
learning

Evaluating a
hypothesis

Evaluating your hypothesis

Just because hypothesis has low training error doesn't mean it's a good hypothesis (hypothesis might overfit data which leads to failing to generalize new examples in training set). Hence training set error is not a good predictor for how well hypothesis will do for new examples. So, How can we tell hypothesis might be overfitting ? One simple solution is to plot and see, But it is not feasible for large no. of features.

Fails to generalize to new examples not in training set.

x_1 = size of

x_2 = no. of bedrooms

x_3 = no. of floors

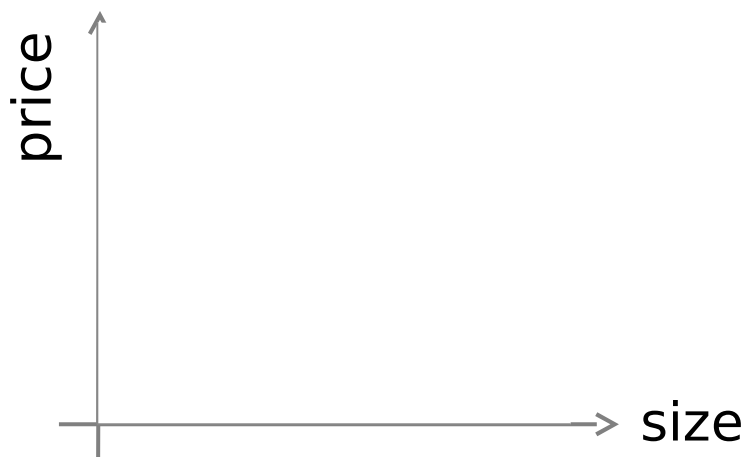
x_4 = age of

x_5 = average income in house

x_6 = kitchen size

\vdots

x_{100}



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

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 |



$$(x^{(1)}, y^{(1)})$$

$$(x^{(2)}, y^{(2)})$$

\vdots

$$(x^{(m)}, y^{(m)})$$



$$(x_{test}^{(1)}, y_{test}^{(1)})$$

$$(x_{test}^{(2)}, y_{test}^{(2)})$$

\vdots

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

Training/testing procedure for linear regression

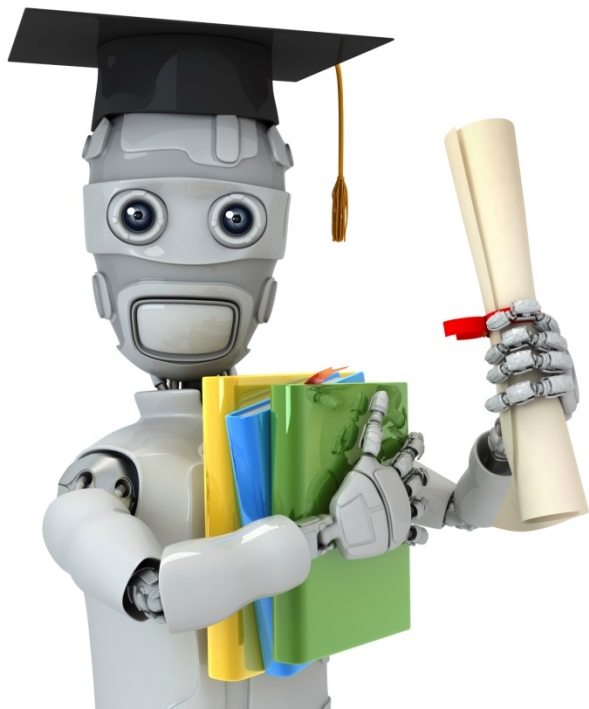
- Learn parameter θ from training data
(minimizing $J(\theta)$ training error)
- Compute test set error:

Training/testing procedure for logistic regression

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

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

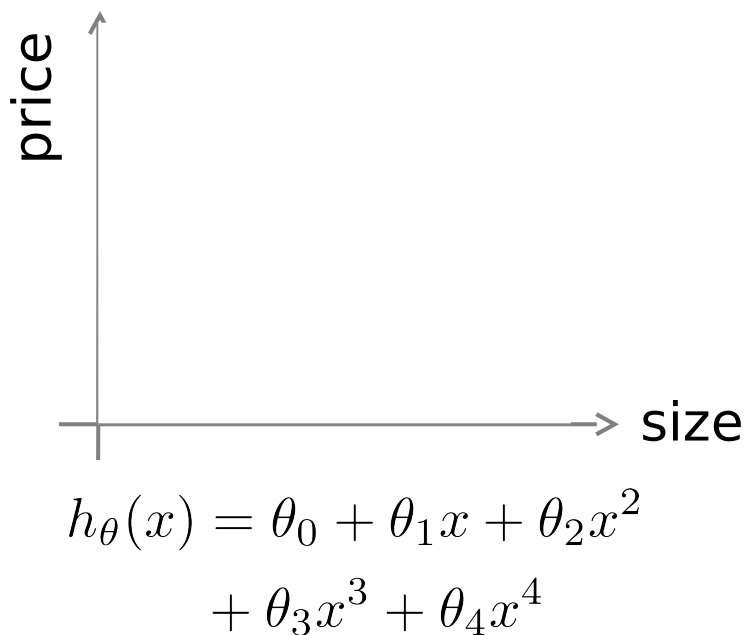
- Misclassification error (0/1 misclassification error):



Machine Learning

Advice for
applying
machine
learning
Model selection and
training/validation/t
est sets

Overfitting example



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

Model selection

1. $h_{\theta}(x) = \theta_0 + \theta_1 x$
2. $h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_3 x^3$
- \vdots
10. $h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{10} x^{10}$

Choose $\theta_0 + \dots + \theta_5 x^5$

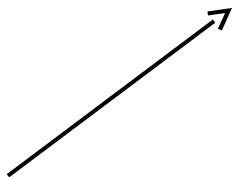
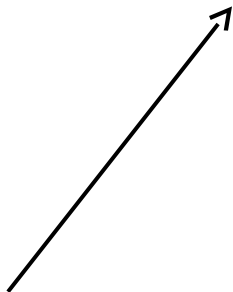
How well does the model generalize? Report test set error $J_{test}(\theta^{(5)})$.

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

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 |



$$(x^{(1)}, y^{(1)})$$

$$(x^{(2)}, y^{(2)})$$

$$\vdots$$

$$(x^{(m)}, y^{(m)})$$

$$(x_{cv}^{(1)}, y_{cv}^{(1)})$$

$$(x_{cv}^{(2)}, y_{cv}^{(2)})$$

$$\vdots$$

$$(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$$

$$(x_{test}^{(1)}, y_{test}^{(1)})$$

$$(x_{test}^{(2)}, y_{test}^{(2)})$$

$$\vdots$$

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

Train/validation/test error

Training error:

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

Cross Validation error:

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

Test error:

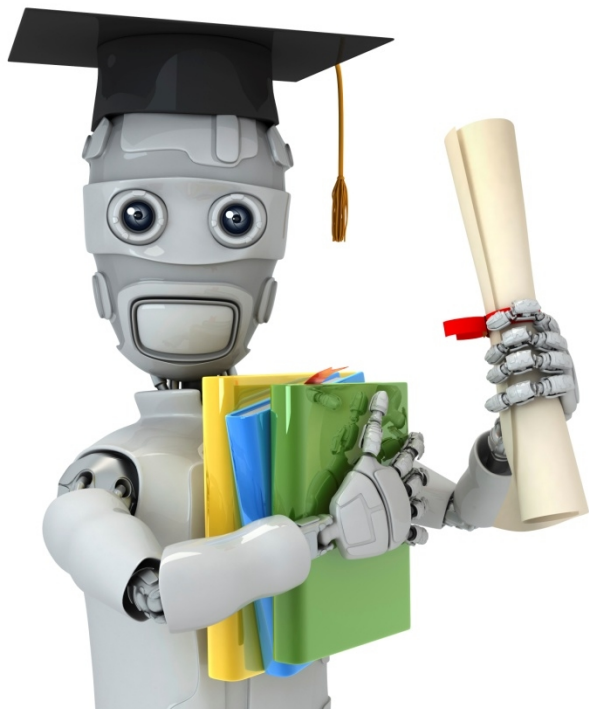
$$J_{test}(\theta) = \frac{1}{2m_{test}} \sum_{i=1}^{m_{test}} (h_{\theta}(x_{test}^{(i)}) - y_{test}^{(i)})^2$$

Model selection

1. $h_{\theta}(x) = \theta_0 + \theta_1 x$
2. $h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
3. $h_{\theta}(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$
- \vdots
10. $h_{\theta}(x) = \theta_0 + \theta_1 x + \cdots + \theta_{10} x^{10}$

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

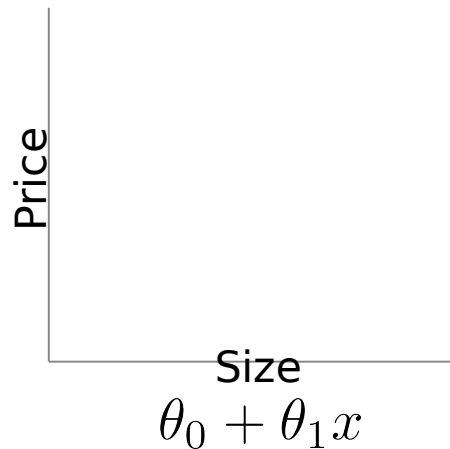
Estimate generalization error for $J_{\text{test}}(\theta^{(4)})$



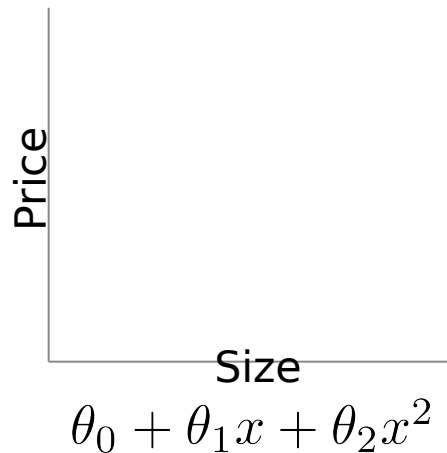
Machine Learning

Advice for applying machine learning Diagnosing bias vs. variance

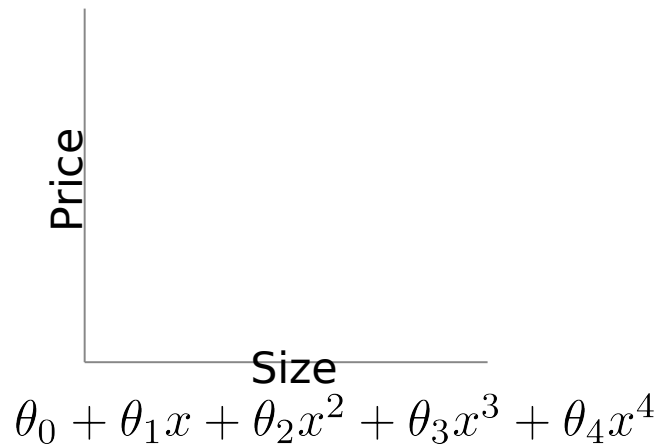
Bias/variance



High bias
(underfit)



“Just right”

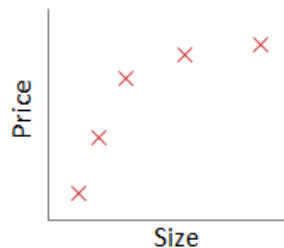
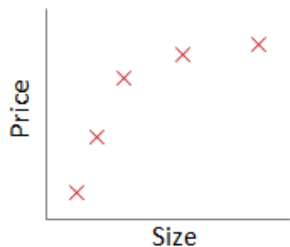
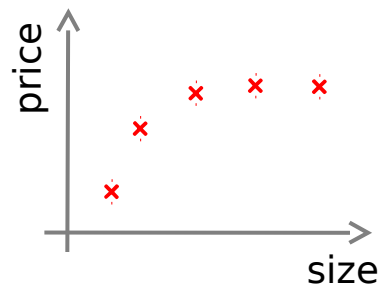


High
variance
(overfit)

Bias/variance

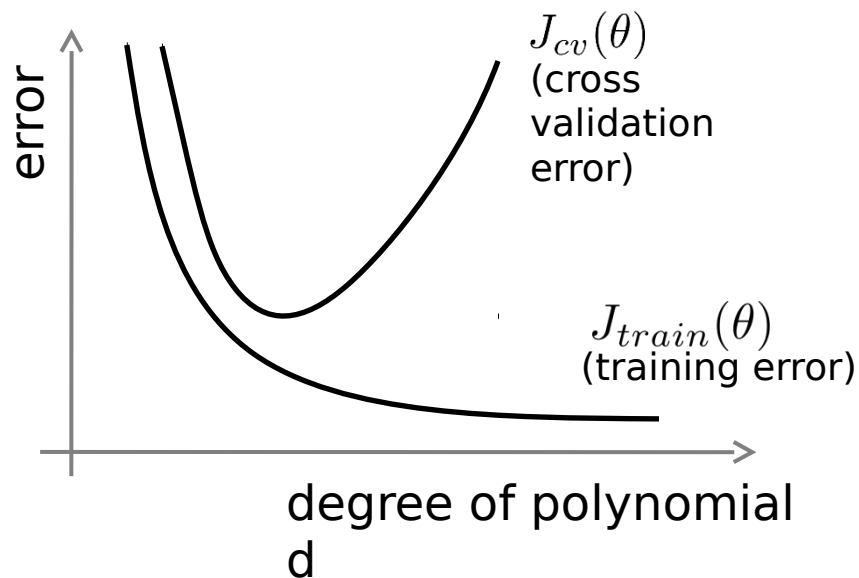
Training error:
$$J_{train}(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

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



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):

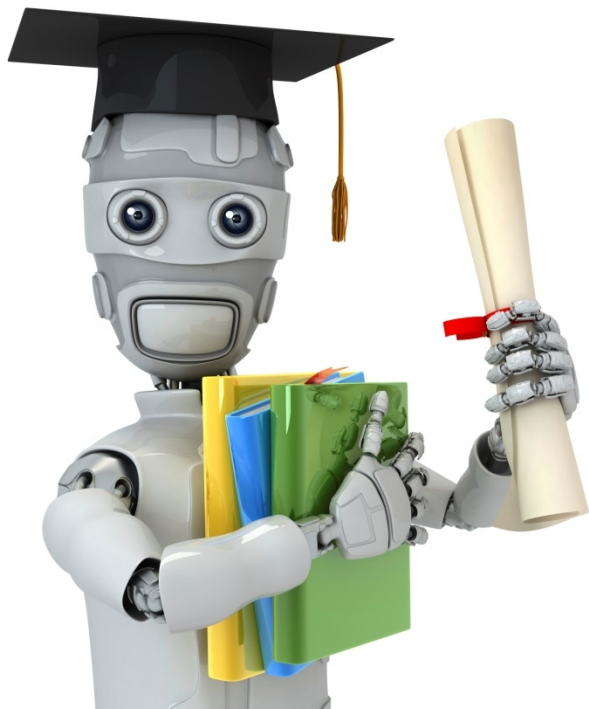
Variance (overfit):

Suppose you have a classification problem. The (misclassification) error is defined as $\frac{1}{m} \sum_{i=1}^m \text{err}(h_{\theta}(x^{(i)}), y^{(i)})$, and the cross validation (misclassification) error is similarly defined, using the cross validation examples $(x_{cv}^{(1)}, y_{cv}^{(1)}), \dots, (x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})$. Suppose your training error is 0.10, and your cross validation error is 0.30. What problem is the algorithm most likely to be suffering from?

- ☐ High bias (overfitting)
- ☐ High bias (underfitting)
- ☒ High variance (overfitting)

Well done!

- ☐ High variance (underfitting)



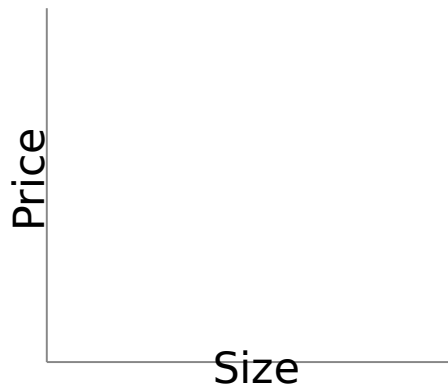
Machine Learning

Advice for
applying
machine
Regularization
and
bias/variance

Linear regression with regularization

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

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

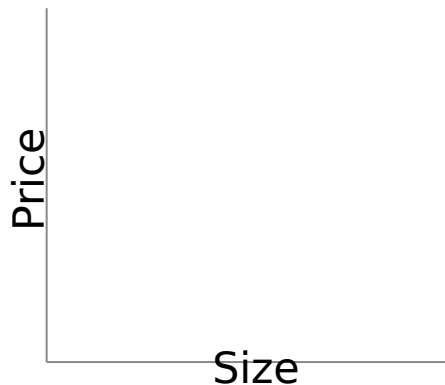


Large λ
High bias

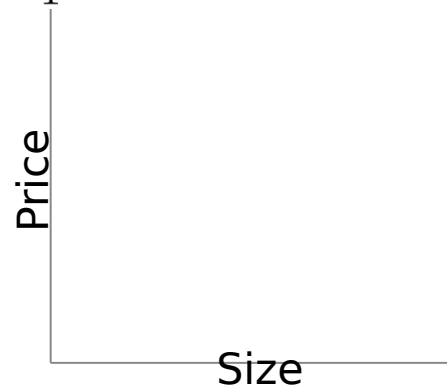
$\lambda = 10000, \theta_1 \approx 0, \dots$

$h_{\theta}(x) \approx \theta_0$

(underfit)



Intermediate λ
"Just right"



Small λ
High variance (overfit)

Choosing the regularization parameter

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

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

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

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

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

Choosing the regularization parameter

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 $\lambda = 0$
2. Try $\lambda = 0.01$
3. Try $\lambda = 0.02$
4. Try $\lambda = 0.04$
5. Try $\lambda = 0.08$
- \vdots
12. Try $\lambda = 10$

Pick (say) $\lambda^{(5)}$. Test error:

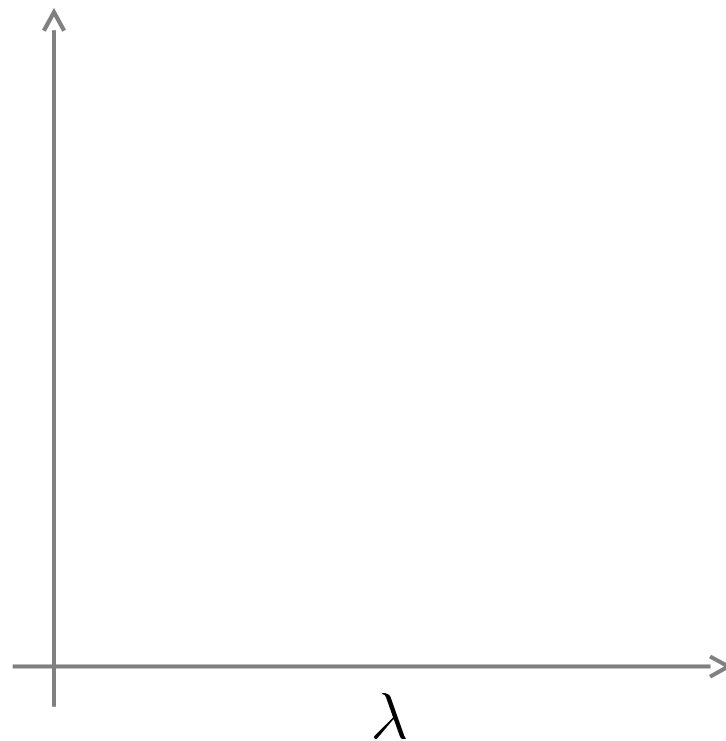
Bias/variance as a function of the regularization parameter

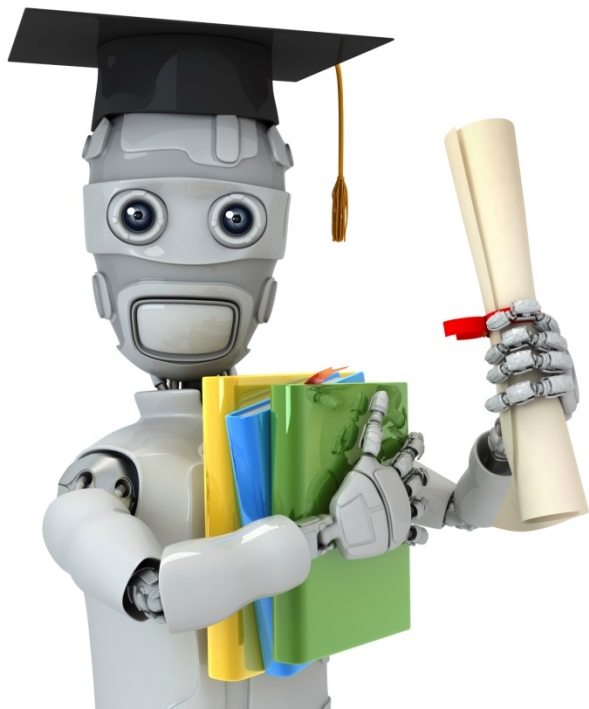
λ

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

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

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





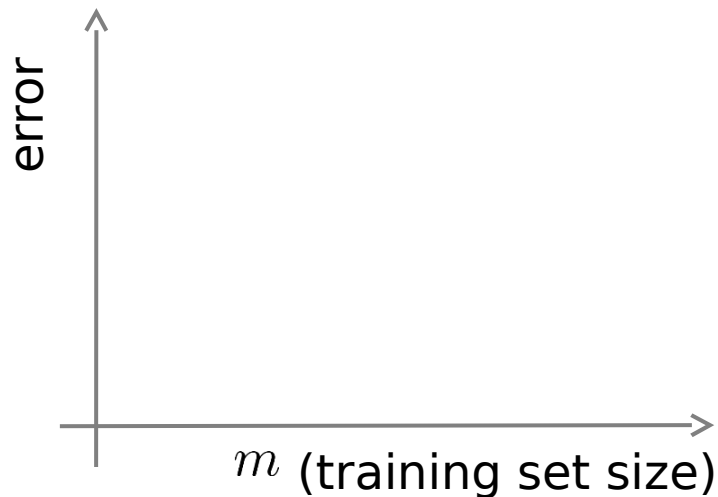
Machine Learning

Advice for applying machine --- learning Learning curves

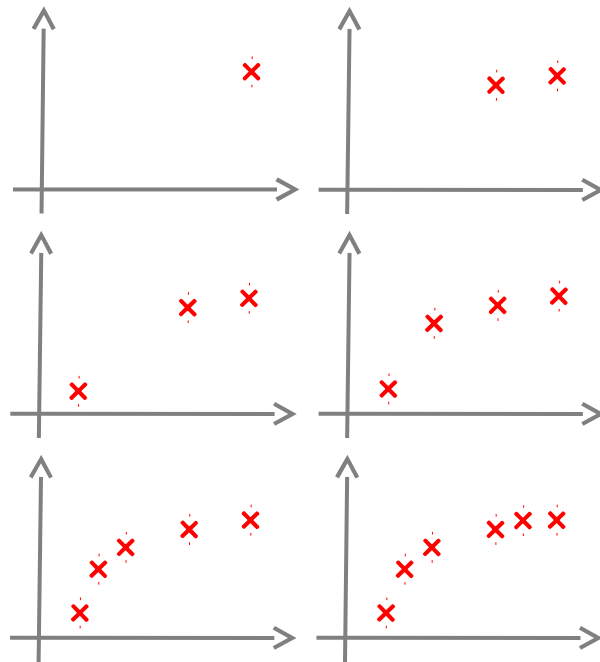
Learning curves

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

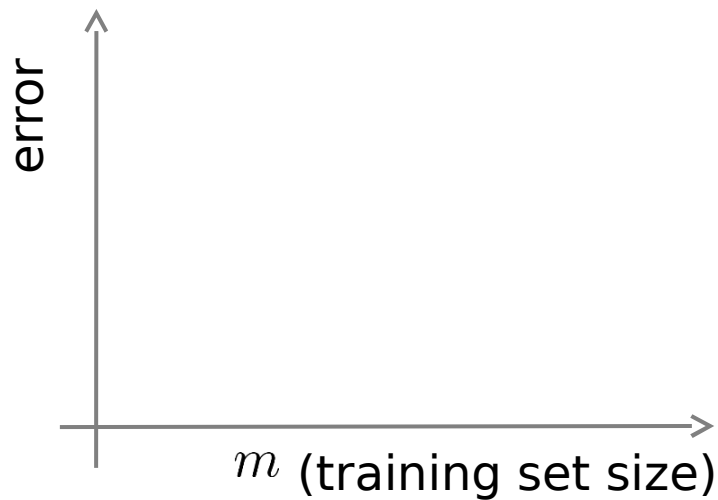
$$J_{cv}(\theta) = \frac{1}{2m_{cv}} \sum_{i=1}^{m_{cv}} (h_{\theta}(x_{cv}^{(i)}) - y_{cv}^{(i)})^2$$



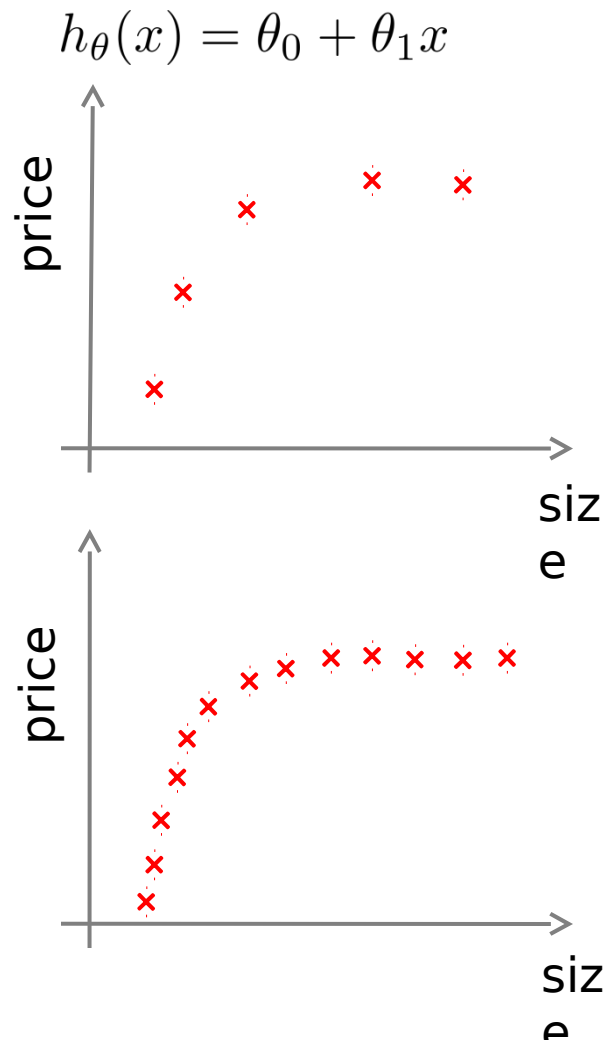
$$h_{\theta}(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$



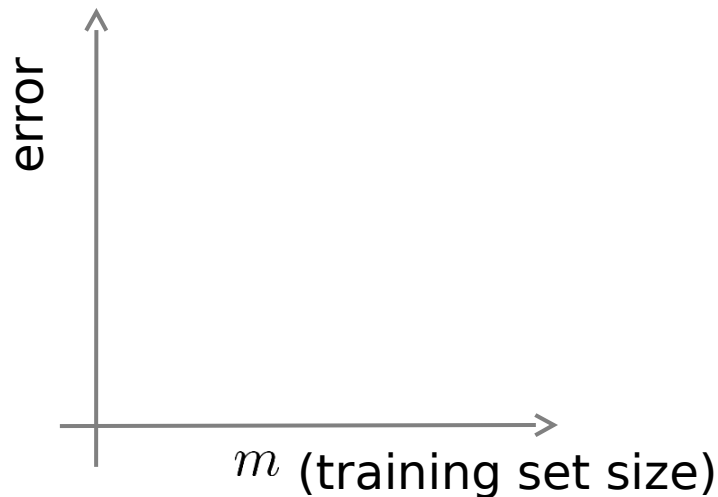
High bias



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



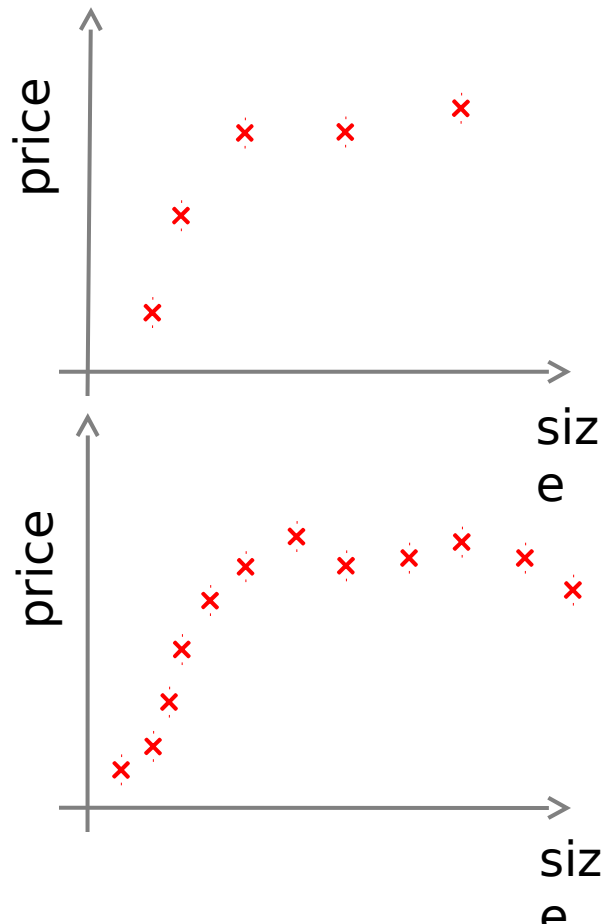
High variance



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

$$h_{\theta}(x) = \theta_0 + \theta_1 x + \dots + \theta_{100} x^{100}$$

(and small λ)



In which of the following circumstances is getting more training data likely to significantly help a learning algorithm's performance?

☐ Algorithm is suffering from high bias.

Well done!

☒ Algorithm is suffering from high variance.

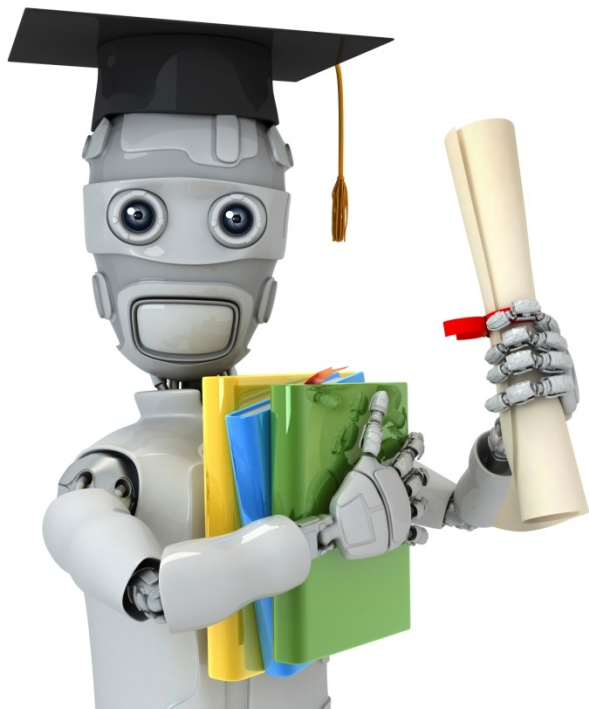
Well done!

☒ $J_{CV}(\theta)$ (cross validation error) is much larger than $J_{train}(\theta)$ (training error).

Well done!

☐ $J_{CV}(\theta)$ (cross validation error) is about the same as $J_{train}(\theta)$ (training error).

Well done!



Machine Learning

Advice for
applying
machine
~~learning~~ deciding what
to try next
(revisited)

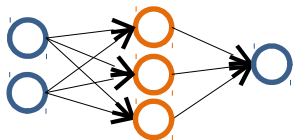
Debugging a learning algorithm:

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
- Try smaller sets of features
- Try getting additional features
- Try adding polynomial features ($w_1^2, w_2^2, x_1 x_2, \text{etc}$)
- Try decreasing λ
- Try increasing λ

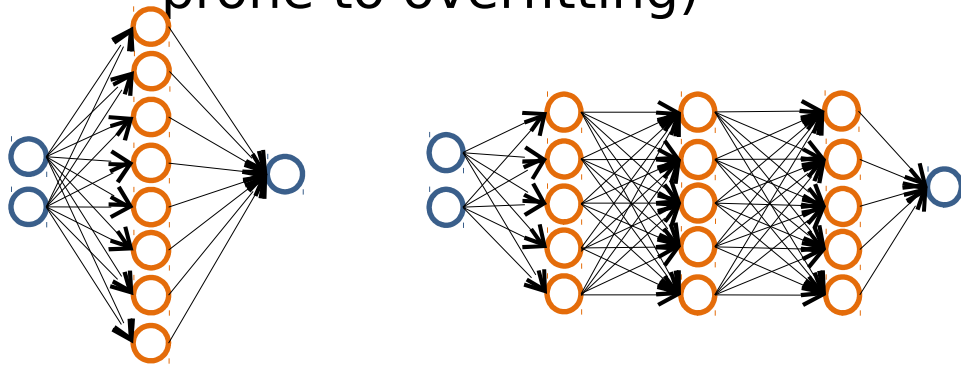
Neural networks and overfitting

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



Computationally
cheaper

“Large” neural network
(more parameters; more
prone to overfitting)



Computationally more expensive.

Use regularization () to address
overfitting.

Suppose you fit a neural network with one hidden layer to a training set. You find that the cross validation error $J_{CV}(\theta)$ is much larger than the training error $J_{train}(\theta)$. Is increasing the number of hidden units likely to help?

- ☐ Yes, because this increases the number of parameters and lets the network represent more complex functions.
- ☐ Yes, because it is currently suffering from high bias.
- ☐ No, because it is currently suffering from high bias, so adding hidden units is unlikely to help.
- ☒ No, because it is currently suffering from high variance, so adding hidden units is unlikely to help.

Correct Response