Diagnostics

Evaluating a Learning Algorithm

Advice for Applying Machine Learning

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?

Actions

- Some actions
  - Get more training examples
  - Try smaller set of features (a small set of features)
  - Try getting additional features (just the opposite)
  - Try adding polynomial features
  - Try increasing lambda
  - Try decreasing lambda
- People generally randomly choose one and try it, which is waste of time most of the time.

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

Exercise

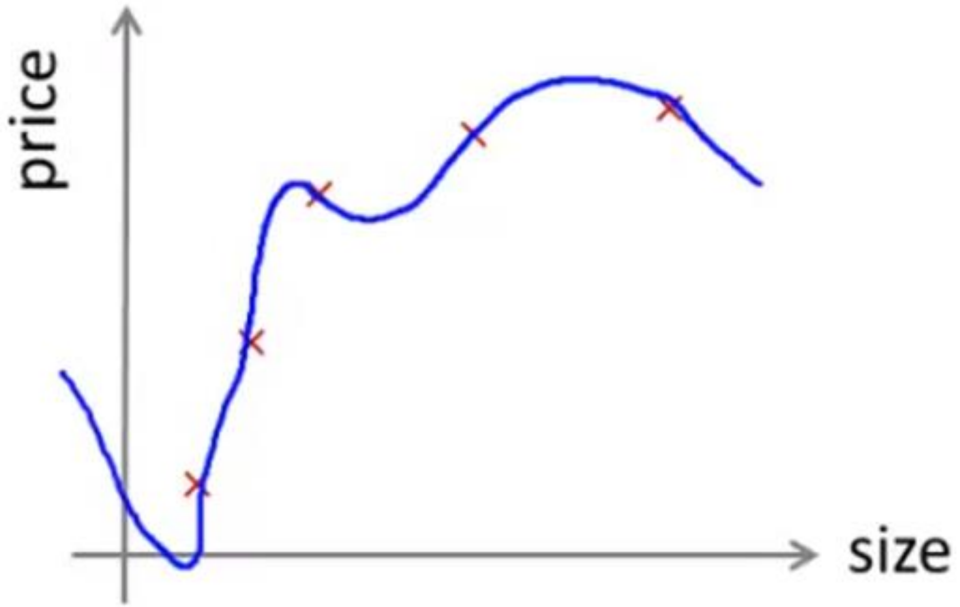
- Which of the following statements about diagnostics are true? Check all that apply.
  - It’s hard to tell what will work to improve a learning algorithm, so the best approach is to go with gut feeling and just see what works.
  - Diagnostics can give guidance as to what might be more fruitful things to try to improve a learning algorithm.
  - Diagnostics can be time-consuming to implement and try, but they can still be a very good use of your time.
  - A diagnostic can sometimes rule out certain courses of action (changes to your learning algorithm) as being unlikely to improve its performance significantly.

Evaluating a Hypothesis
Train Set, Test Set

Evaluating a Learning Algorithm

Advice for Applying Machine Learning

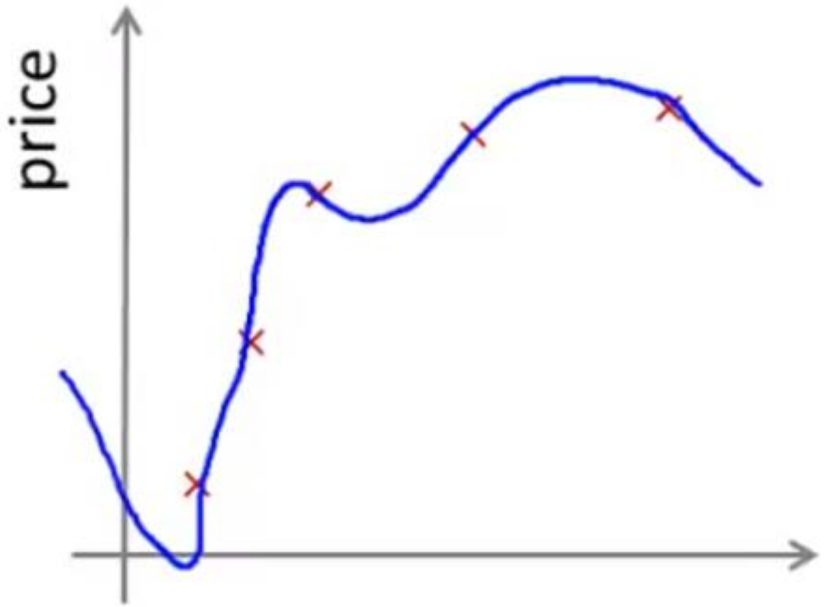
Evaluating your hypoth



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

How to plot many features?

Evaluating your hypothesis



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

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}

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% Training set

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

30% Test set

$$(x^{(1)}_{test}, y^{(1)}_{test})$$
$$(x^{(2)}_{test}, y^{(2)}_{test})$$
$$\vdots$$
$$(x^{(m_{test})}_{test}, y^{(m_{test})}_{test})$$

Exercise

- Suppose an implementation of linear regression (without regularization) is badly overfitting the training set.
- In this case, we would expect:
  - The training error $J(\theta)$ to be **low** and the test error $J_{test}(\theta)$ to be **high**
  - The training error $J(\theta)$ to be **low** and the test error $J_{test}(\theta)$ to be **low**
  - The training error $J(\theta)$ to be **high** and the test error $J_{test}(\theta)$ to be **low**
  - The training error $J(\theta)$ to be **high** and the test error $J_{test}(\theta)$ to be **high**

Procedure

- Training/testing procedure for linear regression
  - Learn parameter θ from training data by minimizing the training error J(θ)
  - Compute the test error for linear regression:

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

  - Computer the test error for logistic regression

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

Training/testing procedure for logistic regression

- Learn parameter θ from training data

m_{test}

- 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):

$$err(h_\theta(x), y) = \begin{cases} 1 & \text{if } h_\theta(x) \geq 0.5,\ y = 0 \\ & \text{or if } h_\theta(x) < 0.5,\ y = 1 \\ 0 & \text{otherwise} \end{cases}$$

error

$$\text{Test error} = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_\theta(x_{test}^{(i)}), y^{(i)}).$$

Model Selection and Train/Validation/Test Sets

Evaluating a Learning Algorithm

Advice for Applying Machine Learning

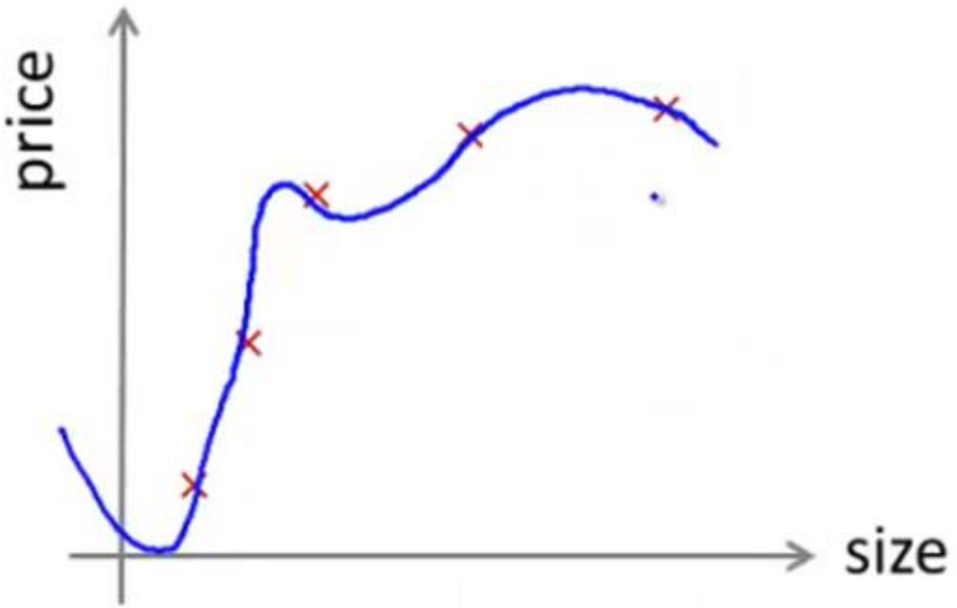
Introduction

- Suppose you are left to decide what degree of polynomial to fit to a data set.
- So that what features to include that gives you a learning algorithm.
- Or suppose you'd like to choose the regularization parameter lambda for learning algorithm
- These are called **model selection** problems.

Introduction

- We've already seen a lot of times the problem of overfitting, in which just because a learning algorithm fits a training set well, that doesn't mean it's a good hypothesis.
- More generally, this is why the training set's error is not a good predictor for how well the hypothesis will do on new example.

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.

Therefore it's an optimistic estimate for the real life 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. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3$

$\vdots$

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

Model selection

d = degree of polynomial

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

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

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

$\vdots$ $\quad\vdots$ $\quad\vdots$

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

Choose $\theta_0 + \ldots \theta_5 x^5 \leftarrow$

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

Model selection

d = degree of polynomial

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

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

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

$\vdots$ $\quad\vdots$ $\quad\vdots$

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

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

$\boxed{\Theta_0, \Theta_1 \ldots}$

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

$\Theta^{(5)}$

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

Evaluating your hypothesis

Dataset:

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

$$\boxed{\begin{array}{c}(x^{(1)}, y^{(1)})\\(x^{(2)}, y^{(2)})\\\vdots\\(x^{(m)}, y^{(m)})\end{array}}$$

$$\boxed{\begin{array}{c}(x_{cv}^{(1)}, y_{cv}^{(1)})\\(x_{cv}^{(2)}, y_{cv}^{(2)})\\\vdots\\(x_{cv}^{(m_{cv})}, y_{cv}^{(m_{cv})})\end{array}}$$

m_{cv} = no. of cv example

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

$$\boxed{\begin{array}{c}(x_{test}^{(1)}, y_{test}^{(1)})\\(x_{test}^{(2)}, y_{test}^{(2)})\\\vdots\\(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})\end{array}}$$

m_{test}

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

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

Model selection

1. $h_\theta(x) = \theta_0 + \theta_1 x \longrightarrow \min_\theta J(\theta) \rightarrow \Theta^{(1)}$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \longrightarrow \Theta^{(2)}$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3 \longrightarrow \Theta^{(3)}$

$\vdots$

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

Find theta's using the test set, i.e., find theta that minimizes the error of the test set.

Model selection

1. $h_\theta(x) = \theta_0 + \theta_1 x \longrightarrow \min_\theta J(\theta) \rightarrow \theta^{(1)} \longrightarrow J_{cv}(\theta^{(1)})$
2. $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \longrightarrow \theta^{(2)} \longrightarrow J_{cv}(\theta^{(2)})$
3. $h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_3 x^3 \longrightarrow \theta^{(3)}$

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

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

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

$\vdots \quad \vdots$ $\qquad J_{cv}(\Theta^{(4)})$

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

$\underline{d=4}$

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

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

Exercise

- Consider the model selection procedure where we choose the degree of polynomial using a cross validation set. For the final model (with parameters θ), we might generally expect $\mathrm{J}_{\mathrm{CV}}(\theta)$ to be lower than $\mathrm{J}_{\mathrm{test}}(\theta)$
  - An extra parameter (d, the degree of the polynomial) has been fit to the cross validation set.
  - An extra parameter (d, the degree of the polynomial) has been fit to the test set.
  - The cross validation set is usually smaller than the test set.
  - The cross validation set is usually larger than the test set.