Evaluation the Learning Algorithm

Prof. Dr. Christina Bauer

christina.bauer@th-deg.de

Faculty of Computer Science

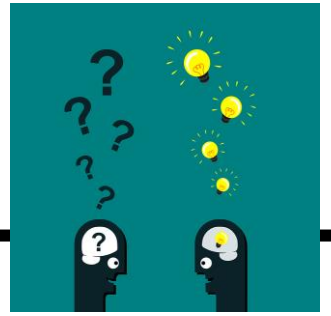
DEBUGGING

- Example:
  - E.g. house prize prediction
  - $J(\theta)\frac{1}{2m}[\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2+\lambda\sum_{j=1}^{n}\theta_j{}^2]$
- Large errors in prediction → What to try next???
  - Get more training examples
  - Try smaller set of features (to avoid overfitting)
  - Try getting additional features
  - Try adding polynomial features (e.g. x_1^2, x_2^2, x_1*x_2)
  - Try decreasing λ
  - Try increasing λ

Debugging

- Machine Learning Diagnostics
- Diagnostic: A test that you can run to get insight what is or is not working with a learning algorithm, and gain guidance as to how best improve its performance.
- Diagnostic can take time to implement, but can help a lot.

QUESTION



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

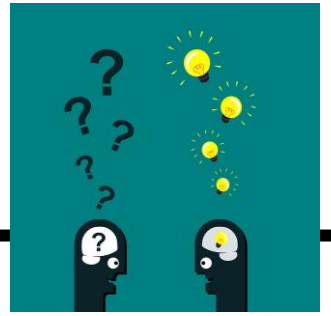
A: It's hard to tell what will work to improve a learning algorithm, so the best approach is to go with gut feeling and just see what works.

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

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

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

QUESTION



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

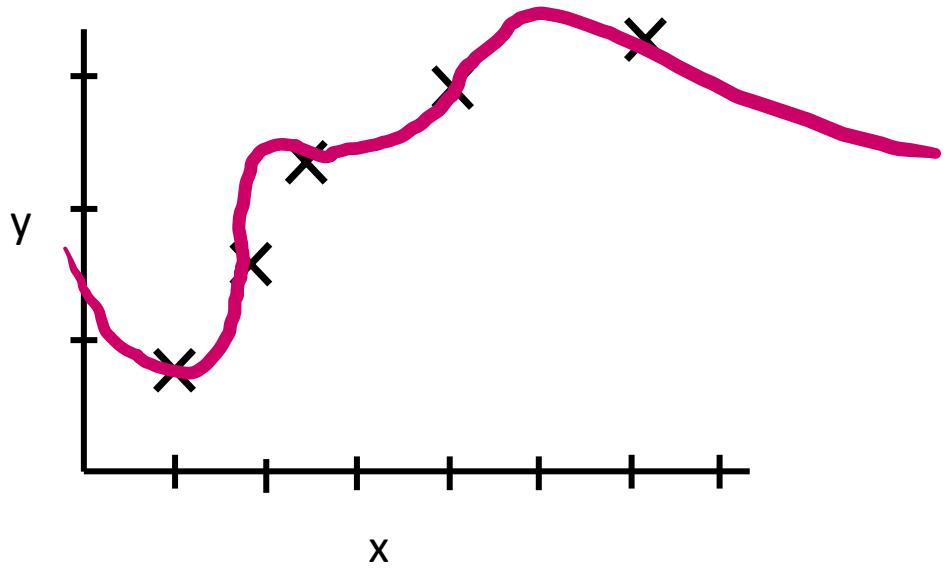
A: It's hard to tell what will work to improve a learning algorithm, so the best approach is to go with gut feeling and just see what works.

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

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

D: 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



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

→ Overfit + high variance

Fails to generalize to new examples that are not part of the training set.

X_1 = size of the house
X_2 = no. of bedrooms
X_3 = no. of floors
X_4 = age of the house
X_5 = average income in neighbourhood
X_6 = kitchen size

….
X_{100}

Evaluating a Hypothesis

70 %

Size in feet²	Price in K $
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

30 %

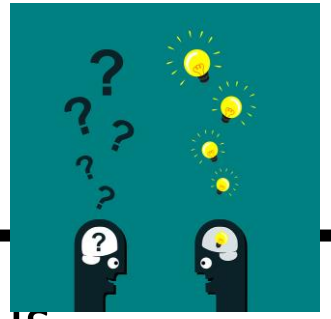
Training set

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

Test set

$(x_{test}^{(1)},y_{test}^{(1)})$
$(x_{test}^{(2)},y_{test}^{(2)})$
....
$(x_{test}^{(m\_test)},y_{test}^{(m\_test)})$

QUESTION



Suppose an implementation of linear regression (without regularization) is badly overfitting the training set. In this case, we would expect:

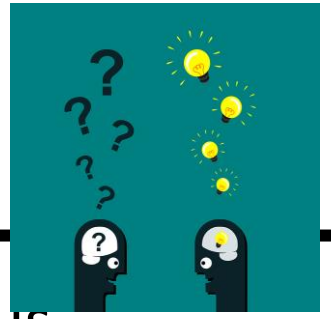
A: The training error $J(\theta)$ to be low and the test error $J_{test}(\theta)$ to be high

B: The training error $J(\theta)$ to be low and the test error $J_{test}(\theta)$ to be low

C: The training error $J(\theta)$ to be high and the test error $J_{test}(\theta)$ to be low

D: The training error $J(\theta)$ to be high and the test error $J_{test}(\theta)$ to be high

QUESTION



Suppose an implementation of linear regression (without regularization) is badly overfitting the training set. In this case, we would expect:

A: The training error $J(\theta)$ to be low and the test error $J_{test}(\theta)$ to be high

B: The training error $J(\theta)$ to be low and the test error $J_{test}(\theta)$ to be low

C: The training error $J(\theta)$ to be high and the test error $J_{test}(\theta)$ to be low

D: The training error $J(\theta)$ to be high and the test error $J_{test}(\theta)$ to be high

Evaluating a Hypothesis

Example linear regression:

- Learn parameter θ from training data (e.g. with 70% of data → minimizing training error $J(\theta)$)
- Compute test set error

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

Evaluating a Hypothesis

Example logistic regression:

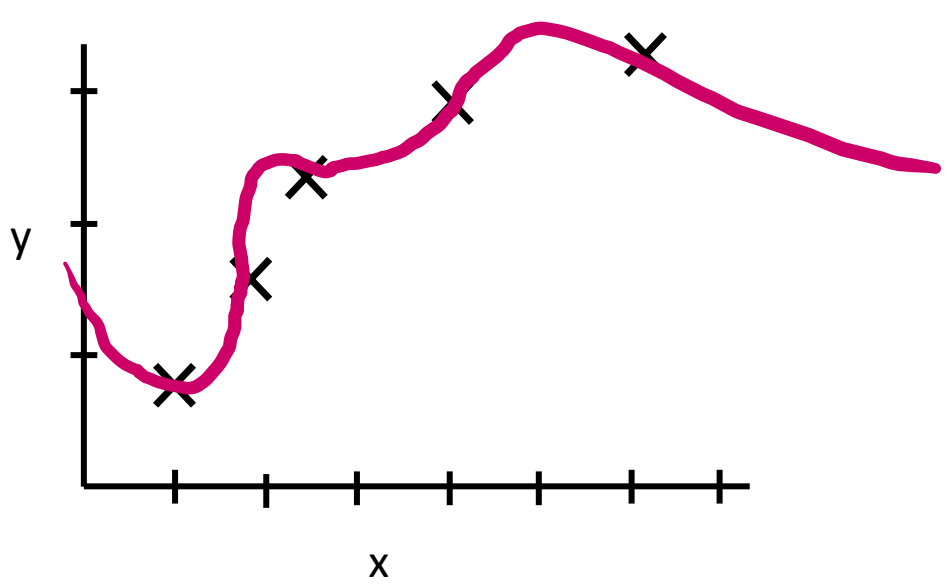
- Learn parameter θ from training data
- Compute test set error
- $J_{test}(\theta) = -\frac{1}{m\_test}[\sum_{i=1}^{m\_test} y_{test}^{(i)} \log(h_\theta(x_{test}^{(i)})) + (1-y_{test}^{(i)})\log(1-h_\theta(x_{test}^{(i)}))]$
- Alternative: Misclassification error (0/1 misclassification error)

  $err(h_\theta(x), y) = 1$ if $h_\theta(x) >= 0.5$ and y = 0 or $h_\theta(x) < 0.5$ and y = 1 (error case)

  0 otherwise

Test error = $\frac{1}{m\_test}\sum_{i=1}^{m\_test} err(h_\theta(x_{test}^{(i)}), y_{test}^{(i)})$

Model Selection



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

→ Overfit

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

Model Selection

d = degree of polynomial

- d= 1: $h_\theta(x) = \theta_0 + \theta_1 x \rightarrow \theta^1 \rightarrow J_{test}(\theta^1)$
- d= 2: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow \theta^2 \rightarrow J_{test}(\theta^2)$
- d= 3: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \rightarrow \theta^3 \rightarrow J_{test}(\theta^3)$
- …
- d= 10: $h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{10} x^{10} \rightarrow \theta^{10} \rightarrow J_{test}(\theta^{10})$

- Choose lowest test set error, e.g. $h_\theta(x) = \theta_0 + \ldots + \theta_5 x^5$
- Report test set error $J_{test}(\theta^5)$
- Problem: d is fit to the test set → $J_{test}(\theta^5)$ is likely to be an overly optimistic estimate of generalization error. Hypothesis is likely to do better on test set than it would on new examples.

Evaluating a Hypothesis

Size in feet²	Price in K $
2104	400
1600	330
2400	369
1416	232
3000	540
1985	300
1534	315
1427	199
1380	212
1494	243

60 % → Training set

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

20 % → Cross validation set

$(x_{cv}^{(1)},y_{cv}^{(1)})$
$(x_{cv}^{(2)},y_{cv}^{(2)})$
....
$(x_{cv}^{(m\_cv)},y_{cv}^{(m\_cv)})$

20 % → Test set

$(x_{test}^{(1)},y_{test}^{(1)})$
$(x_{test}^{(2)},y_{test}^{(2)})$
....
$(x_{test}^{(m\_test)},y_{test}^{(m\_test)})$

Evaluating a Hypothesis

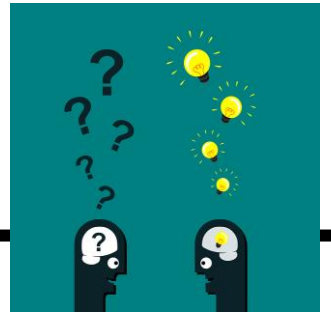
- Training error:
  - $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2$
- Cross validation error:
  - $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i)-{y_{cv}}^i\right)^2$
- Test error:
  - $J_{test}(\theta)\frac{1}{2m\_test}\sum_{i=1}^{m\_test}\left(h_\theta({x_{test}}^i)-{y_{test}}^i\right)^2$

Model Selection

- d= 1: $h_\theta(x) = \theta_0 + \theta_1 x \rightarrow \theta^1 \rightarrow J_{cv}(\theta^1)$
- d= 2: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \rightarrow \theta^2 \rightarrow J_{cv}(\theta^2)$
- d= 3: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 \rightarrow \theta^3 \rightarrow J_{cv}(\theta^3)$
- …
- d= 10: $h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{10} x^{10} \rightarrow \theta^{10} \rightarrow J_{cv}(\theta^{10})$

- Choose lowest test cross validation error, e.g. $h_\theta(x) = \theta_0 + \ldots + \theta_4 x^4$
- Estimate generalization error for test set, e.g. $J_{test}(\theta^4)$

QUESTION



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 $J_{CV}(\theta)$ to be lower than $J_{test}(\theta)$ because:

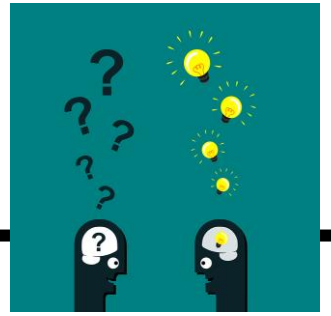
A: An extra parameter (d, the degree of the polynomial) has been fit to the cross validation set.

B: An extra parameter (d, the degree of the polynomial) has been fit to the test set.

C: The cross validation set is usually smaller than the test set.

D: The cross validation set is usually larger than the test set.

QUESTION



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 $J_{CV}(\theta)$ to be lower than $J_{test}(\theta)$ because:

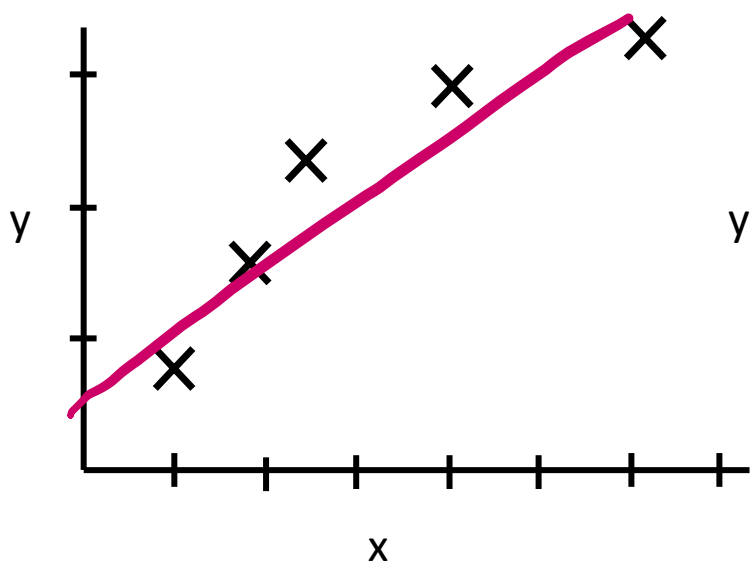
A: An extra parameter (d, the degree of the polynomial) has been fit to the cross validation set.

B: An extra parameter (d, the degree of the polynomial) has been fit to the test set.

C: The cross validation set is usually smaller than the test set.

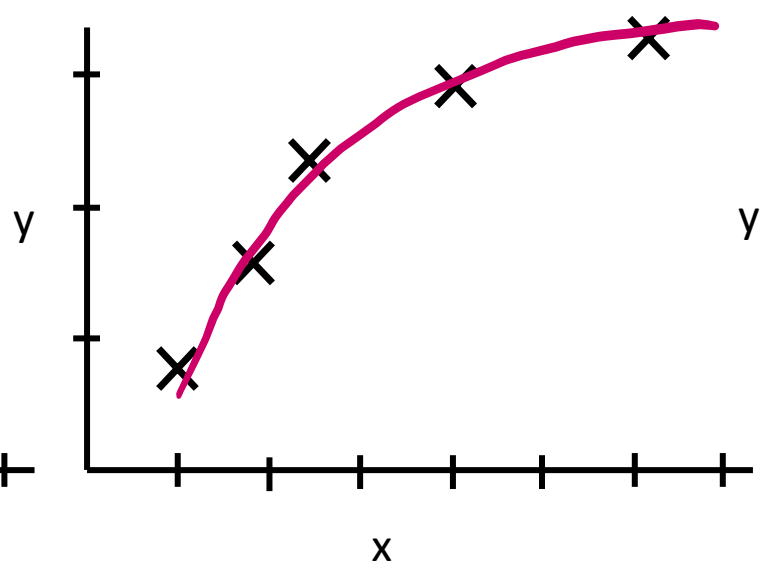
D: The cross validation set is usually larger than the test set.

Bias and Variance

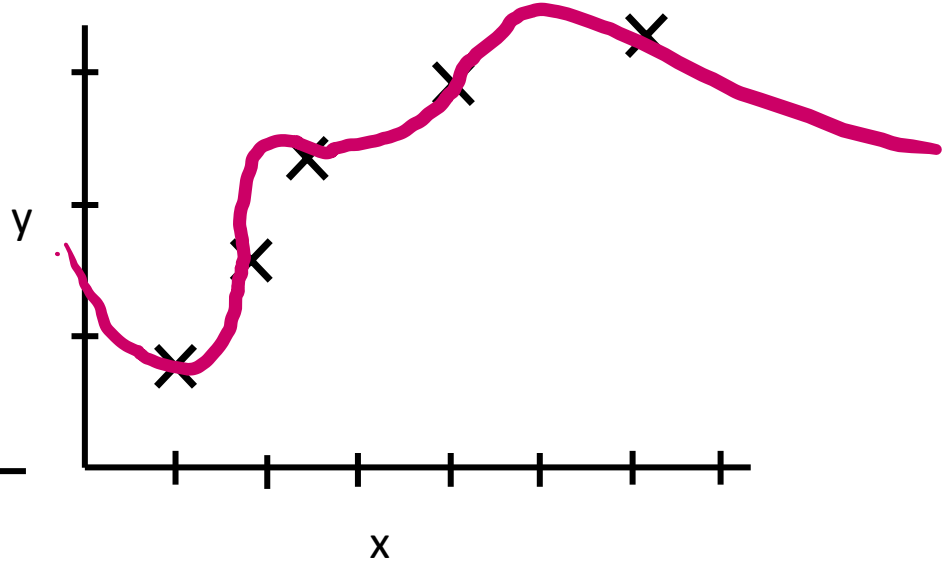


$h_\theta(x) = \theta_0 + \theta_1 x$

→ Underfit + high bias



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

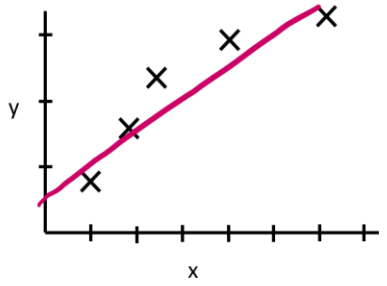


$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

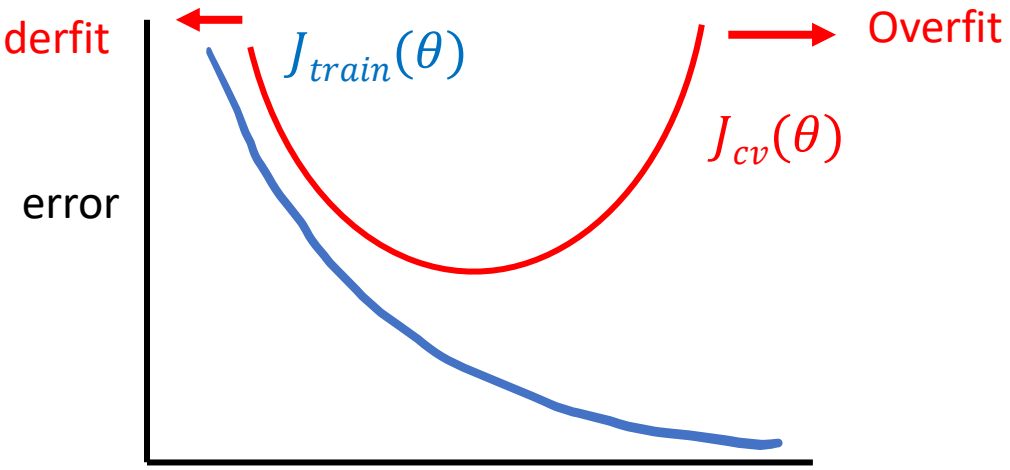
→ Overfit + high variance

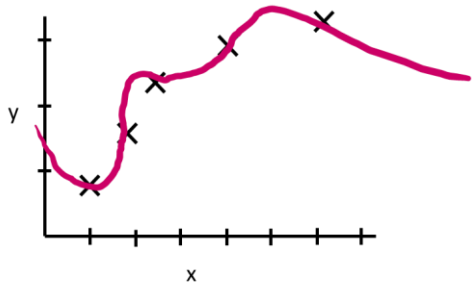
Bias and Variance

- Training error:
  - $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2$
- Cross validation error:
  - $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i)-{y_{cv}}^i\right)^2$ $(similiar\ for\ Jtest\ (\theta))$



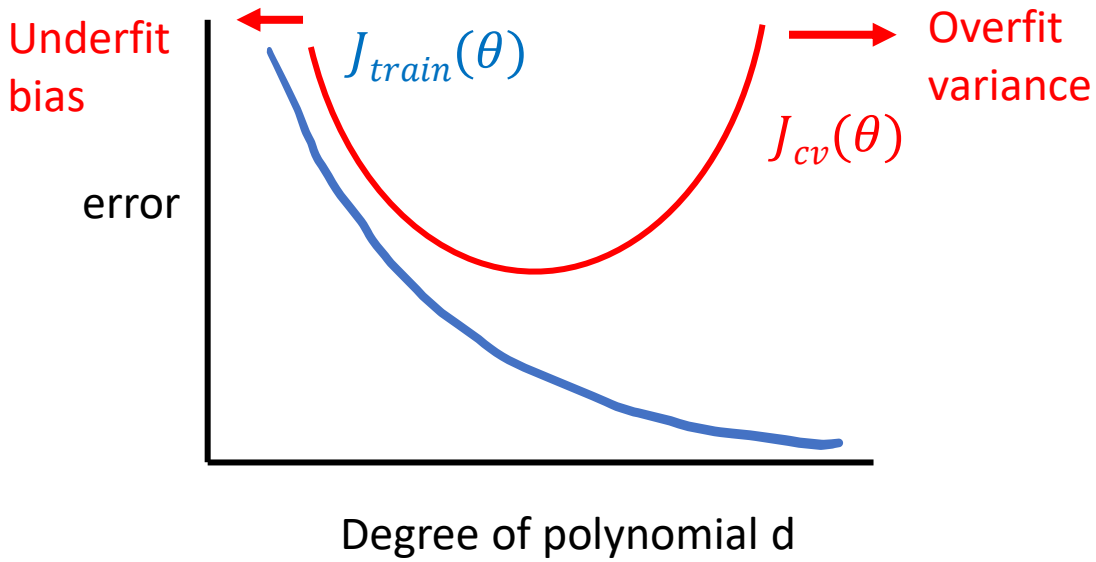
d = 1





Diagnosing Bias and Variance

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



Bias (underfit):

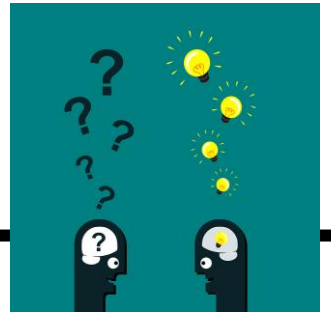
$J_{train}(\theta)$ and $J_{cv}(\theta)$ will be high

Variance (overfit):

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

$J_{cv}(\theta) >> J_{train}(\theta)$

QUESTION



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 $(x_{cv}^{(1)}, y_{cv}^{(1)}), \ldots, (x_{cv}^{(m\_cv)}, y_{cv}^{(m\_cv)})$

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?

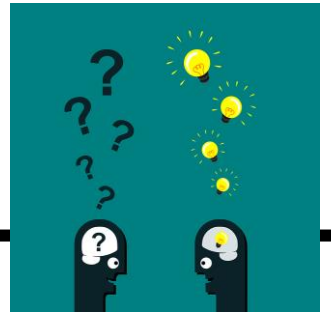
A: High bias (overfitting)

B: High bias (underfitting)

C: High variance (overfitting)

D: High variance (underfitting)

QUESTION



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 $(x_{cv}^{(1)}, y_{cv}^{(1)}), \ldots, (x_{cv}^{(m\_cv)}, y_{cv}^{(m\_cv)})$

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?

A: High bias (overfitting)

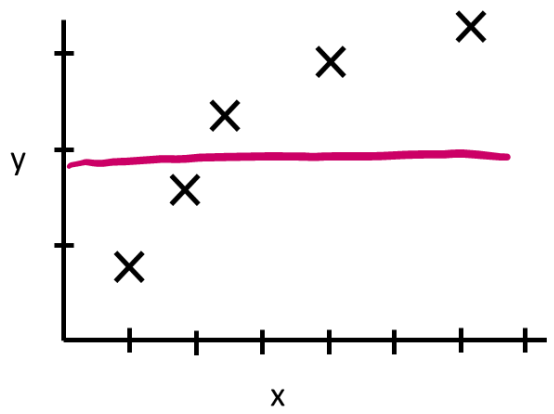
B: High bias (underfitting)

C: High variance (overfitting)

D: High variance (underfitting)

REGULARIZATION

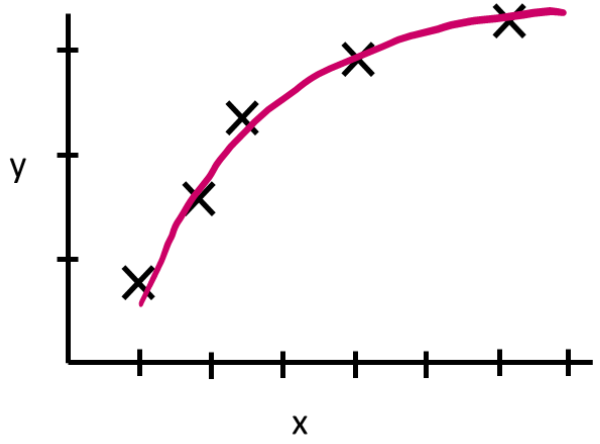
Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4 \rightarrow J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$



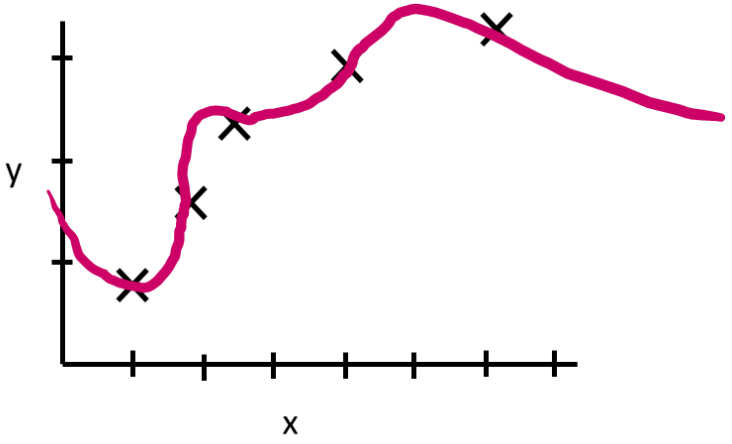
Large λ → Underfit + high bias

→$\lambda = 10000$ → $\theta_1 \sim 0$, $\theta_2 \sim 0$, ..

→$h_\theta(x) = \theta$



Intermediate λ



Small λ → high variance (overfit)

$\lambda \sim 0$

CHOOSING λ

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
- Cost function: $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$

- Training error:
  - $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2$
- Cross validation error:
  - $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i) - {y_{cv}}^i\right)^2$
- Test error:
  - $J_{test}(\theta)\frac{1}{2m\_test}\sum_{i=1}^{m\_test}\left(h_\theta({x_{test}}^i) - {y_{test}}^i\right)^2$

CHOOSING λ

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$
- Cost function: $J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i) - y^i\right)^2 + \frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$

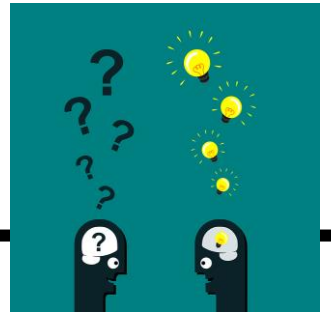
1. Try $\lambda = 0 \rightarrow \min_\theta J(\theta) \rightarrow \theta^1 \rightarrow J_{cv}(\theta^1)$
2. Try $\lambda = 0.01 \rightarrow \min_\theta J(\theta) \rightarrow \theta^2 \rightarrow J_{cv}(\theta^2)$
3. Try $\lambda = 0.02 \rightarrow \min_\theta J(\theta) \rightarrow \theta^3 \rightarrow J_{cv}(\theta^3)$
4. Try $\lambda = 0.04 \rightarrow \min_\theta J(\theta) \rightarrow \theta^4 \rightarrow J_{cv}(\theta^4)$
5. Try $\lambda = 0.08 \rightarrow \min_\theta J(\theta) \rightarrow \theta^5 \rightarrow J_{cv}(\theta^5)$

…

12. Try $\lambda = 10 \rightarrow \min_\theta J(\theta) \rightarrow \theta^{12} \rightarrow J_{cv}(\theta^{12})$

Pick e.g. θ^5 (as the best results on the cv set) → Test error: $J_{test}(\theta^5)$

QUESTION

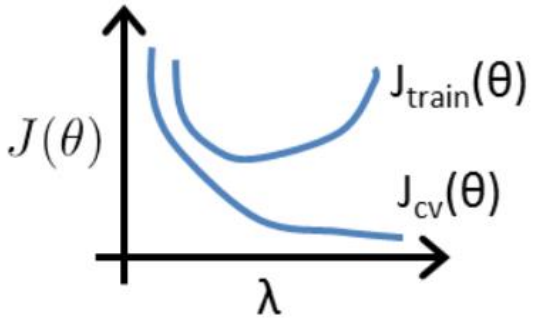


Consider regularized logistic regression. Let

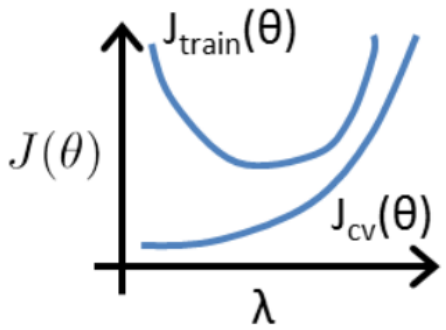
- $J(\theta)=\frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2+\frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$
- $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta({x_{train}}^i)-{y_{train}}^i\right)^2$
- $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i)-{y_{cv}}^i\right)^2$

Suppose you plot J_{train} and J_{CV} as a function of the regularization parameter λ. which of the following plots do you expect to get?

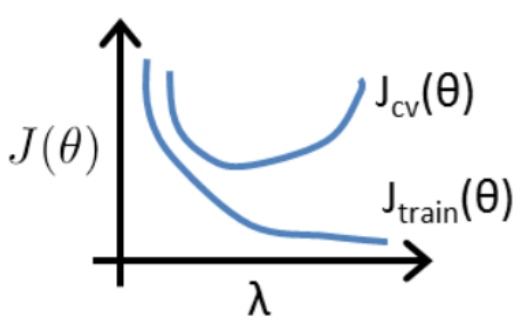
A:



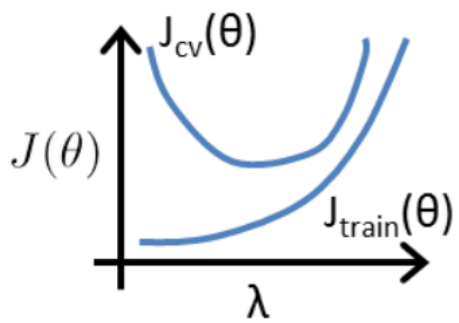
B:



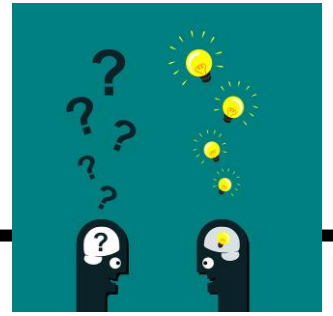
C:



D:



QUESTION

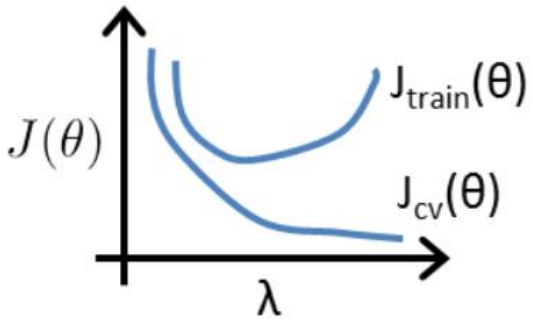


Consider regularized logistic regression. Let

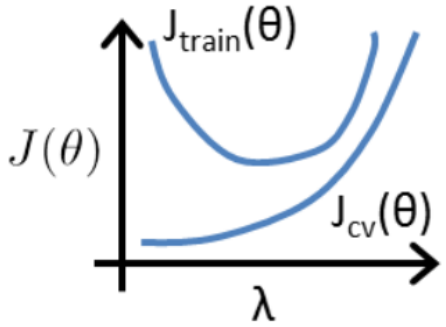
- $J(\theta)=\frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2+\frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$
- $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta({x_{train}}^i)-{y_{train}}^i\right)^2$
- $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i)-{y_{cv}}^i\right)^2$

Suppose you plot J_{train} and J_{CV} as a function of the regularization parameter λ. which of the following plots do you expect to get?

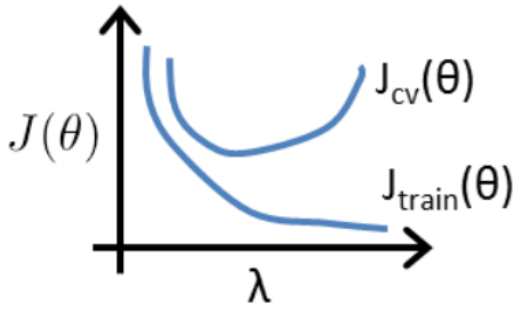
A:



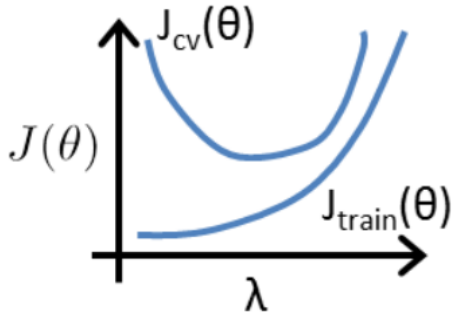
B:



C:

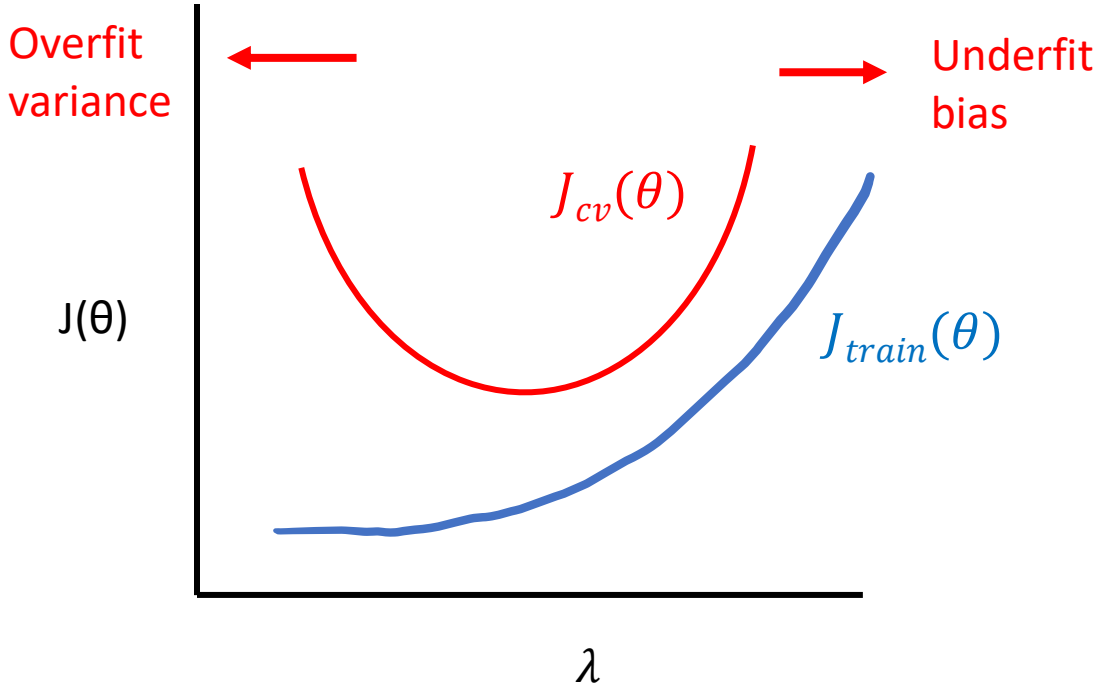


D:



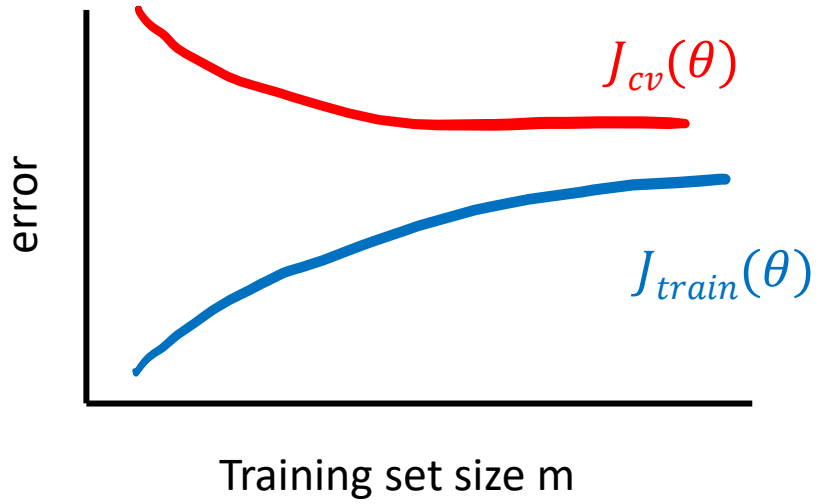
BIAS AND VARIANCE AS A FUNCTION OF λ

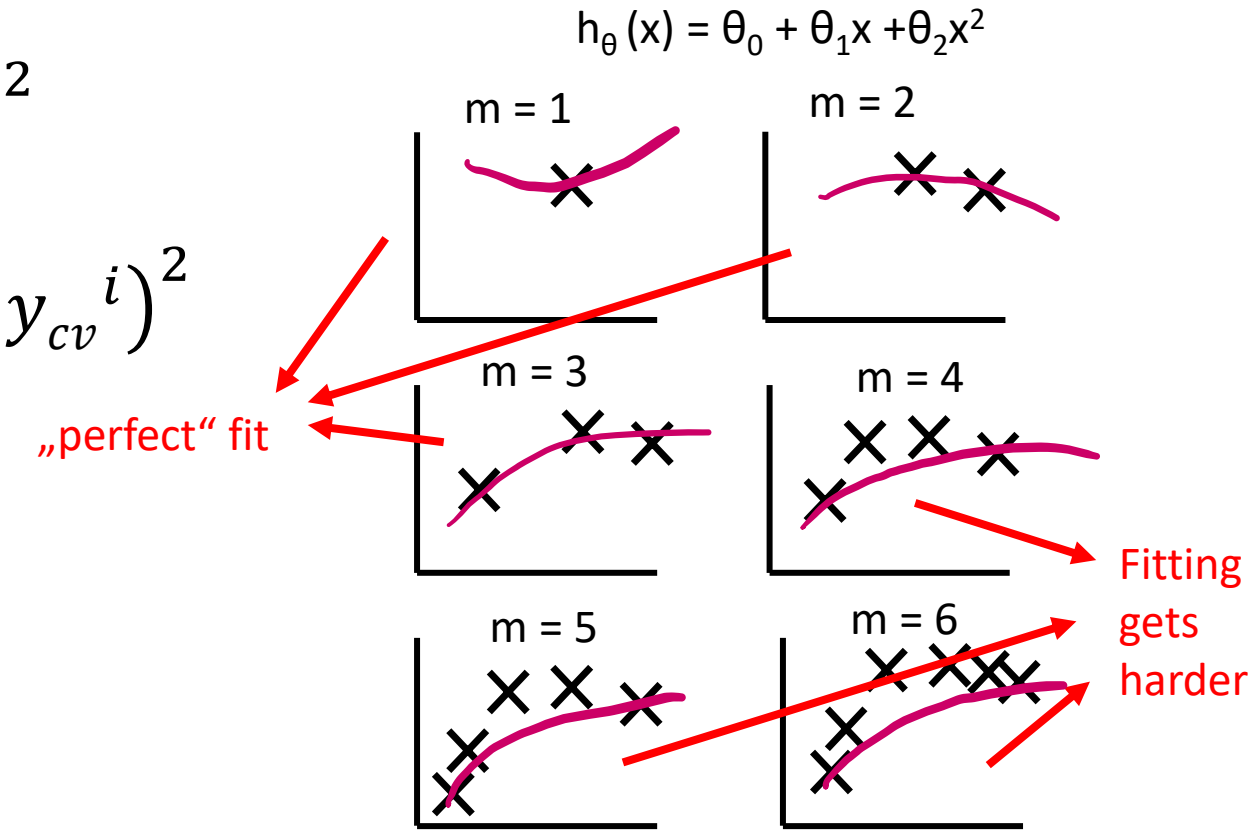
- $J(\theta)=\frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2+\frac{\lambda}{2m}\sum_{i=1}^{n}\theta_j^2\right]$
- $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta({x_{train}}^i)-{y_{train}}^i\right)^2$
- $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta({x_{cv}}^i)-{y_{cv}}^i\right)^2$



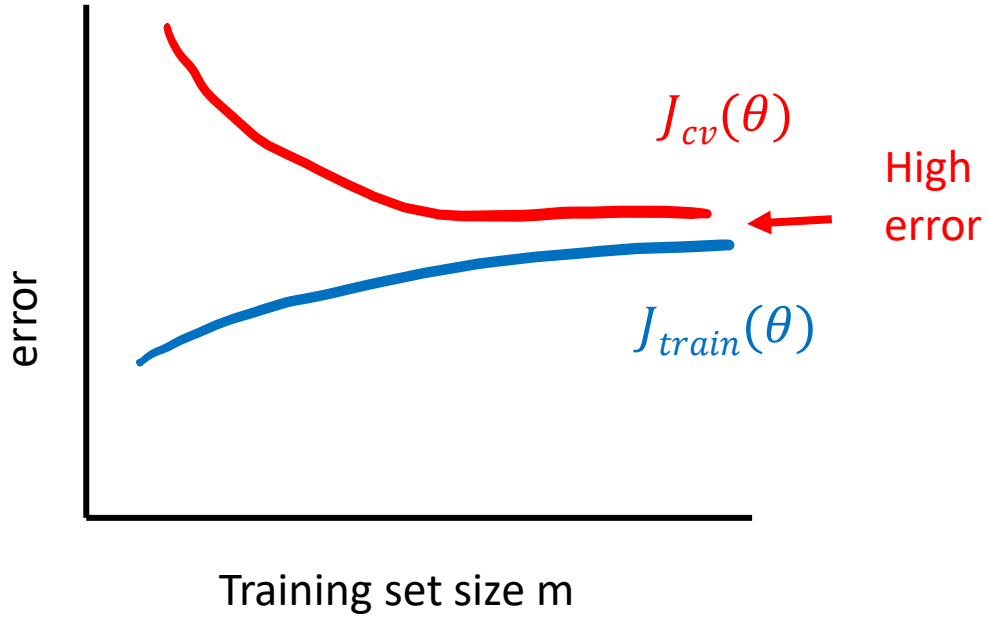
Learning Curves

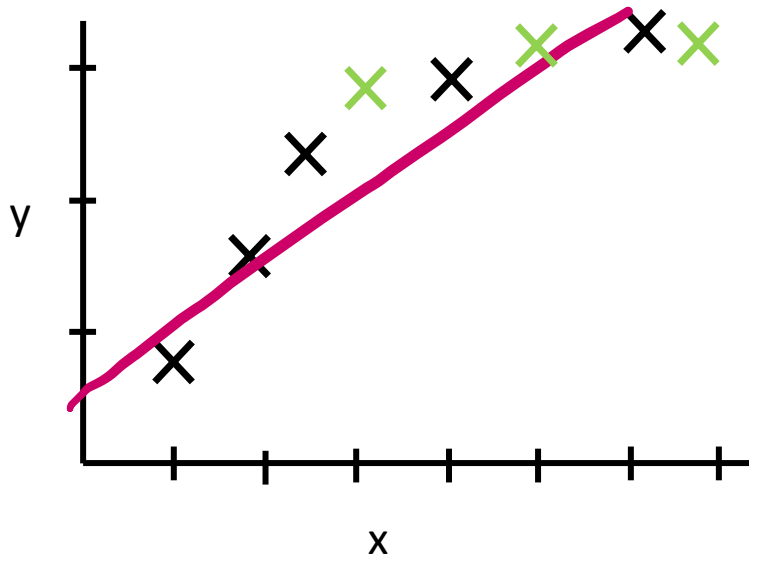
- $J_{train}(\theta)\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^i)-y^i\right)^2$
- $J_{cv}(\theta)\frac{1}{2m\_cv}\sum_{i=1}^{m\_cv}\left(h_\theta(x_{cv}{}^i)-y_{cv}{}^i\right)^2$





Learning Curves – high bias - underfit

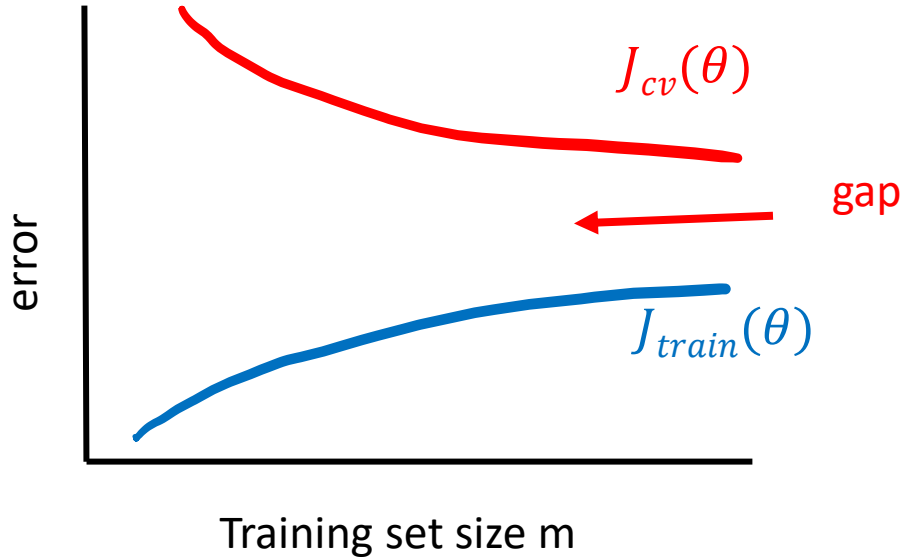


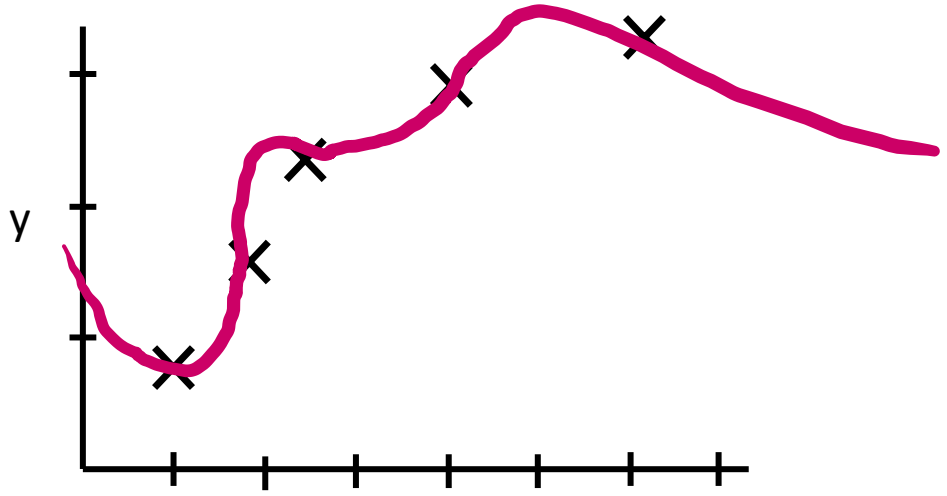


$h_\theta(x) = \theta_0 + \theta_1 x$

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

Learning Curves – high variance – Overfit

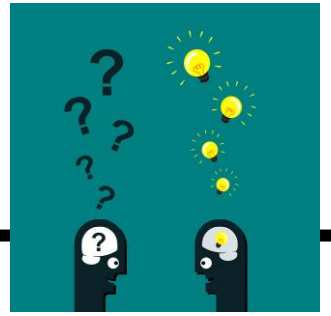




$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{100} x^{100}$ (and small λ)

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

QUESTION



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

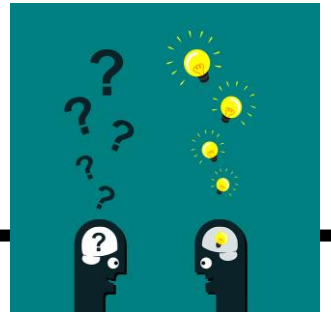
A: Algorithm is suffering from high bias.

B: Algorithm is suffering from high variance.

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

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

QUESTION



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

A: Algorithm is suffering from high bias.

B: Algorithm is suffering from high variance.

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

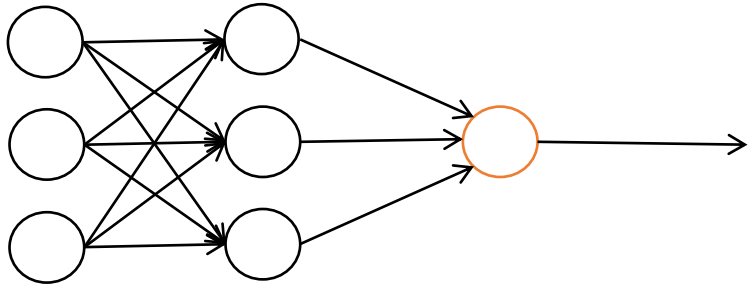
D: $J_{CV}(\theta)$ (cross validation error) is about the same as $J_{train}(\theta)$ (training error).

DEBUGGING

- Large errors in prediction → What to try next???
  - Get more training examples → fixes high variance
  - Try smaller set of features (to avoid overfitting) → fixes high variance
  - Try getting additional features → fixes high bias
  - Try adding polynomial features (e.g. x_1^2, x_2^2, $x_1 * x_2$) → fixes high bias
  - Try decreasing λ → fixes high bias
  - Try increasing λ → fixes high variance

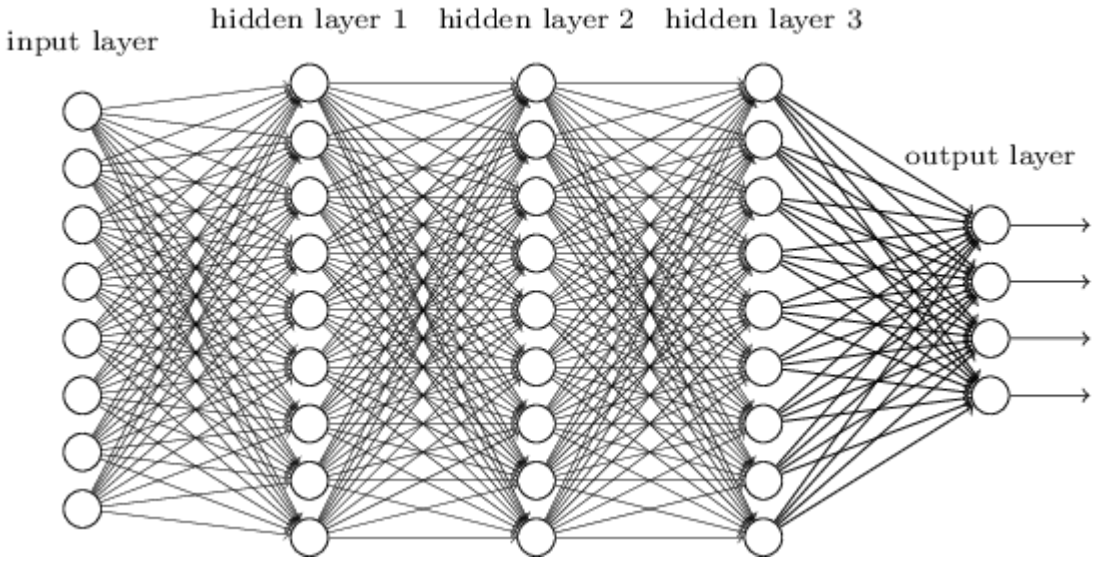
Neural Networks and Fitting Problems

- "Small" network
- fewer parameters
- more prone to underfitting
- Computationally cheaper



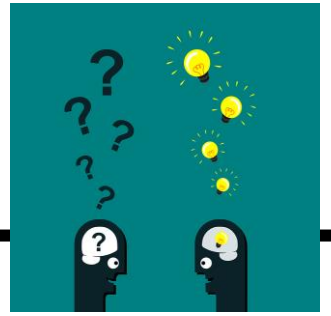
- "Large" network
- more parameters
- more prone to overfitting
- Computationally more expensive

→ User regularization to avoid overfitting



http://neuralnetworksanddeeplearning.com/chap5.html

QUESTION



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?

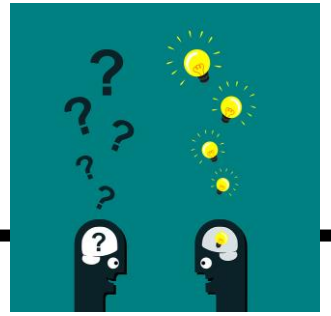
A: Yes, because this increases the number of parameters and lets the network represent more complex functions.

B: Yes, because it is currently suffering from high bias.

C: No, because it is currently suffering from high bias, so adding hidden units is unlikely to help.

D: No, because it is currently suffering from high variance, so adding hidden units is unlikely to help.

QUESTION



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?

A: Yes, because this increases the number of parameters and lets the network represent more complex functions.

B: Yes, because it is currently suffering from high bias.

C: No, because it is currently suffering from high bias, so adding hidden units is unlikely to help.

D: No, because it is currently suffering from high variance, so adding hidden units is unlikely to help.

Wrap-up

Evaluating a Hypothesis

- A hypothesis may have a low error for the training examples but still be inaccurate (because of overfitting). Thus, to evaluate a hypothesis, given a dataset of training examples, we can split up the data into two sets: a training set and a test set. Typically, the training set consists of 70 % of your data and the test set is the remaining 30 %.
- The new procedure using these two sets is then
  - Learn Θ and minimize $J_{train}(\Theta)$ using the training set
  - Compute the test set error $J_{test}(\Theta)$

WRAP-UP

Evaluating a Hypothesis

- The test set error
  - For linear regression: $J_{test}(\theta)\frac{1}{2m\_test}\sum_{i=1}^{m\_test}\left(h_\theta(xtest^i) - {y_{test}}^i\right)^2$
  - For classification ~ Misclassification error (aka 0/1 misclassification error):
    - err(h_θ(x),y) = 1 if h_θ(x)>=0.5 and y = 0 or h_θ(x)<0.5 and y = 1 (error case)

      0 otherwise
    - This gives us a binary 0 or 1 error result based on a misclassification. The average test error for the test set is:
    - Test error = $\frac{1}{m\_test}\sum_{i=1}^{m\_test}$ err $(h_\theta({x_{test}}^{(i)}), {y_{test}}^{(i)})$
    - This gives us the proportion of the test data that was misclassified.

WRAP-UP

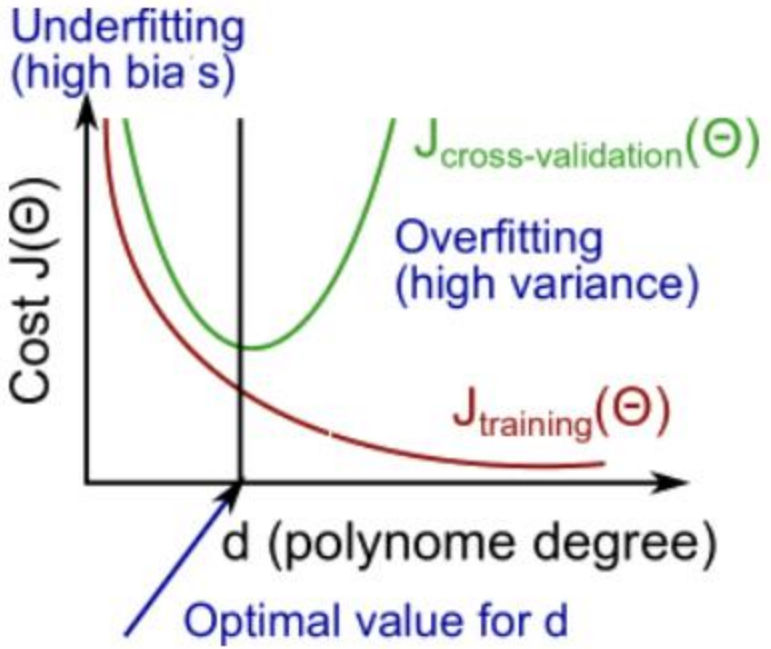
Model Selection and Train/Validation/Test Sets

- Just because a learning algorithm fits a training set well, that does not mean it is a good hypothesis. It could overfit and as a result your predictions on the test set would be poor. The error of your hypothesis as measured on the data set with which you trained the parameters will be lower than the error on any other data set.
- Given many models with different polynomial degrees, we can use a systematic approach to identify the 'best' function. In order to choose the model of your hypothesis, you can test each degree of polynomial and look at the error result.
- One way to break down our dataset into the three sets is:
  - Training set: 60%
  - Cross validation set: 20%
  - test set: 20%
- We can now calculate three separate error values for the three different sets using the following method:
  - Optimize the parameters in Θ using the training set for each polynomial degree.
  - Find the polynomial degree d with the least error using the cross validation set.
  - Estimate the generalization error using the test set with $J_{test}(\Theta^{(d)})$, (d = theta from polynomial with lower error);
  - This way, the degree of the polynomial d has not been trained using the test set.

Wrap-Up

Diagnosing Bias vs. Variance

- We have to examine the relationship between the degree of the polynomial d and the underfitting or overfitting of our hypothesis.
- We need to distinguish whether bias or variance is the problem contributing to bad predictions.
- High bias is underfitting and high variance is overfitting. Ideally, we need to find a golden mean between these two.
- The training error will tend to decrease as we increase the degree d of the polynomial.
- At the same time, the cross validation error will tend to decrease as we increase d up to a point, and then it will increase as d is increased, forming a convex curve.
- **High bias (underfitting):** both $J_{train}(\Theta)$ and $J_{CV}(\Theta)$ will be high. Also, $J_{CV}(\Theta) \approx J_{train}(\Theta)$.
- **High variance (overfitting):** $J_{train}(\Theta)$ will be low and $J_{CV}(\Theta)$ will be much greater than $J_{train}(\Theta)$.



WRAP-UP

Regularization and Bias/Variance

As λ increases, our fit becomes more rigid. On the other hand, as λ approaches 0, we tend to overfit the data. In order to choose the model and the regularization term λ, we need to:

- Create a list of lambdas (i.e. $\in\{0,0.01,0.02,0.04,0.08,0.16,0.32,0.64,1.28,2.56,5.12,10.24\}$)
- Create a set of models with different degrees or any other variants.
- Iterate through the λs and for each λ go through all the models to learn some Θ.
- Compute the cross validation error using the learned Θ (computed with λ) on the $J_{CV}(\Theta)$ without regularization or $\lambda = 0$.
- Select the best combo that produces the lowest error on the cross validation set.
- Using the best combo Θ and λ, apply it on $J_{test}(\Theta)$ to see if it has a good generalization of the problem.

WRAP-UP

Learning Curves

Training an algorithm on a very few number of data points (such as 1, 2 or 3) will easily have 0 errors because we can always find a e. g. quadratic curve that touches exactly those number of points.

- As the training set gets larger, the error for a quadratic function increases.
- The error value will plateau out after a certain m, or training set size.

Experiencing high bias:

- **Low training set size:** causes $J_{train}(\Theta)$ to be low and $J_{CV}(\Theta)$ to be high.
- **Large training set size:** causes both $J_{train}(\Theta)$ and $J_{CV}(\Theta)$ to be high with $J_{train}(\Theta) \approx J_{CV}(\Theta)$.
- If a learning algorithm is suffering from high bias, getting more training data will not (by itself) help much.

Experiencing high variance:

- **Low training set size:** $J_{train}(\Theta)$ will be low and $J_{CV}(\Theta)$ will be high.
- **Large training set size:** $J_{train}(\Theta)$ increases with training set size and $J_{CV}(\Theta)$ continues to decrease without flattening. Also, $J_{train}(\Theta) < J_{CV}(\Theta)$ and the difference between them remains significant.
- If a learning algorithm is suffering from high variance, getting more training data is likely to help.

Wrap-up

Deciding What to Do Next

- Getting more training examples: Fixes high variance
- Trying smaller sets of features: Fixes high variance
- Adding features: Fixes high bias
- Adding polynomial features: Fixes high bias
- Decreasing λ: Fixes high bias
- Increasing λ: Fixes high variance

Wrap-up

Diagnosing Neural Networks

- A neural network with fewer parameters is prone to underfitting. It is also computationally cheaper.
- A large neural network with more parameters is prone to overfitting. It is also computationally expensive. In this case you can use regularization (increase λ) to address the overfitting.
- Using a single hidden layer is a good starting default. You can train your neural network on a number of hidden layers using your cross validation set. You can then select the one that performs best.

Model Complexity Effects:

- Lower-order polynomials (low model complexity) have high bias and low variance. In this case, the model fits poorly consistently.
- Higher-order polynomials (high model complexity) fit the training data extremely well and the test data extremely poorly. These have low bias on the training data, but very high variance.
- In reality, we would want to choose a model somewhere in between, that can generalize well but also fits the data reasonably well.

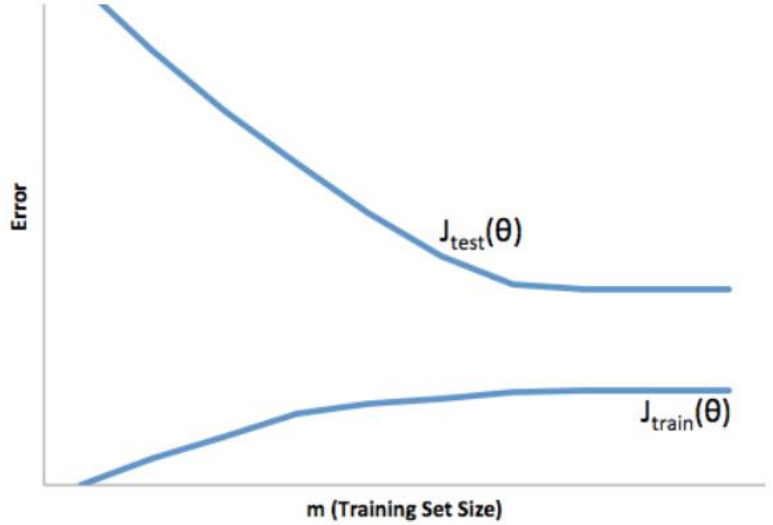
Quiz - Question 1

You train a learning algorithm and find that it has unacceptably high error on the test set. You plot the learning curve and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?

A: High bias

B: Neither

C: High variance



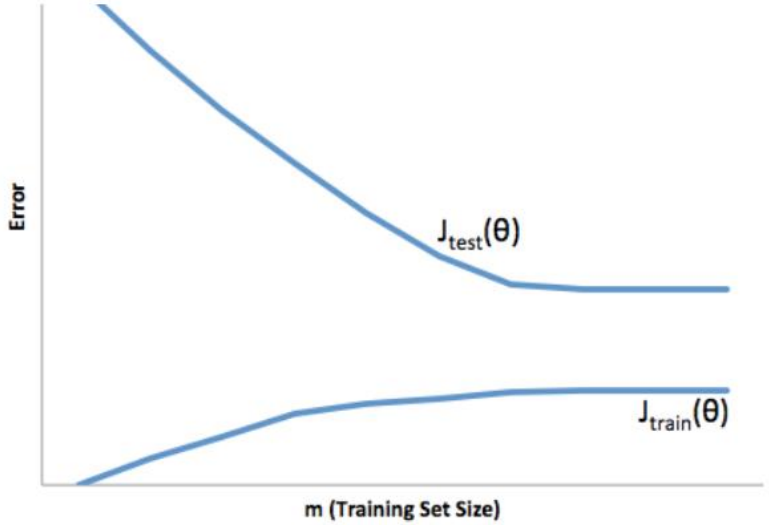
Quiz - Question 1

You train a learning algorithm and find that it has unacceptably high error on the test set. You plot the learning curve and obtain the figure below. Is the algorithm suffering from high bias, high variance, or neither?

A: High bias

B: Neither

C: High variance



Quiz - Question 2

Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs well (has low error) on the training set. Which of the following are promising steps to take? Check all that apply.

A: Try evaluating the hypothesis on a cross validation set rather than the test set.

B: Try using a smaller set of features.

C: Try decreasing the regularization parameter λ.

D: Try increasing the regularization parameter λ.

Quiz - Question 2

Suppose you have implemented regularized logistic regression to classify what object is in an image (i.e., to do object recognition). However, when you test your hypothesis on a new set of images, you find that it makes unacceptably large errors with its predictions on the new images. However, your hypothesis performs well (has low error) on the training set. Which of the following are promising steps to take? Check all that apply.

A: Try evaluating the hypothesis on a cross validation set rather than the test set.

B: Try using a smaller set of features.

C: Try decreasing the regularization parameter λ.

D: Try increasing the regularization parameter λ.

Quiz - Question 3

Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that it makes unacceptably large errors in its predictions. Furthermore, the hypothesis performs poorly on the training set. Which of the following might be promising steps to take? Check all that apply.

A: Try decreasing the regularization parameter λ.

B: Try adding polynomial features.

C: Try evaluating the hypothesis on a cross validation set rather than the test set.

D: Use fewer training examples.

Quiz - Question 3

Suppose you have implemented regularized logistic regression to predict what items customers will purchase on a web shopping site. However, when you test your hypothesis on a new set of customers, you find that it makes unacceptably large errors in its predictions. Furthermore, the hypothesis performs poorly on the training set. Which of the following might be promising steps to take? Check all that apply.

A: Try decreasing the regularization parameter λ.

B: Try adding polynomial features.

C: Try evaluating the hypothesis on a cross validation set rather than the test set.

D: Use fewer training examples.

Quiz - Question 4

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

A: Suppose you are training a logistic regression classifier using polynomial features and want to select what degree polynomial to use. After training the classifier on the entire training set, you decide to use a subset of the training examples as a validation set. This will work just as well as having a validation set that is separate (disjoint) from the training set.

B: It is okay to use data from the test set to choose the regularization parameter λ, but not the model parameters (θ).

C: A typical split of a data set into training, validation and test sets might be 60% training set, 20% validation set, and 20% test set.

D: Suppose you are using linear regression to predict housing prices, and your dataset comes sorted in order of increasing sizes of houses. It is then important to randomly shuffle the dataset before splitting it into training, validation and test sets, so that we don't have all the smallest houses going into the training set, and all the largest houses going into the test set.

Quiz - Question 4

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

A: Suppose you are training a logistic regression classifier using polynomial features and want to select what degree polynomial to use. After training the classifier on the entire training set, you decide to use a subset of the training examples as a validation set. This will work just as well as having a validation set that is separate (disjoint) from the training set.

B: It is okay to use data from the test set to choose the regularization parameter λ, but not the model parameters (θ).

C: A typical split of a data set into training, validation and test sets might be 60% training set, 20% validation set, and 20% test set.

D: Suppose you are using linear regression to predict housing prices, and your dataset comes sorted in order of increasing sizes of houses. It is then important to randomly shuffle the dataset before splitting it into training, validation and test sets, so that we don't have all the smallest houses going into the training set, and all the largest houses going into the test set.

Quiz - Question 5

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

A: A model with more parameters is more prone to overfitting and typically has higher variance.

B: If the training and test errors are about the same (but high), adding more features will not help improve the results.

C: If a learning algorithm is suffering from high variance, adding more training examples is likely to improve the test error.

D: If a learning algorithm is suffering from high bias, only adding more training examples may not improve the test error significantly.

Quiz - Question 5

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

A: A model with more parameters is more prone to overfitting and typically has higher variance.

B: If the training and test errors are about the same (but high), adding more features will not help improve the results.

C: If a learning algorithm is suffering from high variance, adding more training examples is likely to improve the test error.

D: If a learning algorithm is suffering from high bias, only adding more training examples may not improve the test error significantly.