Machine Learning - Week 6

赵燕

目录

1 Evalutaing a Learning Algorithm

1.1 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 $(x_1^2, x_2^2, x_1x_2, \text{etc.})$
- Try decreasing λ
- Try increasing λ

机器学习诊断法:

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.

图 1: 机器学习诊断法

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.

未选择的是正确的

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

正确

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

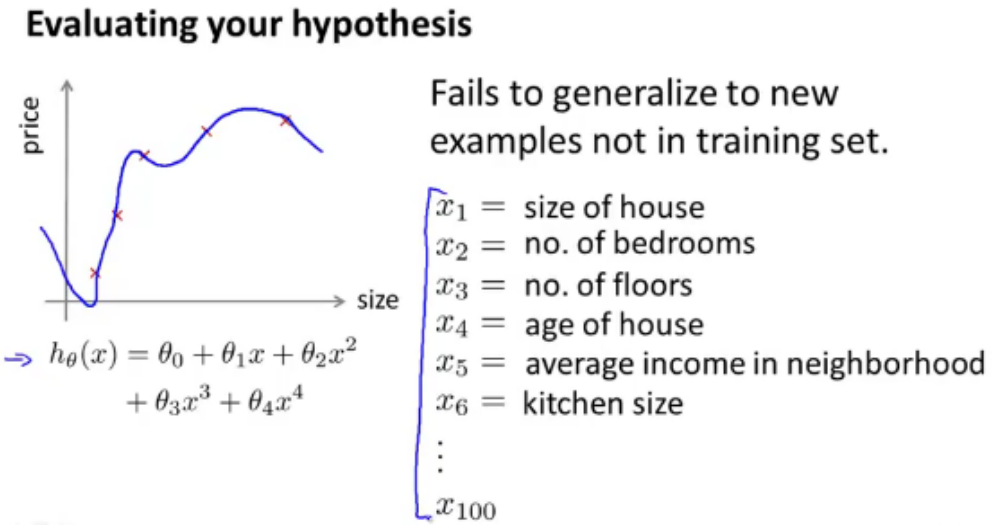
正确

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

正确

1.2 Evaluating a Hypothesis

怎样用你学过的算法来评估假设函数。



当确定学习算法的参数的时候，我们考虑的是选择参量来使训练误差最小化，有人 认为得到一个非常小的训练误差一定是一件好事，但已经知道，仅仅是因为这个假设具 有很小的训练误差，并不能说明它就一定是一个好的假设函数。而且也学习了过拟合假 设函数的例子，所以这推广到新的训练集上是不适用的。

那么，该如何判断一个假设函数是过拟合的呢？对于这个简单的例子，我们可以对假 设函数 h(x) 进行画图，然后观察图形趋势,但对于特征变量不止一个的这种一般情况，还 有像有很多特征变量的问题，想要通过画出假设函数来进行观察，就会变得很难甚至是不可 能实现。

因此，需要另一种方法来评估我们的假设函数过拟合检验。

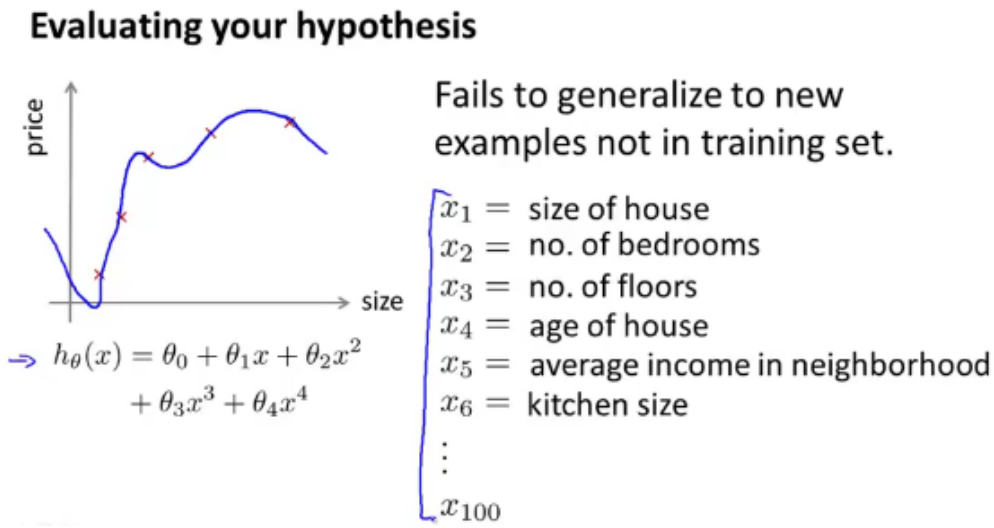


图 2: 训练集和测试集

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 $J_{\text{test}}(\theta)$ to be **high**

  正确

- The training error $J(\theta)$ to be **low** and the test error $J_{\text{test}}(\theta)$ to be **low**
- The training error $J(\theta)$ to be **high** and the test error $J_{\text{test}}(\theta)$ to be **low**
- The training error $J(\theta)$ to be **high** and the test error $J_{\text{test}}(\theta)$ to be **high**

按如下步骤训练和测试学习算法:

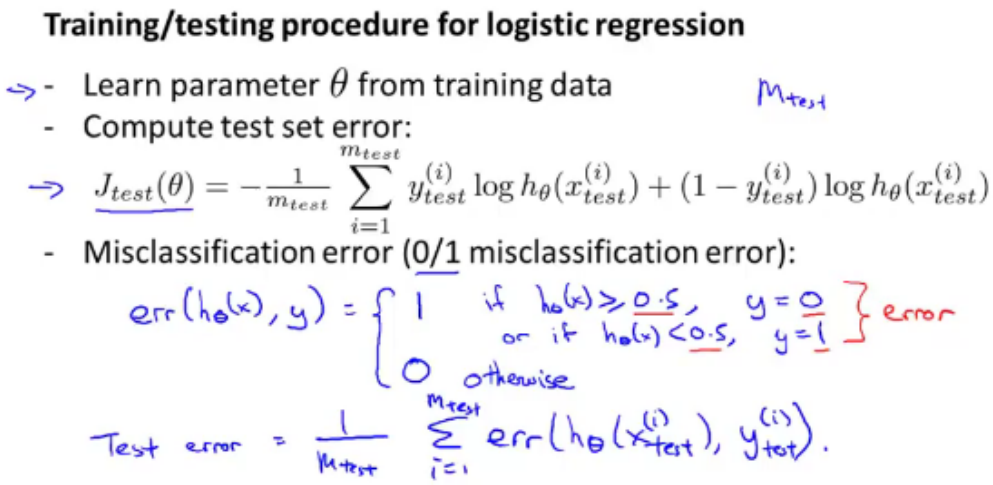


图 3: 训练集和测试集

1.2.1 Evaluating a Hypothesis

Once we have done some trouble shooting for errors in our predictions by:

- Getting more training examples
- Trying smaller sets of features
- Trying additional features
- Trying polynomial features
- Increasing or decreasing λ

We can move on to evaluate our new 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:

1.Learn Θ and minimize $J_{train}(\Theta)$ using the training set

2.Compute the test set error $J_{test}(\Theta)$

1.2.2 The test set error

1.For linear regression，利用测试集数据计算代价函数:

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

2.For classification　Misclassification error (aka 0/1 misclassification error):（误分率）

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

然后对计算结果求平均。

This gives us a binary 0 or 1 error result based on a misclassification. The average test error for the test set is:

$$Test \quad Error = \frac{1}{m_{test}} \sum_{i=1}^{m_{test}} err(h_\Theta(x_{test}^{(i)}), y_{test}^{(i)}) \tag{2}$$

This gives us the proportion of the test data that was misclassified.

1.3 Model Section and Train/Validation/Test Sets

（模型选择）怎样选择正确的特征来构造学习算法或者正确选择学习算法中的正则化参数λ:

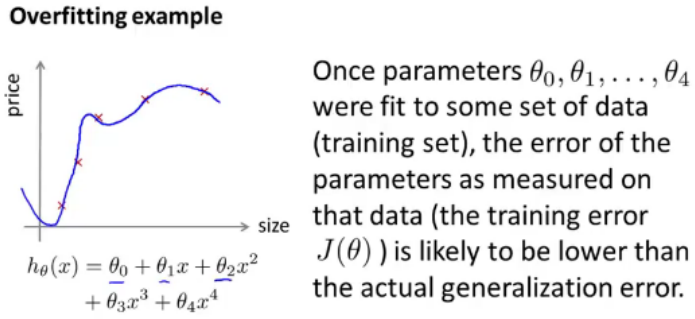


图 4: 过拟合示例

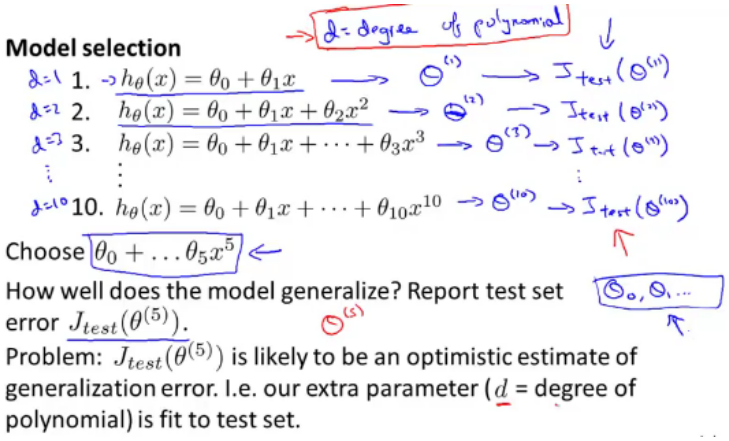


图 5: 模型选择

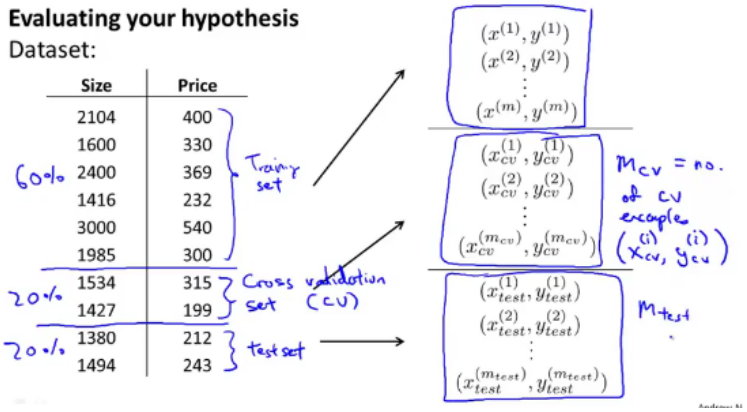


图 6: 训练集，交叉验证集和测试集

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:

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

图 7: 训练，交叉验证和测试误差

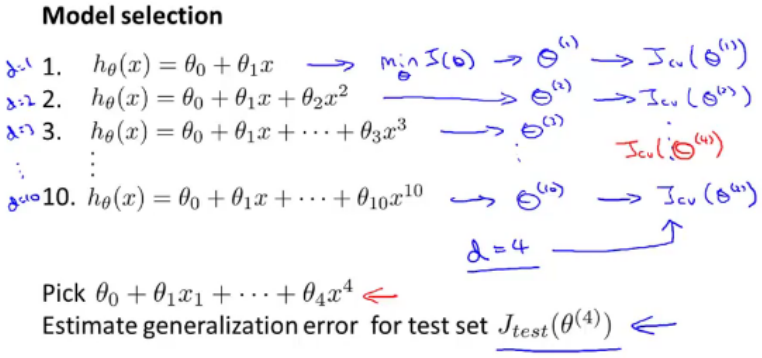


图 8: 模型选择

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

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

  正确

- 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.

Just because a learning algorithm fits a training set well, that does not mean it is a good hypothesis. It could over fit 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.

2 Bias vs. Variance

2.1 Diagnosing Bias vs. Variance

判断一个算法是偏差还是方差问题:

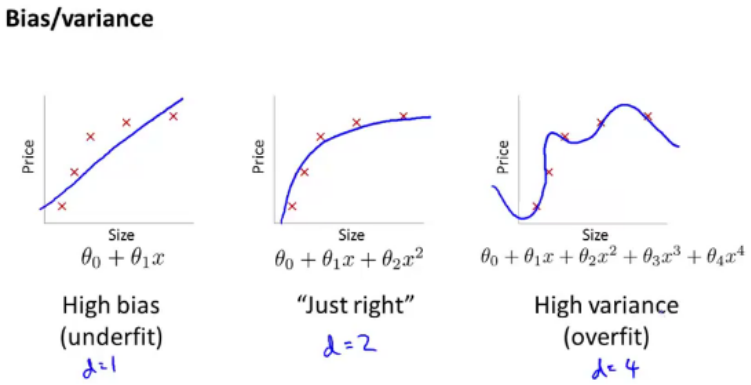


图 9: Bias/variance

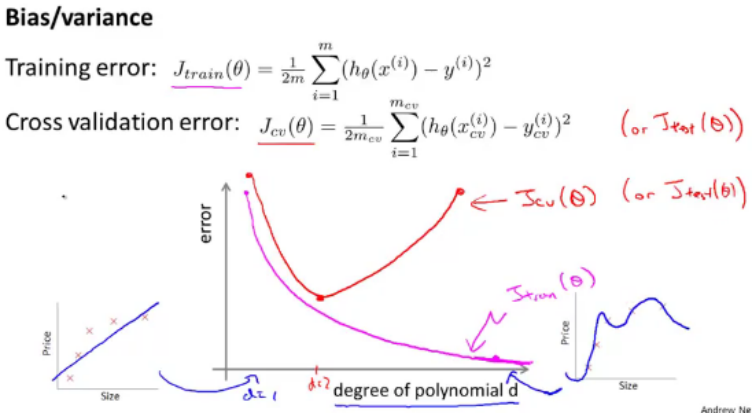


图 10: Bias/variance error

如何判断此时的算法处于哪个状态?

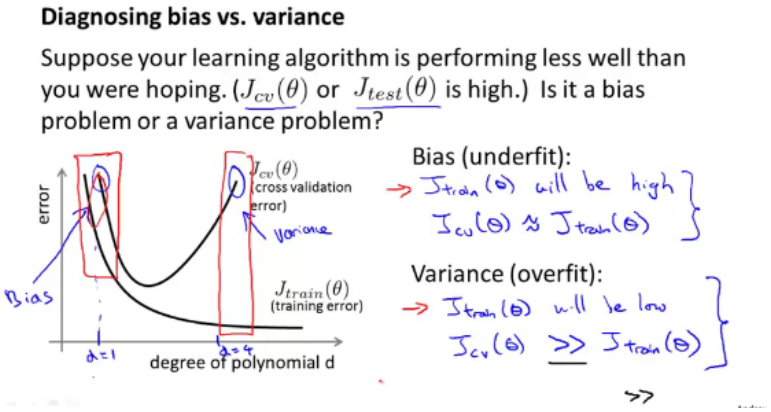


图 11: Diagnosing Bias/variance

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 $(x_{\text{cv}}^{(1)}, y_{\text{cv}}^{(1)}), \ldots, (x_{\text{cv}}^{(m_{\text{cv}})}, y_{\text{cv}}^{(m_{\text{cv}})})$. 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)

正确

- High variance (underfitting)

In this section we 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 $J_{train}(\Theta)$.

The is summarized in the figure below:

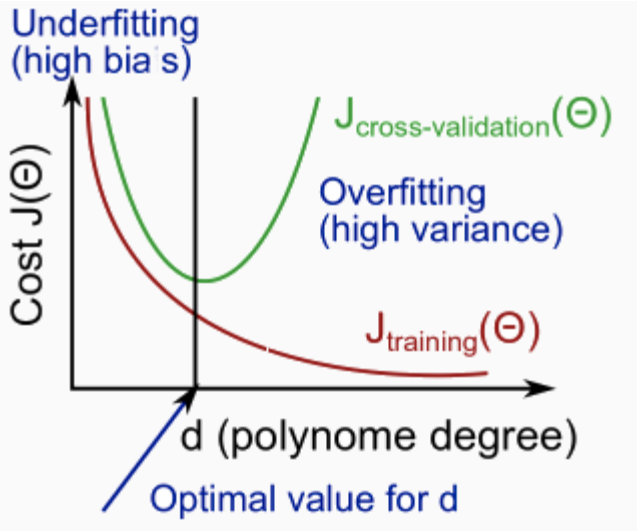


图 12: Diagnosing Bias/variance

2.2 Regularization and Bias/Variance

算法正则化与方差偏差的关系:

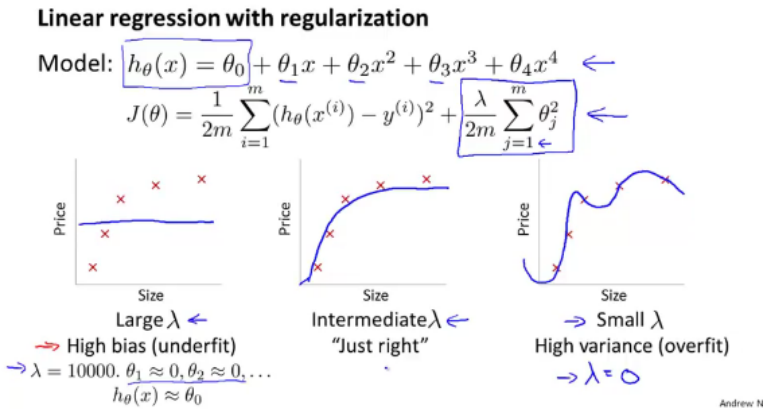
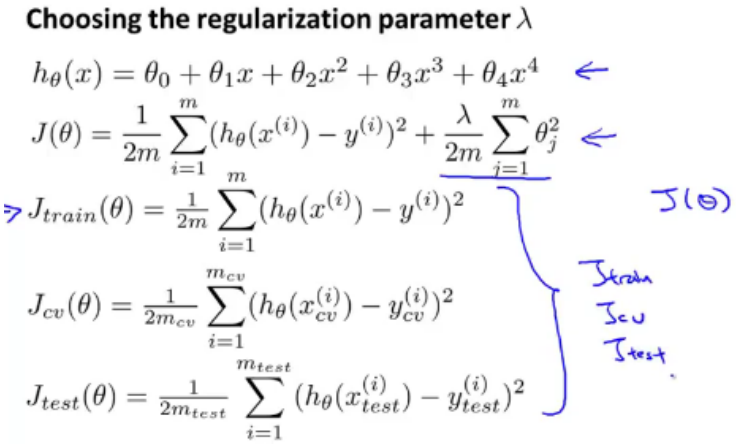


图 13: 线性回归与正则化



In the figure above, we see that as λ increases, our fit becomes more rigid. On the other hand, as λ approaches 0, we tend to over overfit the data. So how do we choose our parameter λ to get it 'just right' ? In order to choose the model and the regularization term λ, we need to:

1.Create a list of lambdas (i.e. $\lambda \in$ 0,0.01,0.02,0.04,0.08,0.16,0.32,0.64,1.28,2.56,5.12,10.24);

2.Create a set of models with different degrees or any other variants.

3.Iterate through the λs and for each λ go through all the models to learn some Θ.

4.Compute the cross validation error using the learned Θ (computed with λ) on the $J_{CV}(\Theta)$ without regularization or $\lambda = 0$.

5.Select the best combo that produces the lowest error on the cross validation set.

6.Using the best combo Θ and λ, apply it on $J_{test}(\Theta)$ to see if it has a good generalization of the problem.

当改变λ时，交叉验证集误差和训练集误差会发生怎样的变化?

我们选择一系列的想要测试的 λ 值,通常是 0-10 之间的呈现 2 倍关系的值(如: 0,0.01,0.02,0.04,0.08,0.15,0.32,0.64,1.28,2.56,5.12,10 共 12 个)。我们同样把数据分为训练集、交叉验证集和测试集。

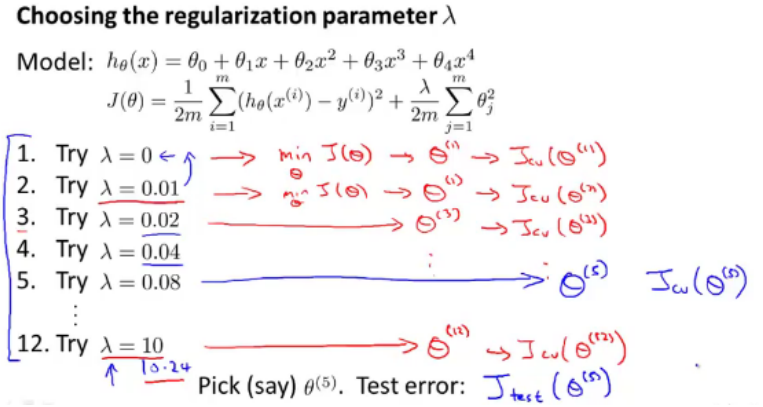


图 14: 选择合适的正则化参数λ

选择λ的方法是:

1. 使用训练集训练出 12 个不同程度正则化的模型

2. 用 12 模型分别对交叉验证集计算的出交叉验证误差

3. 选择得出交叉验证误差最小的模型

4. 运用步骤 3 中选出模型对测试集计算得出推广误差,我们也可以同时将训练集和交叉验证集模型的代价函数误差与 λ 的值绘制在一张图表上:

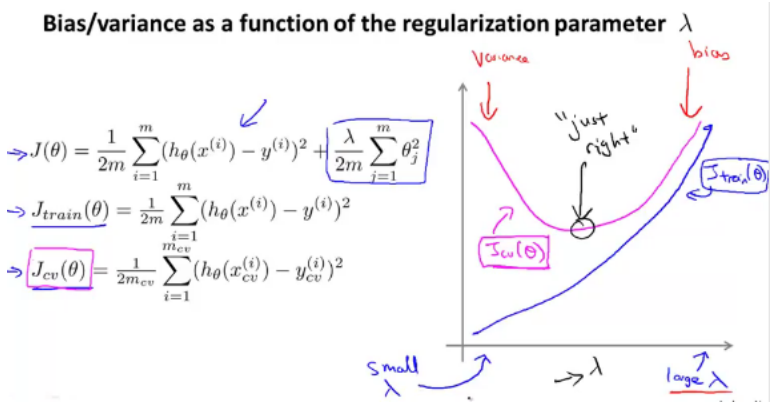


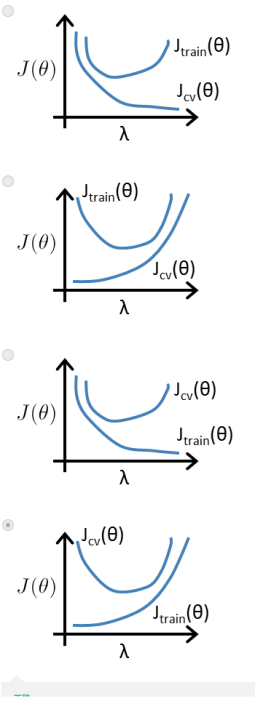
图 15: 正则化参数λ与偏差和方差

- 当 λ 较小时,训练集误差较小(过拟合)而交叉验证集误差较大
- 随着 λ 的增加,训练集误差不断增加(欠拟合),而交叉验证集误差则是先减小后 增加

Consider regularized logistic regression. Let

- $J(\theta) = \frac{1}{2m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=2}^{n}\theta_j^2]$
- $J_{\text{train}}(\theta) = \frac{1}{2m_{\text{train}}}[\sum_{i=1}^{m_{\text{train}}}(h_\theta(x_{\text{train}}^{(i)}) - y_{\text{train}}^{(i)})^2]$
- $J_{\text{CV}}(\theta) = \frac{1}{2m_{\text{CV}}}[\sum_{i=1}^{m_{\text{CV}}}(h_\theta(x_{\text{CV}}^{(i)}) - y_{\text{CV}}^{(i)})^2]$

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?



2.3 Learning Curves

学习曲线就是一种很好的工具,我经常使用学习曲线来判断某一个学习算法是否处于偏 差、方差问题。学习曲线是学习算法的一个很好的合理检验(sanity check)。学习曲线是将训练集误差和交叉验证集误差作为训练集实例数量(m)的函数绘制的图表。

即,如果我们有 100 行数据,我们从 1 行数据开始,逐渐学习更多行的数据。

思想是:当训练较少行数据的时候,训练的模型将能够非常完美地适应较少的训练数据,但是训练出来的模型却不能很好地适应交叉验证集数据或测试集数据。

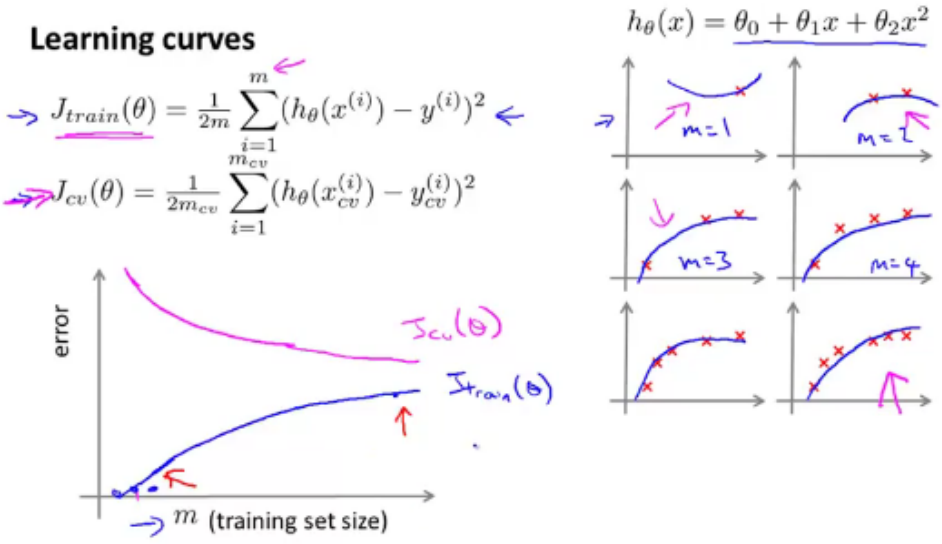


图 16: 学习曲线

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 quadratic curve that touches exactly those number of points. Hence:

1.As the training set gets larger, the error for a quadratic function increases.

2.The error value will plateau out after a certain m, or training set size.

1.训练集误差随着m的增大而增大

2.当训练集较小时，泛化程度不会很好，不能很好的适应新样本，因此这个假设不是一个理想的假设，只有当我使用一个更大的训练集时，才有可能得到一个能够更好拟合数据的 可能的假设，因此验证集误差和测试集误差都会随着训练集样本容量m的增加而减小，因为使用的数据越，越能获得更好地泛化表现，或者说对新样本的适应能力更强，因此数据越多 越能拟合出合适的假设。

当处于高方差或者高偏差的情况时这些曲线又会变成什么样子?

2.3.1 High bias

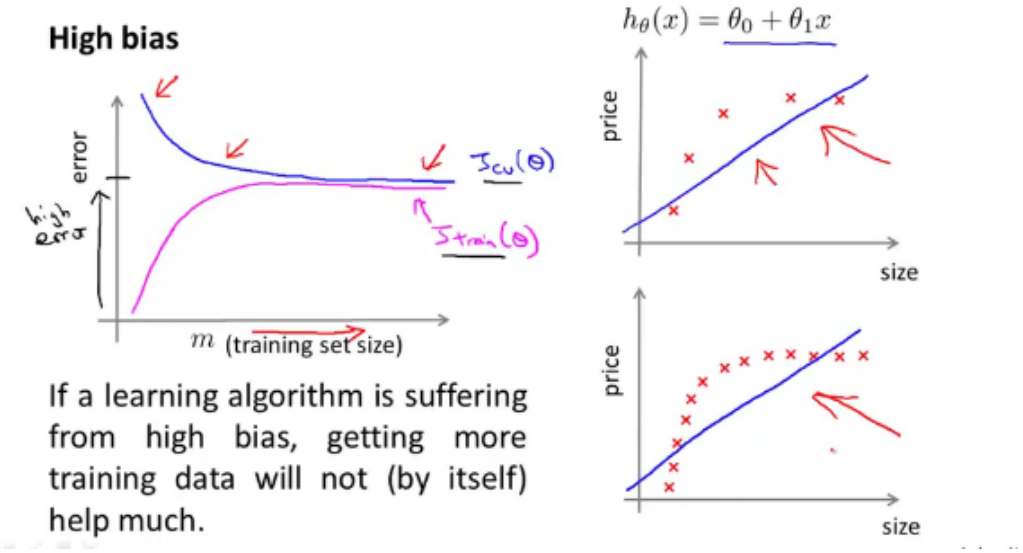


图 17: 高偏差学习曲线

最左端表示训练集样本容量很小，比如说只有一组样本，那么表现当然很不好，而随着你增大训练集样本数，当达到某一个容量值的时候 你就会找到那条最有可能拟合数据的那条线，并且此时即便你继续增大训练集的样本容量， 即使你不断增大m的值，你基本上还是会得到的一条差不多的直线，因此交叉验证集误差我把它标在这里，或者测试集误差，将会很快变为水平而不再变化。只要训练集样本容量值达到或超过了那个特定的数值，交叉验证集误差和测试集误差就趋于不变，这样你会得到最能拟合数据的那条直线;

那么，训练误差又如何呢?

在高偏差的情形中，你会发现训练集误差会逐渐增大，一直趋于接近交叉验证集误差，这是因为你的参数很少，但当m很大的时候，数据太多，此时训练集和交叉验证集的预测效果将会非常接近。

高偏差的情形反映出的问题是,交叉验证集和训练集误差都很大,也就是说,你最终会得到一个值比较大$J_{CV}(\Theta)$和$J_{train}(\Theta)$。

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.

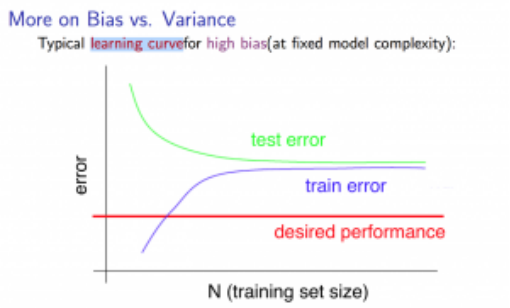


图 18: 高偏差学习曲线

2.3.2 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 leveling off. Also, $J_{train}(\Theta) < J_{CV}(\Theta)$ but the difference between them remains significant.

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

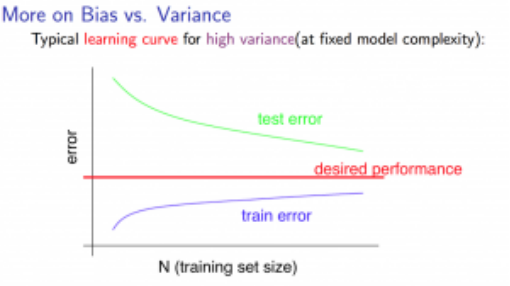


图 19: 高方差学习曲线1

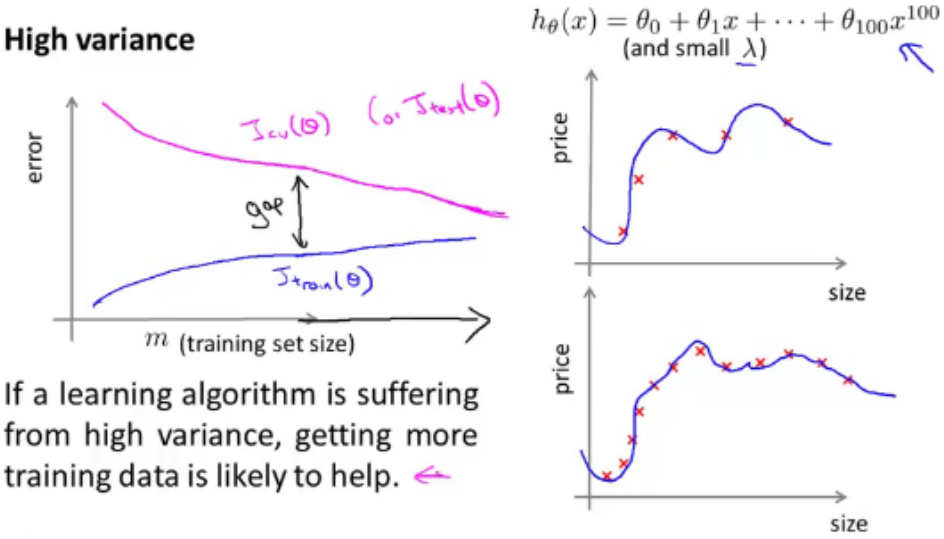


图 20: 高方差学习曲线2

如果继续增大训练样本的数量，将曲线向右延伸，交叉验证集误差将会逐渐下降，所以在高方差的情形中，使用更多的训练集数据对改进算法的表现，事实上是有效果的，这同样也体现出，知道你的算法正处于高方差的情形也是非常有意义的，因为它能告诉你，是否有必要花时间来增加更多的训练集数据。

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.

未选择的是正确的

- [x] Algorithm is suffering from high variance.

正确

- [x] $J_{CV}(\theta)$ (cross validation error) is much larger than $J_{train}(\theta)$ (training error).

正确

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

未选择的是正确的

2.4 Deciding What to Do Next Revisited

我们已经介绍了怎样评价一个学习算法,我们讨论了模型选择问题,偏差和方差的问题。 那么这些诊断法则怎样帮助我们判断,哪些方法可能有助于改进学习算法的效果,而哪些可 能是徒劳的呢?

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 → fixes high variance
- Try smaller sets of features → fixes high variance
- Try getting additional features → fixes high bias
- Try adding polynomial features $(x_1^2, x_2^2, x_1x_2, \text{etc})$ → fixes high bias.
- Try decreasing λ → fixes high bias
- Try increasing λ → fixes high variance.

让我们再次回到最开始的例子,在那里寻找答案,这就是我们之前的例子。回顾之前所学的 提出的六种可选的下一步,让我们来看一看我们在什么情况下应该怎样选择:

1.获得更多的训练实例——解决高方差

2. 尝试减少特征的数量——解决高方差

3. 尝试获得更多的特征——解决高偏差

4. 尝试增加多项式特征——解决高偏差

5. 尝试减少正则化程度 λ ——解决高偏差

6. 尝试增加正则化程度 λ ——解决高方差

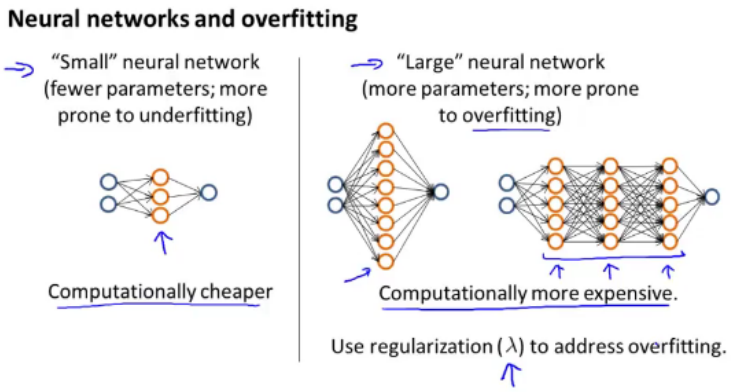


图 21: 诊断神经网络

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.

一般来说,使用一个大型的神经网络并使用正则化来修正过拟合问题,通常比使用一个小型的神经网络效果更好,但主要可能出现的问题 是计算量相对较大,最后还需要选择隐藏层的层数，你是应该用一个隐藏层呢，还是应该用三个呢，就像我们这里画的，或者还是用两个隐藏层呢？

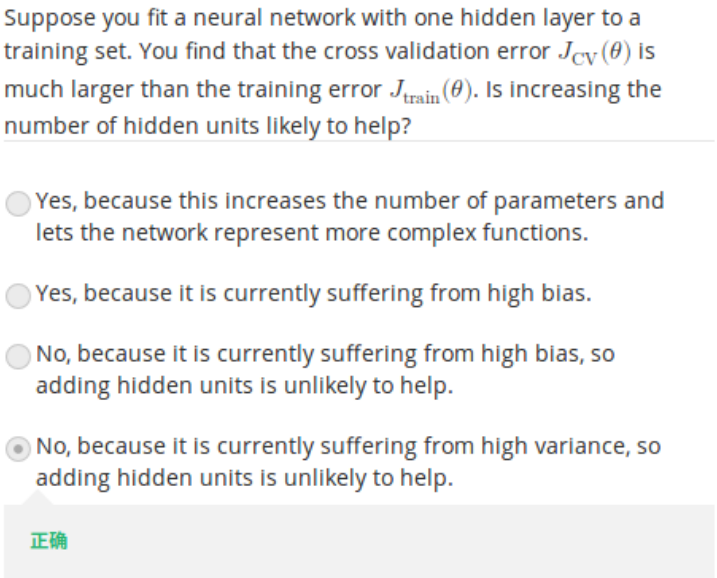
Using a single hidden layer is a good starting default. You can train your neural network on a number of hidden layers using your cross validation set. You can then select the one that performs best.

通常来说,正如我在前面的视频中讲过的,默认的情况是 ,使用一个隐藏层 ,但是如果你确实想要选择多个隐藏层 ,你也可以试试把数据分割为训练集, 验证集和测试集 ,然后使用交叉验证的方法 ,比较一个隐藏层的神经网络 ,然后试试两个 ,三个隐藏层 ,以此类推 ,然后看看哪个神经网络 ,在交叉验证集上表现得最理想 ,也就是说 ,你得到了三个神经网络模型 ,分别有一个 ,两个, 三个隐藏层 ,然后你对每一个模型都用交叉验证集数据进行测试 ,算出三种情况下的 ,交叉验证集误差$J_{CV}(\Theta)$然后选出你认为最好的神经网络结构。

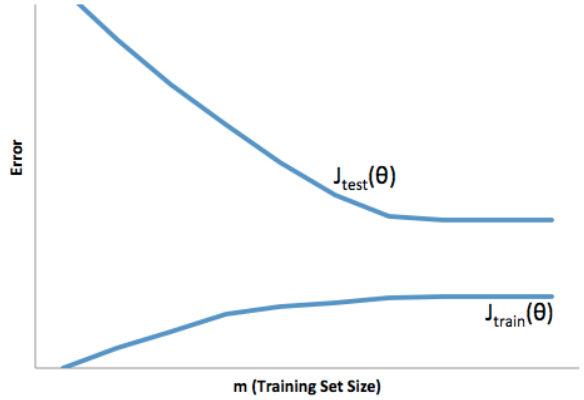
这就是偏差和方差问题以及诊断该问题的学习曲线方法，在改进学习算法的表现时 ，你可以充分运用以上这些内容来判断哪些途径可能是有帮助的，而哪些方法可能是无意义的。

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.



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?



- (●) High variance
- ( ) Neither
- ( ) High bias

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.

- [x] Get more training examples.
- [x] Try using a smaller set of features.
- [ ] Try adding polynomial features.
- [ ] Use fewer training examples.

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.

- [x] Try adding polynomial features.
- [x] Try to obtain and use additional features.
- [ ] Try using a smaller set of features.
- [ ] Try increasing the regularization parameter λ.

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

- [x] Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **cross validation** error.
- [ ] Suppose you are training a regularized linear regression model. The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **test set** error.
- [ ] Suppose you are training a regularized linear regression model.The recommended way to choose what value of regularization parameter λ to use is to choose the value of λ which gives the lowest **training set** error.
- [x] The performance of a learning algorithm on the training set will typically be better than its performance on the test set.

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

- [x] If a learning algorithm is suffering from high bias, only adding more training examples may **not** improve the test error significantly.
- [x] A model with more parameters is more prone to overfitting and typically has higher variance.
- [x] When debugging learning algorithms, it is useful to plot a learning curve to understand if there is a high bias or high variance problem.
- [ ] If a neural network has much lower training error than test error, then adding more layers will help bring the test error down because we can fit the test set better.

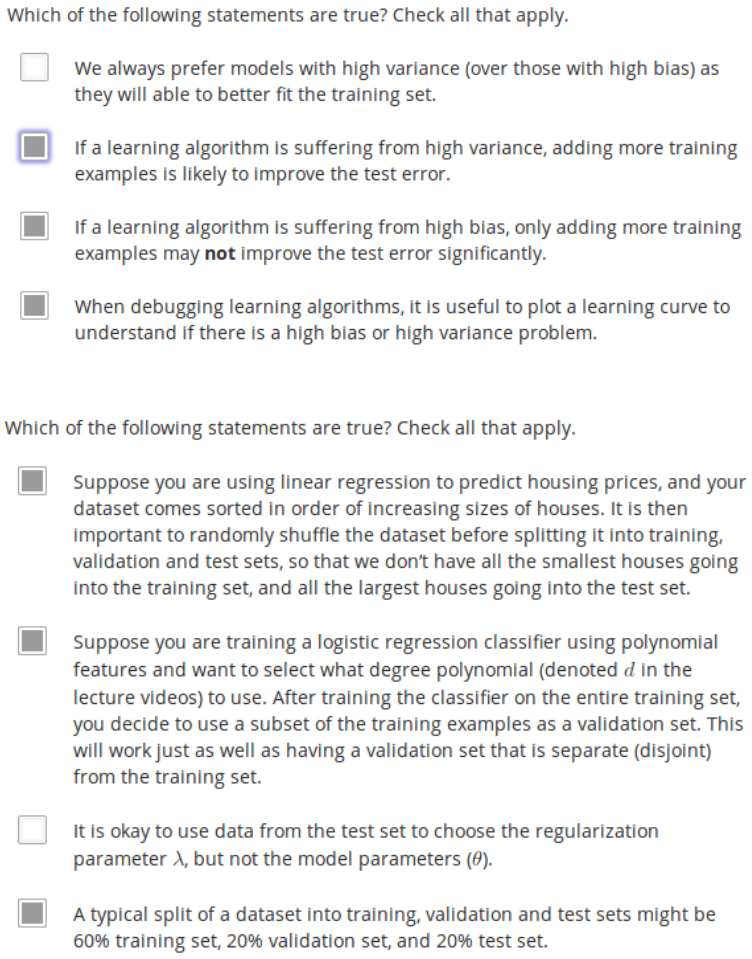


图 22: 此题有误

3 Building a Spam Classifier

3.1 Prioritizing What to Work On

3.2 Error Analysis

4 Handing Skewed Data

4.1 Error Metrics for Skewed Classes

4.2 Trading Off Precision and Recall

5 Using Large Data Sets

5.1 Data For Machine Learning