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

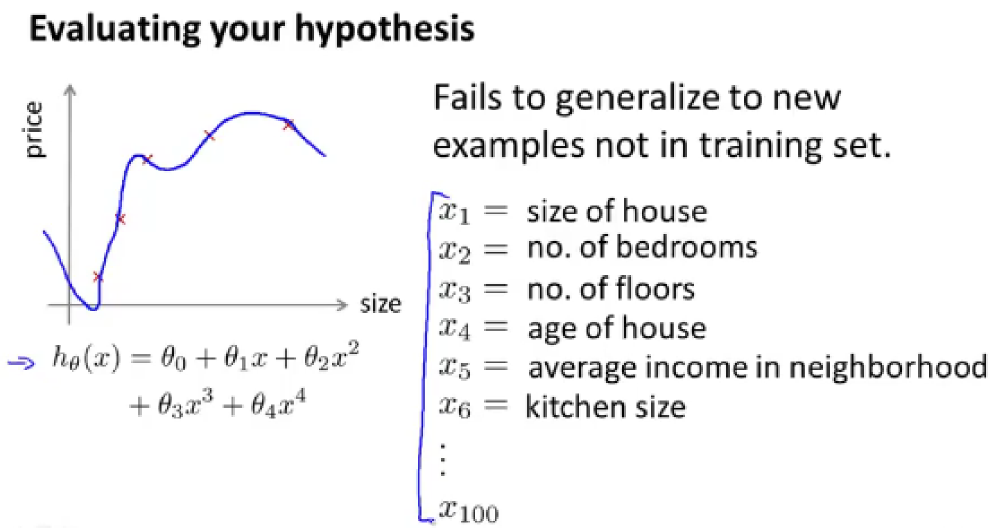
  正确

- [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。

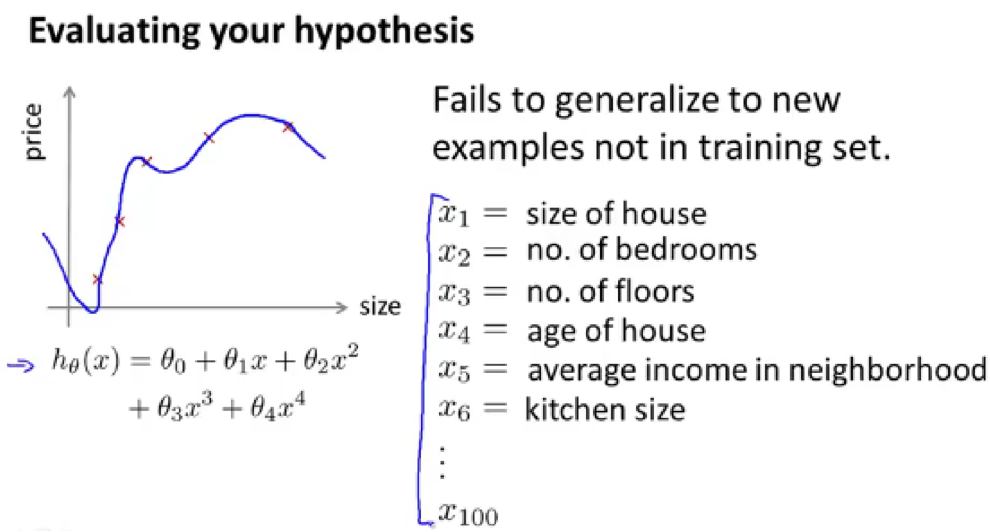


图 2: 训练集和测试集

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

- (selected) 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：

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):

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

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

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）

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

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

2 Bias vs. Variance

2.1 Diagnosing Bias vs. Variance

2.2 Regularization and Bias/Variance

2.3 Learning Curves

2.4 Deciding What to Do Next Revisited

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