Week 10

July 27, 2021

Debugging a learning algorithm

Imagine you've used regularized linear regression to forecast home prices:

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

- Trained it.
- However, when tested on new data, it produces unacceptably high errors in its predictions.
- What should your next step be?
  - Obtain additional training data.
  - Try a smaller set of features.
  - Consider getting more features.
  - Add polynomial features.
  - Change the value of λ.

Evaluating a hypothesis

- Split data into two portions: training set and test set.
- Learn parameters θ from training data, minimizing $J(\theta)$ using 70% of the training data.
- Compute the 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 and training validation test sets

- How should a regularization parameter or polynomial degree be chosen?
- We've previously discussed the issue of overfitting.
- This is why, in general, training set error is a poor predictor of hypothesis accuracy for new data (generalization).
- Try to determine the degree of polynomial that will fit data.

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

$\vdots$

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

- Introduce a new parameter d, which represents the degree of polynomial you want to use.
- Model 1 is minimized using training data, resulting in a parameter vector θ^1 (where d =1).
- Same goes for other models up to n.
- Using the previous formula, examine the test set error for each computed parameter $J_{test}(\theta^k)$.
- Minimize cost function for each of the models as before.
- Test these hypothesis on the cross validation set to generate the cross validation error.
- Pick the hypothesis with the lowest cross validation error.

Training error:

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

Cross Validation error:

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

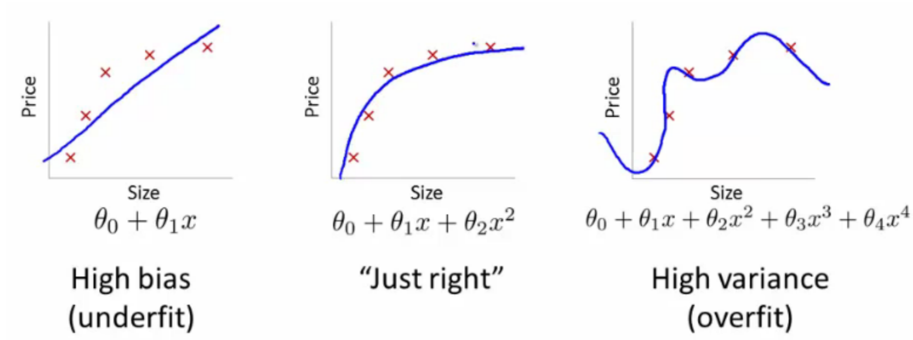
Test error:

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

Model selection and training validation test sets

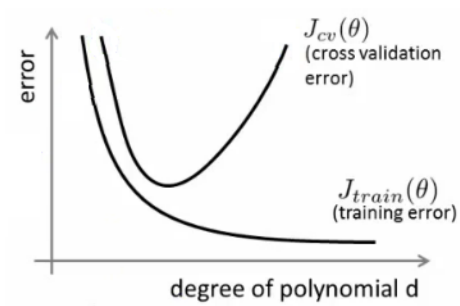
Bad results are generally the consequence of one of the following:

- High bias - under fitting problem.
- High variance - over fitting problem.



Now plot

- x = degree of polynomial d
- y = error for both training and cross validation (two lines)



- For the high bias case, we find both cross validation and training error are high
- For high variance, we find the cross validation error is high but training error is low

Regularization and bias/variance

Linear regression with regularization:

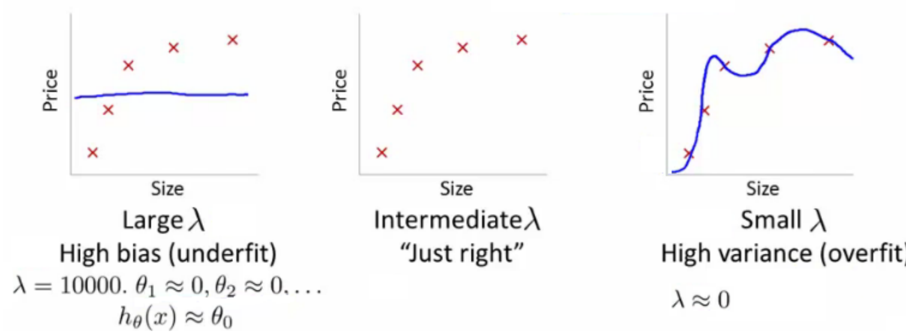
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

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

The above equation describes the fitting of a high order polynomial with regularization (used to keep parameter values small).

1. λ is large (high bias $->$ under fitting data)

2. λ is intermediate (good)
3. λ is small (high variance $->$ overfitting)

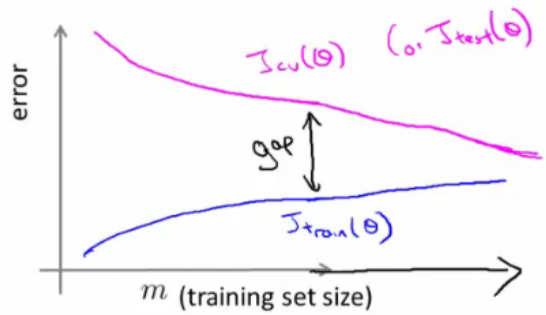


- Have a set or range of values to use (for example from 0 to 15).
- For each λ_i minimize the cost function. Result is $\theta^{(i)}$.
- For each $\theta^{(i)}$ measure average squared error on cross validation set.
- Pick the model which gives the lowest error.

Learning curves

Plot J_{train} (average squared error on training set) and J_{cv} (average squared error on cross validation set) against m (number of training examples).

- J_{train} on smaller sample sizes is smaller (as less variance to accommodate).
- As training set grows your hypothesis generalize better and J_{cv} gets smaller.



- A small gap between training error and cross validation error might indicate high bias. Here, more data will not help.
- A large gap between training error and cross validation error might indicate high variance. Here, more data will probably help.