Note: Regularization

Sun Zhao

January 11, 2013

1 Over-fitting

Over-fitting problem occurs when the learned hypothesis fits the training data set very well, but fails to generalize to new examples. It is mainly coursed by the trained model is excessively complex, such as having too many parameters relative to the number of observations. Fig. 1 shows three linear regression hypothesises for the house prices predicting problem. The hypothesis shown in Fig. 1A is called "under-fitting" which means it is quite bias from the right one. Fig. 1C is an example of over-fitting which has high variance. Though it predicts the prices perfectly for every examples in training data, it can not be generalized to new input. Moreover, Fig. 2 shows "under-fitting", "right one", "over-fitting" logistic regression hypothesis separately.

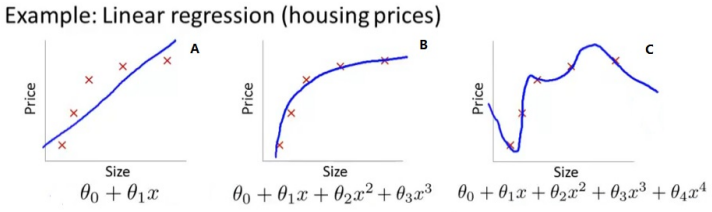

Figure 1:

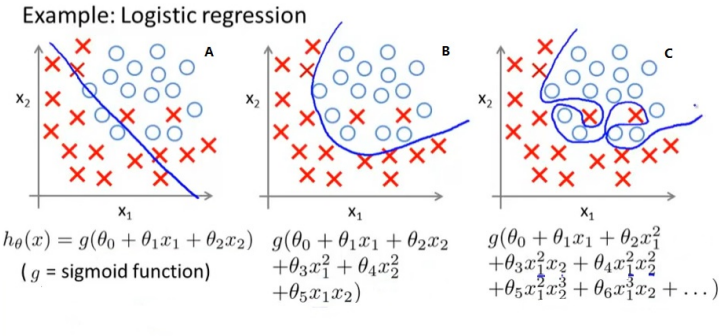

Figure 2:

2 Regularized Cost Function

The intuition of regularization is to penalize large parameters of θ and keep the hypothesis simple. In Fig. 1C, if we can penalize θ_4 and shrink it to zero, the hypothesis will much closer to the "right" one. Regularization adds a term of

$\frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$ to the regression cost function. So, the new cost function is as (1).

$$J^{'}(\theta) = J(\theta) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2 \tag{1}$$

The larger value of θ_j, the larger cost of the corresponding hypothesis, hence the gradient descent algorithm will penalize large θ_j. Note j starts from 1 instead of 0, it is because θ_0 is the const value of the hypothesis function and does not influence the complexity.

3 Regularized Gradient Descent

Calculating the partial derivatives of (1) for each of θ_j, we can infer the regularized gradient descent for regression problem. The only difference of regularized gradient descent for linear regression and logistic regression is the hypothesis function.

Algorithm1: Regularized Gradient Descent for Regression Problem

Repeat{

$\theta_0 = \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_0^{(i)}$

$\theta_j = \theta_j(1 - \frac{\lambda}{\alpha m}) - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})\cdot x_j^{(i)}$ (j=1 ... n)

}

4 Regularized Normal Equation

Normal equation is an alternative for solving linear regression problem. All θs can be calculated by a system of linear equations. When using the new regularized cost function shown in (1), we have the regularized normal equation shown in (2).

$$\theta = (X^T X + \lambda(I_n - e_0))^{-1}X^T Y \tag{2}$$

In (2), I_n is n × n dimensional elementary matrix and e_0 equals $[1 \quad 0\ldots 0]^T_{1\times n}$.

5 Summary

Over-fitting is a common problem when using machine learning algorithms. Regularization penalizes large θs and keep the hypothesis simple to overcome this problem. In practice, regularization produces good outcome.