Note: Linear Regression with Multiple Variables

Sun Zhao

December 21, 2012

1 Multiple Features

Univariate linear regression takes only one variable to predict values. Mostly, values are relevant to multiple features other than one feature. Multiple features are powerful than single feature to predict values. An example of multiple features are shown in Table1. A house's price is now related to the size, number of bedrooms, number of floors, and age of it. Each row is an training example with left three feature columns and right one value column. We use n to denote the number of features, $x^{(i)}$ to denote the features of the i^{th} training example and $x_j^{(i)}$ to denote the value of feature j in i^{th} training example. For example, $x^{(2)}$ equals $[1416 \quad 3 \quad 2 \quad 40]^T$ and $x_3^{(2)}$ equals 2. Since the hypothesis of univariate linear regression is $h_\theta(x) = \theta_0 + \theta_1 x$, that of multivariate linear regression becomes formula1.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n \tag{1}$$

Let $x_0^{(i)} = 1$, $\theta = [\theta_0 \quad \theta_1 \ldots \theta_n]^T$, and $x = [x_0 \quad x_1 \quad \ldots x_n]^T$, then formula1 is refined as formula2.

$$h_\theta(x) = \theta^T x \tag{2}$$

2 Gradient Descent

Refer to formula1 we can infer the cost function of multivariate linear regression shown in formula3.

$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2 \tag{3}$$

And the gradient descent algorithm is adapted to:

```
    Algorithm1: Gradient for Multivariate Linear Regression
    Repeat{
$\theta_j = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$
}
```

3 Practical Issues

3.1 Feature Scaling

The idea of feature scaling is to make sure that all features are on similar scale. If one feature's range is significantly dominating the others', the regression process will converge at a very slow speed. Generally, make the scale of all features is around [-1, 1] is a good choice. Methods for normalization includes dividing values by its ranges or first decreasing by its mean values and then dividing by its range. In a word, we can normalize x form [min, max] with x/(max-min) or (x-average([min, max]))/(max-min).

3.2 Learning Rate

The core function of gradient descent is $\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$. When running gradient descent algorithm, we need to choose an appropriate α and make sure the algorithm is working correctly. Teacher recommended to plot the value of $J(\theta)$ against the number of regression iterations. If the plotted curve decreases faster as Figure1(A), then we can confirm α and the algorithm all works well. If the plotted curve decreases slow as Figure1(A), we need to increase the value of α a little more. Otherwise, the plotted curve increase as Figure1(C) or vibrates as Figure1(D), we have to decrease the value of α.

3.3 Choosing Features