Machine Learning Week-3

ramay7

February 20, 2017

Contents

1 Classification and Representation

1.1 Hypothesis Representation

Want $0 \leq h_\theta(x) \leq 1$:

$$h_\theta(x) = g(X\theta)$$

where $g(z) = \frac{1}{1+e^{-z}}$ is called **sigmoid/logistic function**. $h_\theta(x)$ will give us the **probability that our output is 1**.

1.2 Decision Boundary

For:

$$g(X\theta) = \frac{1}{1+e^{-X\theta}}$$

assume $X\theta = 0$, and then get an equation.

2 Logistic Regression Model

2.1 Cost Function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} CostFunction(h_\theta(x^{(i)}), y^{(i)})$$

For linear regression:

$$Cost(h_\theta(x), y) = \frac{1}{2}(h_\theta(x) - y)^2$$

For logistic regression:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if y = 1} \\ -\log(1 - h_\theta(x)) & \text{if y = 0} \end{cases} \tag{1}$$

This form of $Cost(h_\theta(x), y)$ guarantees that $J(\theta)$ is **convex** for logistic regression.
We can change the cost function of logistic regression into another form:

$$Cost(h_\theta(x), y) = -y\log(h_\theta(x)) - (1 - y)\log(1 - h_\theta(x))$$

Hence, we can get:

$$J(\theta) = \frac{1}{m}(-y^T \log(h) - (1 - y)^T \log(1 - h))$$

To make it concrete:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^{(i)}\log(h_\theta(x^{(i)})) + (1 - y^i)\log(1 - h_\theta(x^i)) \tag{2}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \tag{3}$$

2.2 A Vectorized Implementation for Gradient Descent

$$\theta := \theta - \frac{\alpha}{m}X^T(g(X\theta) - y)$$

3 Multiclass Classfication: One-vs-all

Train a logistic regression classifier $h_\theta^{(i)}$ for each class i to predict the probability that $y = 1$.
On a new input x to make a prediction, pick the class i that maximizes $\max_i h_\theta^{(i)}(x)$

4 Solving the Problem of Overfitting

4.1 The Problem of Overfitting

Underfitting is when the form of our hypothesis function h maps poorly to the trend of the data. It is usually caused by a function that is too simple or uses too few features. As the other extremes, overfitting is caused by a hypothesis function that fits the available data but does not generalize well to predict new data. It is usually caused by a complicated function that creates a lot of unnecessary curves and angles unrelated to the data.

This terminology is applied to both linear and logistic regression. There are two main options to adress the issue of overfitting:

- Reduce the number of features:
  - Manually select which features to keep
  - Use a model selection algorithm
- Regularization
  - Keep all the features, but reduce the magnitude of parameters θ_j
  - Regularization works well when we have a lot of slightly useful features

4.2 Cost Function in Regularization

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

where λ is called **regularization parameter**.

Please take care that the index of the quadratic sum of θ starts from 1. That is to say, ignore θ_0^2.

If λ is set to an extremely large value (perhaps for too large for our problem, say $\lambda = 10^{10}$), it may smooth out the function too much and cause underfitting.

4.3 Regularized Linear Regression

The Gradient Descent is: **repeat until convergence:{**

$$\theta_0 := \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)} \tag{4}$$

$$\theta_j := \theta_j - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} + \lambda\theta_j \qquad for\ j \geq 1 \tag{5}$$

}

4.4 Normal Equation

$$\theta = (X^T X + \lambda L)^{-1} X^T y$$

where L is a $(n+1)$ by $(n+1)$ matirx:

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

If $m \leq n$, then $X^T X$ is non-invertible. However, when we add the term L, then $X^T X + \lambda L$ becomes invertiable.

4.5 Regularized Logistic Regression

Cost Function:

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

Please take care that the index of the quadratic sum of θ starts from 1. That is to say, ignore θ_0^2.

Gradient descent is the same with regularized linear regression except that $h_\theta(X) = \frac{1}{1+e^{-\theta^T X}}$.

$$\frac{\partial J(\theta)}{\partial \theta_0} := \theta_0 - \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_0^{(i)} \tag{6}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} := \left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j \qquad for\ j \geq 1 \tag{7}$$

5 The End

The most wonderful job is getting full marks for Exercise-2.