Notes on UFLDL Tutorial from Stanford University

June 29, 2015

The content can be found in http://deeplearning.stanford.edu/tutorial/.

1 Linear regression

1. Linear function: $h_\theta(x) = \sum_j \theta_j x_j$
2. The cost function: $J(\theta) = \frac{1}{2}\sum_i (h_\theta(x^{(i)}) - y^{(i)})^2$
3. Gradient descent is used to minimize the cost function
4. The gradient of the cost function $\nabla_\theta J(\theta)$ is the differential of function $J(\theta)$ and it is a vector that points in the direction of steepest increase as a function of θ.

$$\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)}(h_\theta(x^{(i)} - y^{(i)}))$$

2 Logistic regression

1. Logistic regression is a simple classification algorithm.

$$P(y=1|x) = h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$$

$$P(y=0|x) = 1 = P(y=1|x) = 1 - h_\theta(x) = 1 - \frac{1}{1+\exp(-\theta^T x)}$$

2. $\frac{1}{1+\exp(-Z)}$ is called the sigmoid or logistic function. The result of this funciton is a S-shape function ranges from [0,1].

3. The cost function is defined as $J(\theta) = -\sum_i (y^{(i)} log(h_\theta(x^{(i)}) + (1-y^{(i)}) log(1-h_\theta(x^{(i)}))$

4. The gradient of the cost function $\nabla_\theta J(\theta) = \sum_i x^{(i)}(h_\theta(x^{(i)}) - y^{(i)})$ and $\frac{\partial J(\theta)}{\partial \theta_j} = \sum_i x_j^{(i)}(h_\theta(x^{(i)}) - y^{(i)})$

3 Gradient checking

1. This is can be used to check whether the gradient is comuted correctly.

2. Given a function $g(\theta)$ which computes $\frac{dJ(\theta)}{d\theta}$, the value can be checked by $g(\theta) \approx \frac{J(\theta+\epsilon)-J(\theta-\epsilon)}{2\epsilon}$, where ϵ is a small value, e.g., 10^{-4}.

3. A more general version is that considering $\theta \in R^n$ (i.e., θ is a vector). Note that $\overrightarrow{e_j}$ is a vector of 0 which has a length equal to θ. The j^{th} element of $\overrightarrow{e_j}$ is 1.

$$g_j(\theta) \approx \frac{J(\theta^{j+}) - J(\theta^{j-})}{2\epsilon}$$
$$\theta^{j+} = \theta + \epsilon \times \overrightarrow{e_j}$$

4 Softmax regression

1. Softmax regression, or multinomial logistic regression, is a generalization of logistic regression for cases of multiple classes. It allows us to handle $y^{(i)} \in \{1, 2, \cdots, K\}$, where K is the number of classes.

2. Hypothesis in logistic regression

   - $h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$ is the probability of $y^{(i)}$ being 1.
   - Cost function is $J(\theta) = -\sum_{i=1}^{m}(y^{(i)} log(h_\theta(x^{(i)})) + (1-y^{(i)}) log(1-h_\theta(x^{(i)}))$

3. In softmax regression, there are K possible classes, hence, the output of $h_\theta(x)$ is a K-dimensional vector, whose elements sum up to 1.

   - $h_\theta(x) = \frac{1}{\sum_i^K \exp(\theta^{(j)T}x)}[\exp(\theta^{(1)T}x), \exp(\theta^{(2)T}x), \cdots, \exp(\theta^{(K)T}x)]$

- Note that $\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(K)}$ are the parameters of the model. The term $\sum_i^K \exp(\theta^{(j)T} x)$ normalizes the distribution so that it sums to one.
- Cost function is $J(\theta) = -\sum_{i=1}^m \sum_{k=1}^K 1\{y^{(i)} = k\} log(\frac{\exp(\theta^{(k)T} x^{(i)})}{\sum_{j=1}^K \exp(\theta^{(j)T} x^{(i)})})$, where $1\{true\, argument\} = 1$ and $1\{false\, argument\} = 0$