(1) Training Data Set

$D = \{(\vec{x}^{(i)}, y^{(i)}), i = 1, 2, \ldots\ldots, m\}$

x = Number of features

The Cost function for stochastic gradient descent.

$J_{train}(\vec{\theta}) = 1/m \sum_{i=1}^{m} cost(\vec{\theta}, x^{(i)}, y^{(i)})$

where the $cost(\vec{\theta}, (x^{(i)}, y^{(i)}))$ is the cost associated with i-th training example.

$$cost(\vec{\theta}, (x^{(i)}, y^{(i)})) = 1/2 (h_\theta(x^{(i)}) - (y^{(i)}))^2$$

$$= 1/2 [\vec{\theta}^T x^{(i)} - y^{(i)}]^2$$

$$= 1/2 [\theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_x (x, x^{(i)})' - y^{(i)}]^2$$

The update rule for the stochastic gradient descent,

$i = 1$ to m

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} cost(\vec{\theta}, (x^{(i)}, y^{(i)}))$$

which turns out

$$\frac{\partial}{\partial \theta_j} = cost(\theta, (x^{(i)}, y^{(i)}))$$

$$= \frac{\partial}{\partial \theta_j} \cdot 1/2 [(\theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_j x_j^{(i)} + \theta_x x_x^{(i)} - y^{(i)}]^2$$

$$= 1/2 \times 2 [\theta_0 + \theta_1 x_1^{(i)} + \ldots + \theta_x x_x^{(i)} - y^{(i)}] x_j^{(i)}$$

1 continued

Stochastic Gradient Descent update rule,

for $i = 1$ to m

$\theta_j = \theta_j - \alpha(\theta_0 + \theta_1 x_1^{(i)} + \dots + \theta_n x_n^{(i)} - y) x_j^{(i)}$

for $i = 1$ to m

$\theta_j = \theta_j - \alpha(\vec{\theta}^T x^{(i)} - y^{(i)}) x_j^{(i)}$

(2)

First, we need to be clear about the probabilistic assumptions you make. When using the results of the analysis, we must be aware that the results may not be relevant in cases where the assumptions do not hold.

Next, we must use the maximum likelihood estimation formula which is defined as x1, x2, ..., xn will be observations of n independent and identically distributed random variables, drawn from a probability Distribution f_0f0, when f_0f0 is known to be from a distributions ff that depend on some parameters θ.

Lets Assume that the target variables and the inputs are related using this equation. $y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$. where $\varepsilon^{(i)}$ is an error term that captures unmodeled effects or random noise. We can then further assume that $\varepsilon^{(i)}$ are distributed IID according to Gaussian distributions with mean zero and some variance σ^2. We then write this assumption as $\varepsilon^{(i)} \sim N(0, \sigma^2)$

$$P(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$$ which implies

$$p(y^{(i)} | x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$p(y^{(i)} | x^{(i)} \theta)$ indicates that $y^{(i)}$ given $x^{(i)}$ is distribution and parameterized by θ. You may also write the distribution of $y^{(i)}$ as $y^{(i)}; \theta \sim N(\theta^T x^{(i)}, \sigma^2)$.

Given X and θ, the probability of the data given by $P(\vec{y} | X; \theta)$. This quantity is typically viewed a function of $\vec{y}$ for a fixed value of θ. When you wish to view this as a function of θ, we will call it a likelihood function.

2 continue

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y} \mid X; \theta)$$

The independent assumption on the $\varepsilon^{(i)}$ can be written as

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

The principle of maximum likelihood says that we should choose θ to make the data as high probability as possible. Instead of maximizing $L(\theta)$, we can also maximize any strictly increasing function of $L(\theta)$. In particular, the derivations will be simpler if we maximize the log likelihood $\ell(\theta)$

$$\ell(\theta) = \log L(\theta) = \log \sum_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$= m \log \frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{\sigma^2} \times \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

Maximizing $\ell(\theta)$ gives the same answer as minimizing $\frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$

Maximizing Likelihood Estimation isn't necessarily optimal, due to other estimation algorithms that achieve better results, it does have its own attractive properties. The most important being consistency. A sequence of Maximizing likelihood estimation on an increasing number of operations will converge to the true value of the parameters.

Using the above two definitions we can say $J\theta$ is a reasonable choice when using θ.

(3)

$$\log h_\theta(x^i) = \log \frac{1}{1+e^{-\theta x^i}} = -\log(1+e^{-\theta x^i})$$

$$\log(1-h_\theta(x^i)) = \log(1-\frac{1}{1+e^{-\theta x^i}})$$

$$= \log(e^{-\theta x^i}) - \log(1+e^{-\theta x^i})$$

$$= -\theta x^i - \log(1+e^{-\theta x^i})$$

The original cost function is in the form of

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

We then plug in the two simplified expressions

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^i(\log(1+e^{-\theta x^i})) + (1-y^i)(-\theta x^i - \log(1+e^{-\theta x^i}))]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}[y_i\theta x^i - \theta x^i - \log(1+e^{-\theta x^i})]$$

$$= -\frac{1}{m}\sum_{i=1}^{m}[y_i\theta x^i - \log(1+e^{\theta x^i})]$$

The second equality

$$-\theta x^i - \log(1+e^{-\theta x^i})$$

$$= -[\log e^{\theta x^i} + \log(1+e^{-\theta x^i})]$$

$$= -\log(1+e^{\theta x^i})$$

And to compute the partial derivatives

$$\frac{\partial}{\partial\theta_j} y_i\theta x^i = y_i x_j^i$$

$$\frac{\partial}{\partial\theta_i}\log(1+e^{\theta x^i}) = \frac{x_j^i e^{\theta x^i}}{1+e^{\theta x^i}} = x_j^i h_\theta(x^i)$$