Linear Regression

Linear Regression

- Minimize the cost function J(θ)

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

- Error term

$$err = y^{(i)} - h_\theta\left(x^{(i)}\right)$$

- Minimize the error

$$repeat\ until\ convergence\{$$

$$\theta_j \leftarrow \theta_j + \eta\sum_{i=1}^{m}\left(y^{(i)} - h_\theta\left(x^{(i)}\right)\right)x_j^{(i)} \quad (\text{for every } j)$$

$$\}$$

Probabilistic interpretation

- Why might the least-squares cost function J be a reasonable choice ?

- Let us assume that

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

  - ✓ where $\epsilon^{(i)}$ is an error term or random noise
  - ✓ $\epsilon^{(i)}$ are distributed independently and identically according to a Gaussian distribution (i.e., Normal distribution)

$$\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$$

- The density of $\epsilon^{(i)}$

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(\epsilon^{(i)}\right)^2}{2\sigma^2}\right)$$

Probabilistic interpretation

- This implies that

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

- $p\left(y^{(i)}|x^{(i)};\theta\right)$ indicates that this is the distribution of y given x and parameterized by θ
  - ✓ Since θ is not a random variable, we should not condition on θ
  - ✓ The probability of the data is given by $p\left(y^{(i)}|x^{(i)};\theta\right)$
  - ✓ This quantity is typically viewed a function of y and x for a fixed value of θ
  - ➔ Change to a function of θ

- Likelihood function

$$\mathcal{L}(\theta) = \mathcal{L}(\theta;\mathrm{X},\vec{y}) = p(\vec{y}|\mathrm{X};\theta)$$

Probabilistic interpretation

- Likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)};\theta\right)$$

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

- Maximum likelihood
  - ✓ We should choose θ so as to make the data as high probability as possible
  - ✓ i.e., we should choose θ to maximize $\mathcal{L}(\theta)$

Probabilistic interpretation

- Log likelihood $\ell(\theta)$

$$\ell(\theta) = log\mathcal{L}(\theta)$$

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

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

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

- Maximizing $\ell(\theta)$

$$minimizing \ \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

Logistic Regression

Classification Problem

- Binary classification
  - ✓ y can take on only two values, 0 and 1
  - ✓ y is called the label for the training example

- Logistic regression
  - ✓ Hypotheses is called the logistic function or the sigmoid function

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

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

$$g'^{(z)} = \frac{d}{dz}\left(\frac{1}{1 + e^{-z}}\right) = \frac{1}{(1 + e^{-z})^2}(e^{-z})$$

$$= \frac{1}{(1 + e^{-z})} \cdot \left(1 - \frac{1}{(1 + e^{-z})}\right) = g(z)(1 - g(z))$$

Classification Problem

- Given the logistic regression model, how do we fit θ?

  ✓ With a set of probabilistic assumption
  ✓ Fit the parameters via maximum likelihood

  ✓ Let us assume that

$$p(y = 1|x; \theta) = h_\theta(x)$$
$$p(y = 0|x; \theta) = 1 - h_\theta(x)$$

  ✓ Then,

$$p(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

Classification Problem

- Assuming that the m training examples were generated independently
- Then, we can write down the likelihood of the parameters

$$\mathcal{L}(\theta) = p(\vec{y}|X;\theta)$$

$$= \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)};\theta\right)$$

$$= \prod_{i=1}^{m} (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

- As before, it will be easier to maximize the log likelihood

$$\ell(\theta) = log\mathcal{L}(\theta)$$

$$= \prod_{i=1}^{m} (h_\theta(x))^y (1-h_\theta(x))^{1-y} = \sum_{i=1}^{m} y^{(i)} log h\left(x^{(i)}\right) + \left(1-y^{(i)}\right) log\left(1-h\left(x^{(i)}\right)\right)$$

Classification Problem

- How do we maximize the likelihood? ➔ use gradient ascent
- Gradient ascent

$$\theta \coloneqq \theta + \alpha \nabla_\theta \ell(\theta)$$

- For one training example:

$$\begin{aligned}
\frac{d}{d\theta_j}\ell(\theta) &= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)\frac{d}{d\theta_j}g(\theta^T x) \\
&= \left(y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}\right)g(\theta^T x)\big(1-g(\theta^T x)\big)\frac{d}{d\theta_j}(\theta^T x) \\
&= \Big(y\big(1-g(\theta^T x)\big) - (1-y)g(\theta^T x)\Big)x_j \\
&= \big(y - h_\theta(x)\big)x_j
\end{aligned}$$

Classification Problem

- Stochastic gradient ascent

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta \left(x^{(i)} \right) \right) x_j^{(i)}$$

- We see that it looks identical; but this is not the same algorithm
- Because $h_\theta \left(x^{(i)} \right)$ is now defined as a non-linear function of $\theta^T x$

E N D