Machine Learning

Logistic Regression

1. Model

logistic function:

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

hypothesis:

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

maximum likelihood:

$$\begin{aligned} P(y = 1|x;\theta) &= h_\theta(x) \\ P(y = 0|x;\theta) &= 1 - h_\theta(x) \\ P(y \mid x;\theta) &= (h_\theta(x))^y (1 - h_\theta(x))^{1-y} \end{aligned}$$

input:

$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)}) \text{ where } y^{(i)} \in \{0, 1\}$$

cost function:

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

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

derivative:

$$
\begin{aligned}
\frac{\partial}{\partial\theta_j}\ell(\theta) &= \sum_{i=1}^{m}\left(y^{(i)}\frac{1}{h_\theta(x^{(i)})} - (1-y^{(i)})\frac{1}{1-h_\theta(x^{(i)})}\right)\frac{\partial}{\partial\theta_j}h_\theta(x^{(i)}) \\
&= \sum_{i=1}^{m}\left(y^{(i)}(1-h_\theta(x^{(i)})) - (1-y^{(i)})h_\theta(x^{(i)})\right)x_j^{(i)} \\
&= \sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}
\end{aligned}
$$

gradient ascent:

$$
\begin{aligned}
\theta_j :&= \theta_j + \alpha\frac{\partial}{\partial\theta_j}\ell(\theta) \\
&= \theta_j + \alpha\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)}
\end{aligned}
$$

stochastic gradient ascent:

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

Newton's method:

$$
\theta := \theta - H^{-1}\nabla_\theta\ell(\theta)
$$

where H is Hessian matrix, and $H_{ij} = \frac{\partial^2\ell(\theta)}{\partial\theta_i\partial\theta_j}$.

2. Softmax Regression

input:

$$
(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)}) \text{ where } y^{(i)} \in \{1, \ldots, K\}
$$

hypothesis:

$$
h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K}\exp(\theta^{(j)\top}x)}\begin{bmatrix} \exp(\theta^{(1)\top}x) \\ \exp(\theta^{(2)\top}x) \\ \vdots \\ \exp(\theta^{(K)\top}x) \end{bmatrix}
$$

cost function:

$$L(\theta) = \prod_{i=1}^{m} \sum_{k=1}^{K} 1\left\{y^{(i)} = k\right\} \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})}$$

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{K} 1\left\{y^{(i)} = k\right\} \log \frac{\exp(\theta^{(k)\top} x^{(i)})}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x^{(i)})}$$

derivative:

$$\nabla_{\theta^{(k)}} \ell(\theta) = \sum_{i=1}^{m} \left[x^{(i)} \left(1\{y^{(i)} = k\} - P(y^{(i)} = k|x^{(i)}; \theta)\right)\right]$$

gradient ascent:

$$\begin{aligned} \theta^{(k)} &:= \theta^{(k)} + \alpha \nabla_{\theta^{(k)}} \ell(\theta) \\ &= \theta^{(k)} + \alpha \sum_{i=1}^{m} \left[x^{(i)} \left(1\{y^{(i)} = k\} - P(y^{(i)} = k|x^{(i)}; \theta)\right)\right] \end{aligned}$$

stochastic gradient ascent:

$$\theta^{(k)} := \theta^{(k)} + \alpha \left[x^{(i)} \left(1\{y^{(i)} = k\} - P(y^{(i)} = k|x^{(i)}; \theta)\right)\right]$$

3. Concavity

$$\begin{aligned} \ell(\theta) = \log L(\theta) &= \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)})) \\ &= \sum_{i=1}^{m} y^{(i)} \cdot (\theta^T x^{(i)}) - \log(1 + \exp(\theta^T x^{(i)})) \end{aligned}$$

Because $f(\theta) = y^{(i)} \cdot (\theta^T x^{(i)})$ is concave, and $f(\theta) = \log(1 + \exp(\theta^T x^{(i)}))$ is also concave, so $\ell(\theta)$ is concave.