linear models

prashant ullegaddi
vidit bansal

July 12, 2013

linear regression

What's the problem?

- Given a set of m training examples $\{X^{(i)}, y^{(i)}\}_{i=1}^{m}$ where $X^{(i)} \in \mathbb{R}^{n+1}$ with $X_0^{(i)} = 1$ and $y^{(i)} \in \mathbb{R}$.
- Learn a hypothesis function $h(\cdot)$ such that for any new test example $x^{(t)}$, y can be predicted as $y^{(t)} = h(x^{(t)})$.
- Fit a hyperplane
  $h_\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$

ordinary least squares method

Optimization

Find parameters θ that:

$$\text{minimize}_\theta J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(X^{(i)}) - y^{(i)})^2$$

gradient descent method

Update rule

$$\theta := \theta - \alpha \nabla_\theta J(\theta)$$

where α is the learning rate, $\nabla_\theta J(\theta)$ is the gradient of cost function $J(\theta)$.

derivation of gradient

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

$$\frac{\partial}{\partial \theta_j} J(\theta) = \sum_{i=1}^{m} (h_\theta(X^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_j} (\theta^T X^{(i)} - y^{(i)})$$

Gradient

$$\frac{\partial}{\partial \theta_j} J(\theta) = \sum_{i=1}^{m} (h_\theta(X^{(i)}) - y^{(i)}) X_j^{(i)}$$

$$\nabla_\theta J(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_0} J(\theta) & \frac{\partial}{\partial \theta_1} J(\theta) & \cdots & \frac{\partial}{\partial \theta_n} J(\theta) \end{bmatrix}^T$$

stochastic gradient method

Data: α, m training examples $\{X^{(i)}, y^{(i)}\}_{i=1}^{m}$
Result: θ parameters
$\theta := \mathbf{0}$;
for $i = 1 \cdots m$ **do**
    for $j = 0 \cdots n$ **do**
        $\theta := \theta - \alpha(h_\theta(X^{(i)}) - y^{(i)})X_j^{(i)}$;
    end
end

Algorithm 1: Stochastic gradient algorithm

closed form: add derivation involving trace

closed form solution

Solution

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

where X is an $m \times (n+1)$ data matrix with $X[:, 1] = 1$ and y is a $m \times 1$ vector of labels

logistic regression

What's the problem?

- Given a set of m training examples $\{X^{(i)}, y^{(i)}\}_{i=1}^{m}$ where $X^{(i)} \in \mathbb{R}^{n+1}$ with $X_0^{(i)} = 1$ and $y^{(i)} \in \{0, 1\}$ is a label.

- Learn a hypothesis function $h(\cdot)$ such that for any new test example $x^{(t)}$, label for it can be predicted as $y^{(t)} = h(x^{(t)})$.

- In logistic regression, we assume:

$$h(x) = \text{sigmoid}(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

property of sigmoid function

$$
\begin{aligned}
\text{Let } g(x) &= \frac{1}{1 + e^{-x}} \\
\text{Then, } g'(x) &= -\frac{1}{(1 + e^{-x})^2} \frac{d}{dx}(1 + e^{-x}) \\
&= \frac{e^{-x}}{(1 + e^{-x})^2} \\
&= \left(\frac{1}{1 + e^{-x}} \right) \left(\frac{e^{-x}}{1 + e^{-x}} \right) \\
&= \left(\frac{1}{1 + e^{-x}} \right) \left(1 - \frac{1}{1 + e^{-x}} \right)
\end{aligned}
$$

Thus,

$$
g'(x) = g(x)(1 - g(x)) \tag{1}
$$

optimization—maximize likelihood of the parameters θ

Assume

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

or more generally

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

Likelihood of parameters

$$L(\theta) = p(\mathbf{y}|X; \theta) = \prod_{i=1}^{m} p(y^{(i)}|X^{(i)}; \theta)$$

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

or equivalently maximize loglikelihood of the parameters θ

It's easier to maximize loglikelihood $\log(L(\theta))$ which is the same as maximizing likelihood $L(\theta)$

Optimization

$$
\begin{aligned}
l(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} [y^{(i)} \log h_\theta(X^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(X^{(i)}))] \\
\max_\theta &\, l(\theta)
\end{aligned}
$$

Gradient ascent

Update rule:

$$
\theta := \theta + \alpha \nabla_\theta l(\theta)
$$

derivation of gradient of $l(\theta)$

$$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^{m} [y^{(i)} \log h_\theta(X^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(X^{(i)}))]$$

$$= \sum_{i=1}^{m} [y^{(i)} \frac{1}{h_\theta(X^{(i)})} \frac{\partial}{\partial \theta_j} h_\theta(X^{(i)})$$

$$+ (1 - y^{(i)}) \frac{1}{(1 - h_\theta(X^{(i)}))} \frac{\partial}{\partial \theta_j} (1 - h_\theta(X^{(i)}))]$$

$$= \sum_{i=1}^{m} [y^{(i)} \frac{1}{h_\theta(X^{(i)})} h'_\theta(X^{(i)}) \frac{\partial}{\partial \theta_j} \theta^T X^{(i)}$$

$$+ (1 - y^{(i)}) \frac{1}{(1 - h_\theta(X^{(i)}))} (-h'_\theta(X^{(i)})) \frac{\partial}{\partial \theta_j} \theta^T X^{(i)}]$$

derivation of gradient of $l(\theta)$

Using Eq. 1,

$$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^{m} [y^{(i)} \frac{1}{h_\theta(X^{(i)})} h_\theta(X^{(i)})(1 - h_\theta(X^{(i)}) \frac{\partial}{\partial \theta_j} \theta^T X^{(i)}$$

$$+ (1 - y^{(i)}) \frac{1}{(1 - h_\theta(X^{(i)}))} (-h_\theta(X^{(i)})(1 - h_\theta(X^{(i)}))) \frac{\partial}{\partial \theta_j} \theta^T X^{(i)}]$$

$$= \sum_{i=1}^{m} [y^{(i)}(1 - h_\theta(X^{(i)}) X_j^{(i)} + (1 - y^{(i)})(-h_\theta(X^{(i)}) X_j^{(i)}]$$

Derviative

$$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^{m} (y^{(i)} - h_\theta(X^{(i)})) X_j^{(i)}$$