Linear Regression
Logistic Regression

M.Stepanov

2 ноября 2012 г.

Linear Regression

Living area (sq. feet)	bedrooms	Price (1000)
2104	3	400
1600	3	330
2400	3	369
1416	2	232
3000	4	540

- Objective is to approximate dataset by some linear function.

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

- More generally for arbitrary dataset we are looking for

$$h_\theta = \sum_{i=0}^{n} \theta_i x_i = \theta^t x$$

- we assume that $x_0 = 1$ - intercept term

Cost Function and Gradient Descent Algorithm

- Least-squares cost function

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

- Gradient descent algorithm starts with some initial θ and repeatedly performs update

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

- Gradient descent
  Repeat until convergence {

$$\theta_j = \theta_j - \alpha \sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \text{ (for every j) .}$$

}

- Stohastic gradient descent
  Loop{
      for i = 1 to m{

$$\theta_j = \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \text{ (for every j) .}$$

      }
  }

Normal Equations

- Least-squares cost function

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

- In matrix form could be rewrited as

$$J(\theta) = \frac{1}{2}(X\theta - \bar{y})(X\theta - \bar{y})$$

- By taken gradient from cost function in the matrix form we get

$$\theta = (X^tX)^{-1}X^t\bar{y}$$

Probabilistic Interpretation

- Let assume that the target variables and the inputs are related via the equation
$$y^{(i)} = \theta^t x^{(i)} + \epsilon^{(i)}$$
- $\epsilon^{(i)}$ is an error term that independently and identically distributed according to a Gaussian distribution ($\mathcal{N}(0, \sigma^2)$)

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

- We can rewrite this as a conditional distribution

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

Probabilistic Interpretation

- To estimate θ lets use likelihood function

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

  since $\epsilon^{(i)}$ is i.i.d.
- The likelihood function is maximaized if

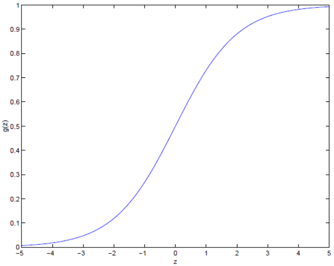
$$\frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \to \min$$

Logistic Regression

- For classification we want $0 \leq h_\theta(x) \leq 1$ since our $y \in \{0, 1\}$
- The natural $h_\theta(x)$ choise is logistic(sigmoid) function

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

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



Logistic Regression Cost Function

- Let us assume that

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

- This can written more compactly as

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

- Likelihood function could be written as

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

- This function could be rewritten as logliklihood function

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

Gradient ascent

- In matrix form
$$\theta := \theta + \alpha \nabla_\theta l(\theta)$$
- Let us start with just one training example (x, y) and take derivates to derive stochastic gradient ascent rule
$$\frac{\partial l(\theta)}{\partial \theta_j} = (y - h_\theta(x))x_j$$
- Stohastic gradient ascent
Loop{
    for i = 1 to m{
$$\theta_j = \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \text{ (for every j)} .$$
    }
}

Bayesian statistics and regularization

- Recently we viewed θ as an unknown parameter and estimate it using maximum likelihood

$$\theta_{ML} = \text{argmax}_\theta \prod_{i=1}^{n} p(y^{(i)}|x^{(i)};\theta)$$

- Lets think of θ as being a random variable distributed by some prior distribution $p(\theta)$
- Given a training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{m}$ lets compute posterior

$$p(\theta|S) = \frac{P(S|\theta)P(\theta)}{p(S)} = \frac{(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(\theta)}{\int_\theta(\prod_{i=1}^{m} p(y^{(i)}|x^{(i)},\theta))p(\theta)d\theta}$$

  - For logistic regression $p(y^{(i)}|x^{(i)},\theta) = (h_\theta(x))^y(1-h_\theta(x))^{1-y}$
- In general it is very hard to estimate $p(\theta|S)$ over θ
- In practice

$$\theta_{MAP} = \text{argmax}_\theta \prod_{i=1}^{m} p(y^{(x_i)}|x^{(y_i)},\theta)p(\theta)$$

- Common choice $\theta(0, \lambda\mathtt{I})$, the norm of θ usually less then that selected by ML

Regularized Linear Regression

- Least-squares cost function

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

- Gradient descent

  Repeat until convergence {

$$\theta_j = \theta_j - \alpha\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_0^{(i)} \text{ (j = 0) }.$$

$$\theta_j = \theta_j - \alpha\sum_{i=1}^{m}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} - \frac{\lambda}{m}\theta_j \text{ (j} \geq \text{1) }.$$

  }

Regularized Logistic Regression

- Regularized cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

- Gradient ascent

Loop{

for i = 1 to m{

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{n}(y^{(i)} - h_\theta(x^{(i)}))x_0^{(i)} \text{ (for j = 0) .}$$

$$\theta_j = \theta_j + \alpha \sum_{i=1}^{n}(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} + \frac{\lambda}{m}\theta_j \text{ (for j} \geq \text{1) .}$$

}

}