

COMP 632: Assignment 1

Due on Tuesday, January 27 2015

Presented to Dr. Doina Precup

Geoffrey Stanley
Student ID: 260645907

Question 1

A)

See source code.

B)

Let a be a vector equal to $[1/\max_i |x_j|, \dots, 1/\max_i |x_n|]$. And let A be a matrix such that $A = aI$. Substituting X by AX in the closed form regression algorithm we obtain:

$$\begin{aligned} 2(AX)^T AXw &= 2(AX)^T Y \\ 2AX^T AXw &= 2AX^T Y \\ X^T AXw &= X^T Y \\ Aw &= (X^T X)^{-1} X^T Y \end{aligned} \tag{1}$$

As such, we can conclude that normalization is equivalent to multiplying the w vector by some scalar.

C)

See source code.

D)

E)

Fold #1 results are:

Training error is : [17633.45681241]

Testing error is : [4795.16430455]

Fold #2 results are:

Training error is : [16317.25933808]

Testing error is : [6362.29874466]

Fold #3 results are:

Training error is : [15395.564805]

Testing error is : [7339.05242593]

Fold #4 results are:

Training error is : [18457.0083633]

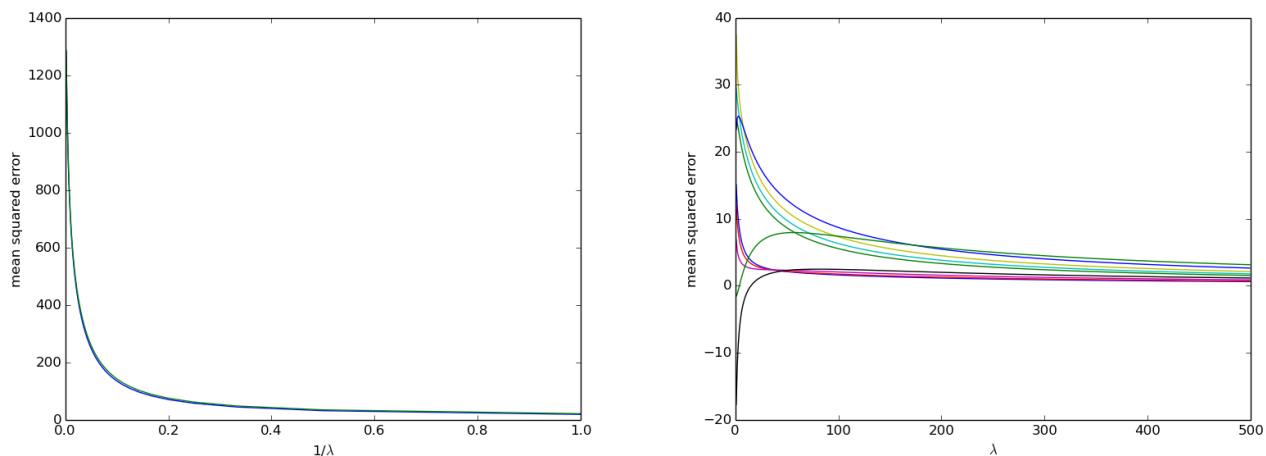
Testing error is : [3879.4511839]

Fold #5 results are:

Training error is : [20136.42642011]

Testing error is : [2188.4234707]

F)



Question 2

When simplifying the maximum likelihood equation for a regression whose variables maintain a constant normal distribution we are able to obtain the sum-squared-error function. However, when the standard deviation varies from one variable to another this simplification is no longer achievable.

$$L(w) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{1}{2}\left(\frac{y_i - h_w(x_i)}{\sigma_i}\right)^2} \quad (2)$$

$$\log L(w) = \sum_{i=1}^m \log \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \right) - \sum_{i=1}^m \frac{1}{2} \left(\frac{y_i - h_w(x_i)}{\sigma_i} \right)^2 \quad (3)$$

$$\log L(w) = \sum_{i=1}^m \log \left((2\pi)^{-1/2} \sigma_i^{-1} \right) - \sum_{i=1}^m \frac{1}{2} \left(\frac{y_i - h_w(x_i)}{\sigma_i} \right)^2 \quad (4)$$

$$\log L(w) = -\frac{1}{2} \log(2\pi) - \sum_{i=1}^m \log \sigma_i - \sum_{i=1}^m \frac{1}{2} \left(\frac{y_i - h_w(x_i)}{\sigma_i} \right)^2 \quad (5)$$

$$\frac{\partial}{\partial w_j} \log L(w) = \frac{\partial}{\partial w_j} \left(-\frac{1}{2} \log(2\pi) - \sum_{i=1}^m \log \sigma_i - \sum_{i=1}^m \frac{1}{2} \left(\frac{y_i - h_w(x_i)}{\sigma_i} \right)^2 \right) \quad (6)$$

$$\frac{\partial}{\partial w_j} \log L(w) = - \sum_{i=1}^m \frac{x_i}{\sigma_i} \left(\frac{y_i - h_w(x_i)}{\sigma_i} \right) \quad (7)$$

Simplifying further while converting standard deviation to variance. Where $\sigma^2 = \Omega$.

$$\frac{\partial}{\partial w_j} \log L(w) = \sum_{i=1}^m \frac{x_i y_i}{\Omega_i} - \sum_{i=1}^m \frac{h_w(x_i) x_i}{\Omega_i} \quad (8)$$

$$\frac{\partial}{\partial w_j} \log L(w) = X^T \Omega^{-1} Y - w X^T \Omega^{-1} X \quad (9)$$

Now setting the gradient to zero.

$$0 = X^T \Omega^{-1} Y - w X^T \Omega^{-1} X \quad (10)$$

$$w^* = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y \quad (11)$$

Question 3

Given the following huberized loss function:

$$L_H(w, \delta) = \begin{cases} (y_i - w^T x_i)^2 / 2 & \text{if } |y_i - w^T x_i| \leq \delta \\ \delta |y_i - w^T x_i| - \delta^2 / 2 & \text{otherwise} \end{cases} \quad (12)$$

A)

Find the derivative:

$$\frac{\partial}{\partial w} L_H(w, \delta) = \begin{cases} \frac{\partial}{\partial w} ((y_i - w^T x_i)^2 / 2) & \text{if } |y_i - w^T x_i| \leq \delta \\ \frac{\partial}{\partial w} (\delta |y_i - w^T x_i| - \delta^2 / 2) & \text{otherwise} \end{cases} \quad (13)$$

The equation $|y_i - w^T x_i|$ is equivalent to $\sqrt{(y_i - w^T x_i)^2}$ as such the derivative becomes:

$$\frac{\partial}{\partial w} L_H(w, \delta) = \begin{cases} -x_i(y_i - w^T x_i) & \text{if } |y_i - w^T x_i| \leq \delta \\ -x_i \delta \frac{y_i - w^T x_i}{|y_i - w^T x_i|} & \text{otherwise} \end{cases} \quad (14)$$

The expression $\frac{y_i - w^T x_i}{|y_i - w^T x_i|}$ is equivalent to the sign of $y_i - w^T x_i$ as such the derivative can be written as:

$$\frac{\partial}{\partial w} L_H(w, \delta) = \begin{cases} -x_i(y_i - w^T x_i) & \text{if } |y_i - w^T x_i| \leq \delta \\ -x_i \delta \text{sign}(y_i - w^T x_i) & \text{otherwise} \end{cases} \quad (15)$$

Where $\text{sign}(y_i - w^T x_i)$ is replaced by 1 or -1 depending on whether the result of the equation is positive or negative.

B)

The algorithm was implemented in python source code and is defined as function GradientDescent.

Data: a dataset containing features and targets, learning rate, lambda, delta

Result: a weight vector

initialize weights to 1 or a random numbers;

while *max iterations not achieved* **do**

 initialize gradient = 0;

while *not at the last row of the data array* **do**

 hypothesis = dot product (weights , row of features);

if $|y_i - w^T x_i| \leq \text{delta}$ **then**

 loss = hypothesis - row of targets;

 J = dot product (transpose of the row features, loss);

 gradient += J + lambda/2 * dot product(transpose of weights, weights);

else

 loss = hypothesis - row of targets;

if *loss is negative* **then**

 | gradient += delta * row of features

else

 | gradient -= delta * row of features

end

end

 go to the next row;

end

if *gradient * learning rate is smaller then desired precision* **then**

 | return weights;

end

 weights = weights + (learning rate) * gradient

end

return weights;

Algorithm 1: Gradient descent using huber loss function

C)

Question 4

$$h_{w1, \dots, wk}(x) = \prod_{k=1}^K \phi_k(x) \quad (16)$$

Where:

$$\phi_k(x) = e^{w_k^T x} = e^{\sum_{i=1}^m w_{ik} x_i} \quad (17)$$

Replacing $\phi_k(x)$

$$h_{w1, \dots, wk}(x) = \prod_{k=1}^K e^{\sum_{i=1}^m w_{ik} x_i} \quad (18)$$

$$\log h_{w1, \dots, wk}(x) = \sum_{k=1}^K \sum_{i=1}^m w_{ik} x_i \quad (19)$$

$$\frac{\partial}{\partial w} \log h_{w1, \dots, wk}(x) = \frac{\partial}{\partial w} \left(\sum_{k=1}^K \sum_{i=1}^m w_{ik} x_i \right) \quad (20)$$

$$\frac{\partial}{\partial w} \log h_{w1, \dots, wk}(x) = \sum_{i=1}^m x_i \quad (21)$$