Regression Theory

Aaron Briel

7/9/2018

1. Data Preprocessing

Printing the dimensions of each partition to verify the number of samples:

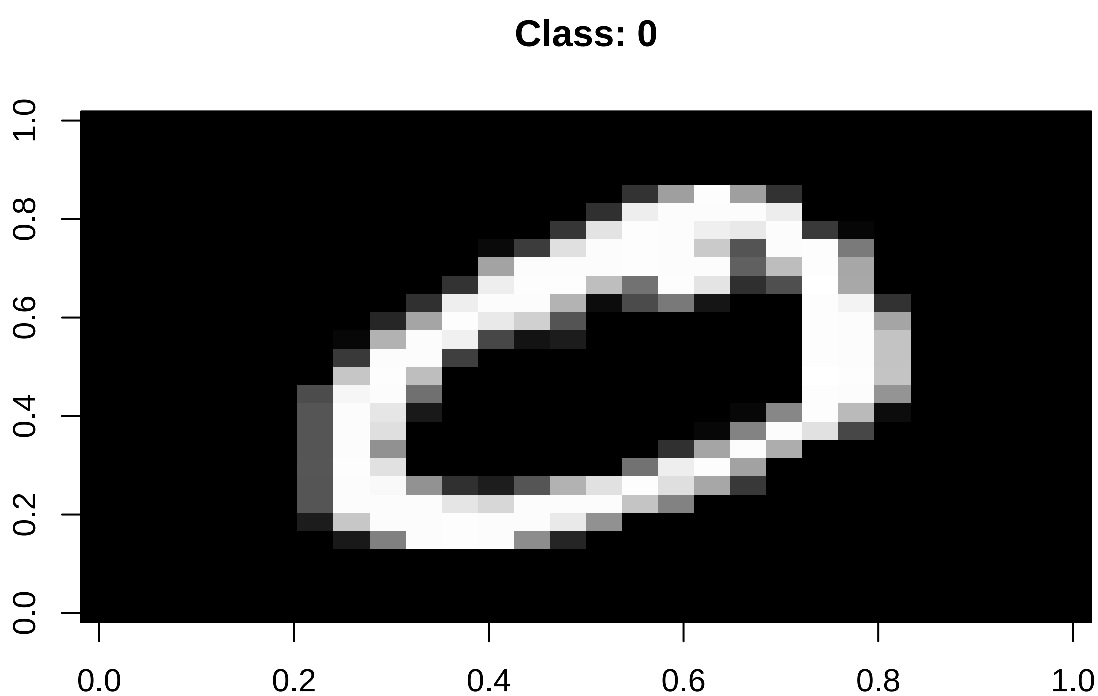
```
## [1]   785 12665
```

```
## [1]   785 11552
```

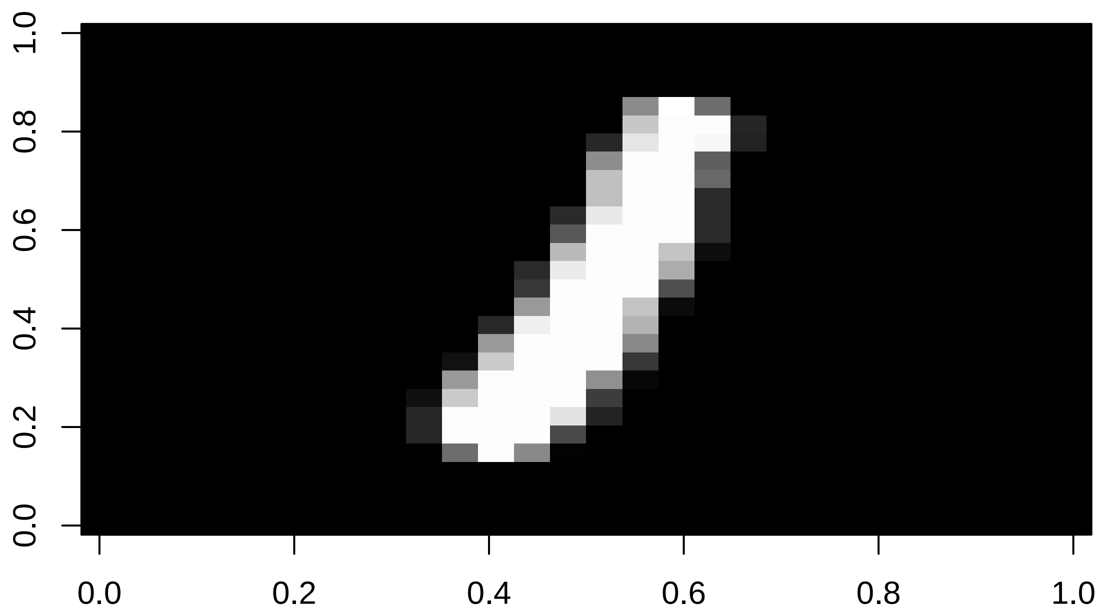
```
## [1]  785 2115
```

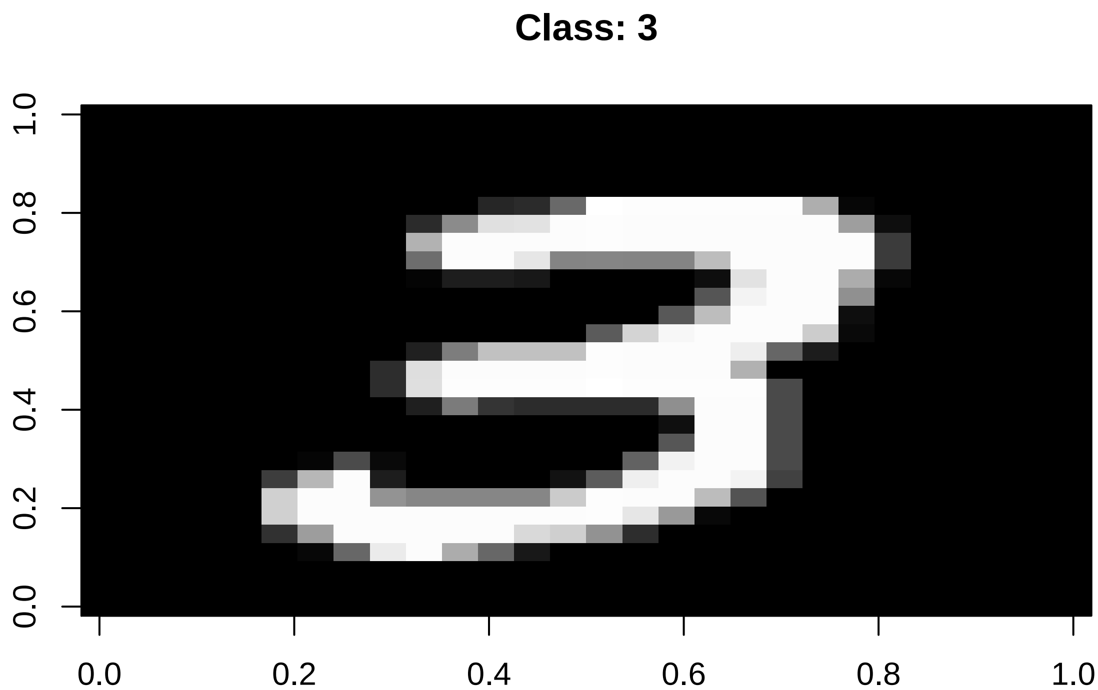
```
## [1]  785 1902
```

Visualizing an image from each class to ensure that the data was processed correctly.

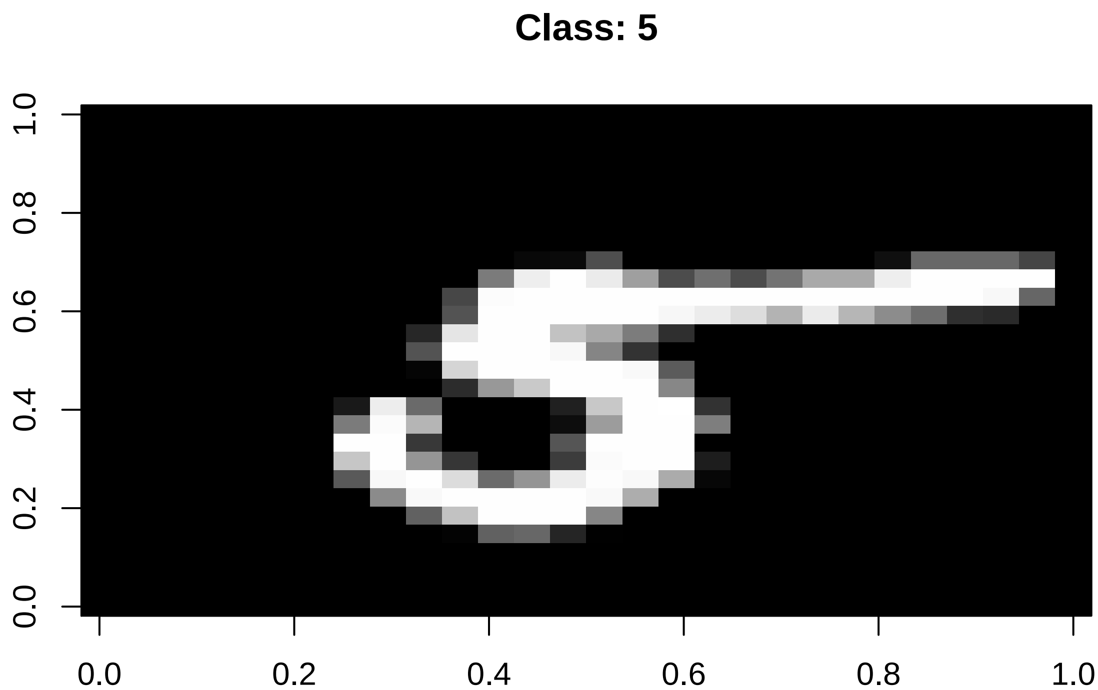


Class: 1





Class: 3
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0



2. Theory

a. The formula for the loss function used in Logistic Regression which we wish to minimize is as follows, where we are assuming that $y^{(i)} \in \{-1, +1\}$:

$$L(\theta) = \underset{\theta}{\operatorname{argmin}} \sum_{n=i}^{n} log\left(1 + \exp(-y^{(i)} \langle \theta, x^{(i)} \rangle)\right)$$

b. The gradient of the loss function with respect to the model pararameters is derived in the following steps, with the assumption that *log* is the natural logarithm.

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \theta_j} &= \frac{1}{1 + \exp(-y^{(i)} \langle \theta, x^{(i)} \rangle)} \cdot \frac{\partial (1 + \exp(-y^{(i)} \langle \theta, x^{(i)} \rangle)}{\partial \theta_j} \\
&= \frac{\exp(-y^{(i)} \langle \theta, x^{(i)} \rangle)}{1 + \exp(-y^{(i)} \langle \theta, x^{(i)} \rangle)} \cdot \Big(- y^{(i)} \cdot \frac{\partial (\langle \theta, x^{(i)} \rangle)}{\partial \theta_j} \Big) \\
&= - \frac{y^{(i)}}{1 + \exp(y^{(i)} \langle \theta, x^{(i)} \rangle)} \cdot \Big(\frac{\partial (\langle \theta, x^{(i)} \rangle)}{\partial \theta_j} \Big)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial (\langle \theta, x^{(i)} \rangle)}{\partial \theta_j} &= \frac{\partial (\theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ... + \theta_j x_j^{(i)} ... + \theta_n x_n^{(i)})}{\partial \theta_j} \\
&= (0 + 0 + ... + \frac{\theta_j x_j^{(i)}}{\partial \theta_j} + ... + 0)
\end{aligned}$$

$$= x_j^{(i)}$$

$$\frac{\partial L(\theta)}{\partial \theta_j} = -\frac{y^{(i)} x_j^{(i)}}{1 + \exp(y^{(i)} \langle \theta, x^{(i)} \rangle)}$$

c. Based on the gradient in (b), we express the Stochastic Gradient Descent (SGD) update rule that uses a single sample at a time as follows:

$$\theta_j \leftarrow \theta_j + \alpha\Big(\frac{y^{(i)} x_j^{(i)}}{1 + \exp(y^{(i)} \langle \theta, x^{(i)} \rangle)}\Big)$$

d. Pseudocode for training a model using Logistic Regression and SGD is below, where "alpha"" is the step size or learning rate, and "n" is the number of samples.

```
alpha = sufficiently small value
theta = array of length n with random values between 0 and 1
theta_old = a large integer repeated n times
theta_current = a small integer repeated n times

do until no value in absolute_value(theta_current - theta_old) < threshold:
  sample_column = array of integers sampled without replacement between 1 and n
  theta_old = theta
  for i = 1 to n
    index = sample_column[i]
    theta = theta + alpha * (y(index)*x(index) /
      (1 + exp(y(index) * dot_product(theta, x(index)))))
  end
  theta_current = theta
end

return theta
```

e. The number of operations per epoch of SGD, where number of samples is n and the dimensionality of each sample is d, can be expressed in Big-O notation as follows: $O(n * d)$

##References:

1. Citation for image creation from matrix:
Author "biomickwatson". (2016, October 6). Retrieved from
https://www.r-bloggers.com/creating-an-image-of-a-matrix-in-r-using-image/

2. Citation for the source of various RMD math notations:* Author R. Prium. (2016, October 16). Retrieved from: https://www.calvin.edu/~rpruim/courses/s341/S17/from-class/MathinRmd.html

3. Citation for RMD argmin math notation:* Answer from user "egreg". (2015, December 20). Retrieved from: https://tex.stackexchange.com/questions/5223/command-for-argmin-or-argmax/5255