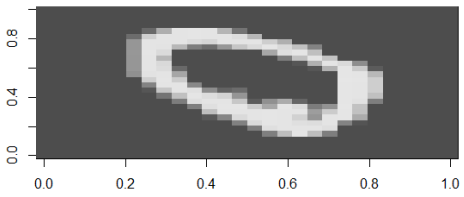
GT Username: sanne31@gatech.edu

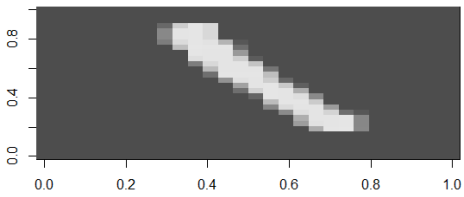
Part 1. Data Preprocessing

Visualization of Each Class in each data partition in both datasets

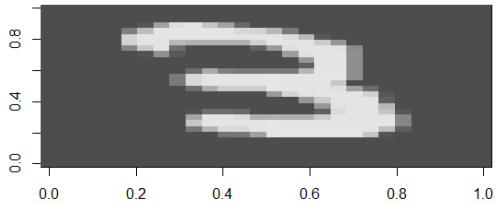
Training Set – Class Label 0



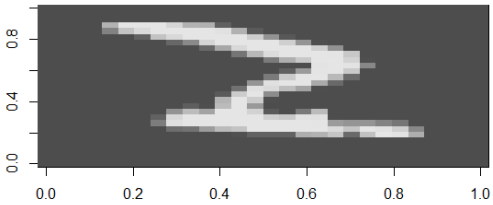
Training Set – Class Label 1



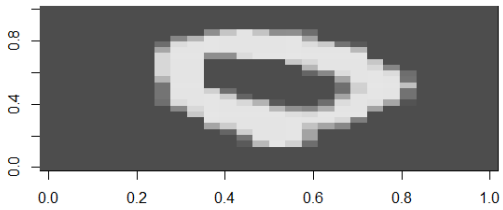
Training Set – Class Label 3



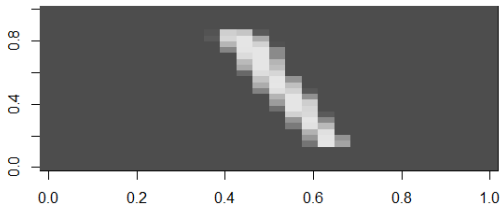
Training Set – Class Label 5



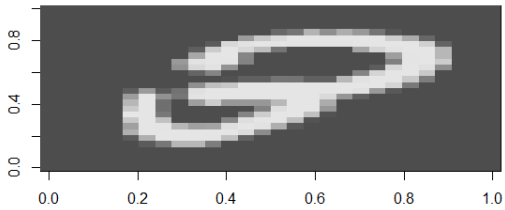
Testing Set – Class Label 0



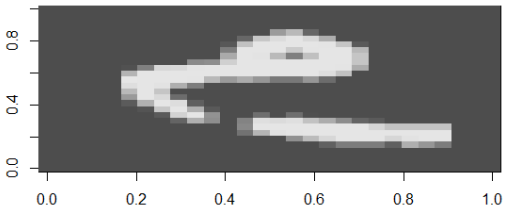
Testing Set – Class Label 1



Testing Set – Class Label 3



Testing Set – Class Label 5



Part 2. Theory

Reference: https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/pdfs/40%20LogisticRegression.pdf

https://www.cs.upc.edu/~marias/teaching/ml/1regression.pdf

Loss Function

$$L(\theta) = \prod_{i=1}^{n} \sigma\left(\theta^T x^i\right)^{y^i} \cdot \left[1 - \sigma\left(\theta^T x^i\right)\right]^{(1-y^i)}$$

$$LL(\theta) = \sum_{i=0}^{n} y^i \, log\, \sigma\left(\theta^T x^i\right) + C, -y^i \, log\left[1 - \sigma\left(\theta^T x^i\right)\right]$$

- ➔ where x is the data point represented
- ➔ Θ is the parameter vector
- ➔ y is the class label and y ϵ {-1,1}

Gradient Descent

$\frac{\partial LL(\theta)}{\partial o_j} = \frac{\partial}{\partial \theta_j} y \, log \, \sigma(\theta^T x) + \frac{\partial}{\partial \theta_j}(1-y) \, log[1-\sigma(\theta^T x)]$ - derivative of sum of terms

$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}\right] \frac{\partial}{\partial \theta_j} \sigma(\theta^T x)$ - derivative of log(f(x))

$= \left[\frac{y}{\sigma(\theta^T x)} - \frac{1-y}{1-\sigma(\theta^T x)}\right] \sigma(\theta^T x)[1-\sigma(\theta^T x)] x_j$ - chain rule + derivative of sigma

$= [y - \sigma(\theta^T x)] x_j$ - cancellation of terms

$= \left[y - \frac{1}{1+e^{-(\theta^T x)}}\right] x_j$ - final formula for gradient descent

Stochastic Gradient Descent

For a single sample at a time, the SGD update rule will become as follows:

$$\theta_j \Leftarrow \theta_j - \alpha \sum_{i=1}^{n} \frac{1}{1+e(-y^i <\theta, x^i>)}$$

Pseudo Code for Training a model using Logistic Regression

- ➔ Given α, $\{ (x^i , y^i) \}^m_{i=1}$
- ➔ Initialize a = $\langle 1, \cdots\cdots 1 \rangle^T$
- ➔ Perform feature scaling on the attribute
- ➔ Repeat until convergence
    - for each j = 0, ….., n;
        - $a'_j = a_j + \alpha \, \Sigma_i \left(y^i - h_a(x^i)\right) x_j^i$
    - for each j = 0, ... , n:
        - $a_j = a_j$
- ➔ Output a

Number of Operations

-- Each gradient descent update iteration requires $2n(d + 1)$ operations.

-- $O(n^2)$ – Big O notation