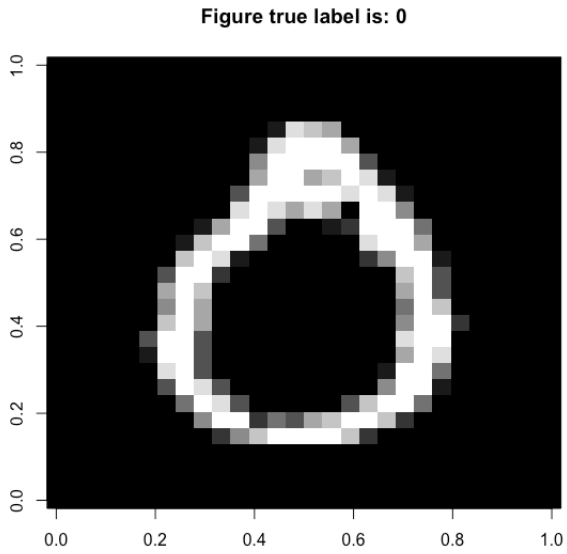
Homework 3

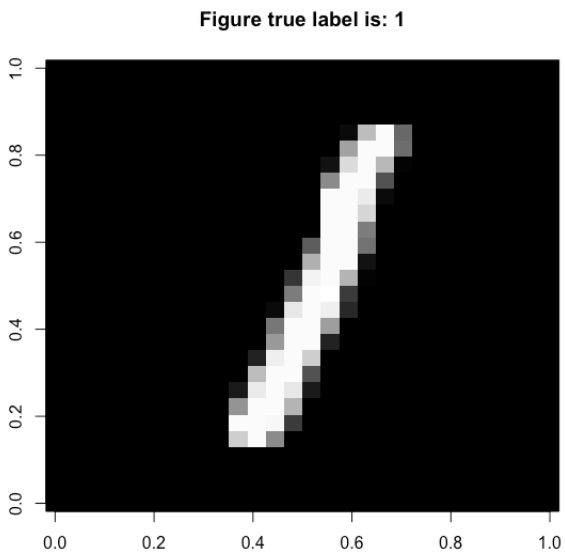
Section 1: Data Preprocessing

1. Report: 4 sample images, one from each class, along with their class labels to demonstrate you've read the data correctly.

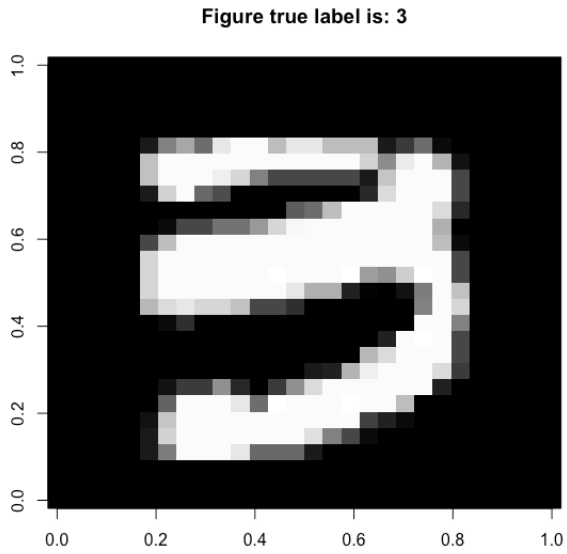
(1) Sample image with label "0":



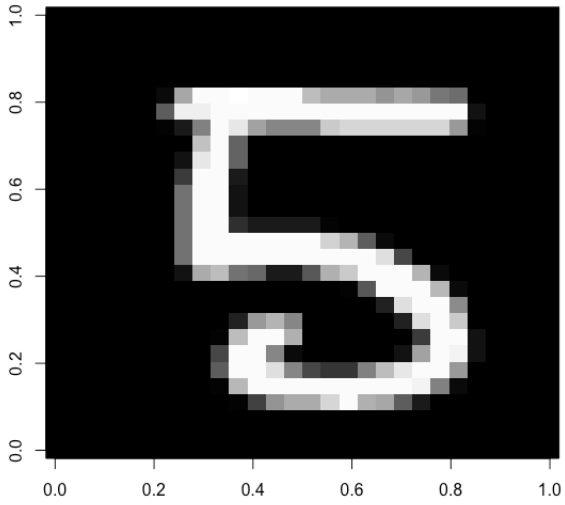
(2) Sample image with label "1":



(3) Sample image with label "3":



(4) Sample image with label "5":



Section 2: Theory

1. Write down the formula for the loss function used in Logistic Regression, the expression that you want to minimize: L(θ)

We first define logistic function as following:

Eq1: $$f(x) = \frac{1}{1+e^{-x}}$$

We assume $y^{(i)} \in \{0,1\}$ and:

Eq2: $$P\left(y^{(i)}\middle|x^{(i)};\,\theta\right) = \begin{cases} f\left(\theta^T x^{(i)}\right) = \frac{1}{1+e^{-\theta^T x^{(i)}}}, & y^{(i)} = 1 \\ 1 - f\left(\theta^T x^{(i)}\right) = \frac{1}{1+e^{\theta^T x^{(i)}}}, & y^{(i)} = 0 \end{cases}$$

Eq2 could be rewritten as following:

Eq3:
$$P\left(y^{(i)}\middle|x^{(i)};\,\theta\right) = \left[f\left(\theta^T x^{(i)}\right)\right]^{y^{(i)}}\left[1 - f\left(\theta^T x^{(i)}\right)\right]^{1-y^{(i)}}$$

According to the definition of the likelihood of parameters we have:

Eq4:
$$Likelihood(\theta) = \prod_{i=1}^{n} P\left(y^{(i)}\middle|x^{(i)};\,\theta\right)$$

Put Eq3 into Eq4:

Eq5:
$$Likelihood(\theta) = \prod_{i=1}^{n}\left[f\left(\theta^T x^{(i)}\right)\right]^{y^{(i)}}\left[1 - f\left(\theta^T x^{(i)}\right)\right]^{1-y^{(i)}}$$

Take natural logarithm on the both sides of Eq5:

Eq6:
$$\ln[Likelihood(\theta)] = \sum_{i=1}^{n} y^{(i)}\,\ln\left[f\left(\theta^T x^{(i)}\right)\right] + (1 - y^{(i)})\ln\left[1 - f\left(\theta^T x^{(i)}\right)\right]$$

Since we want to maximize the likelihood and minimize the loss function L(θ), the loss function should be the negative $\ln[Likelihood(\theta)]$

Eq7:

$$L(\theta) = -\ln[Likelihood(\theta)] = -\sum_{i=1}^{n} y^{(i)} \ln\left[f\left(\theta^T x^{(i)}\right)\right] + (1 - y^{(i)}) \ln\left[1 - f\left(\theta^T x^{(i)}\right)\right]$$

$$= \sum_{i=1}^{n} -y^{(i)} \ln\left[f\left(\theta^T x^{(i)}\right)\right] - \ln\left[1 - f\left(\theta^T x^{(i)}\right)\right] + y^{(i)} \ln\left[1 - f\left(\theta^T x^{(i)}\right)\right]$$

$$= \sum_{i=1}^{n} y^{(i)} \ln\frac{1 - f\left(\theta^T x^{(i)}\right)}{f(\theta^T x^{(i)})} - \ln\left[1 - f\left(\theta^T x^{(i)}\right)\right]$$

$$= \sum_{i=1}^{n} y^{(i)} \ln e^{-\theta^T x^{(i)}} - \ln\left[\frac{1}{1 + e^{\theta^T x^{(i)}}}\right]$$

$$= \sum_{i=1}^{n} -y^{(i)} \theta^T x^{(i)} + \ln\left[1 + e^{\theta^T x^{(i)}}\right]$$

2. Derive the gradient of the loss function with respect to model parameters: dL(θ)dθ or ∂L(θ)∂θj (Hint: Use chain rule to successively differentiate the terms.)

We first define $u = 1 + e^{\theta^T x^{(i)}}$, $v = \theta^T x^{(i)}$ and we have:

Eq8:

$$\frac{d \ln u}{du} = \frac{1}{u} = \frac{1}{1 + e^{\theta^T x^{(i)}}}$$

Eq9:

$$\frac{du}{dv} = e^v = e^{\theta^T x^{(i)}}$$

Eq10:

$$\frac{dv}{d\theta_j} = \frac{d(\theta_1 x_{(i)1} + \theta_2 x_{(i)2} + \cdots + \theta_j x_{(i)j} + \cdots + \theta_d x_{(i)d})}{d\theta_j} = x_{(i)j}$$

So the gradient of the loss function should be:

Eq11:

$$\frac{dL(\theta)}{d\theta_j} = \sum_{i=1}^{n} -y^{(i)} \frac{d\theta^T x^{(i)}}{d\theta_j} + \frac{d \ln\left[1 + e^{\theta^T x^{(i)}}\right]}{d\theta_j}$$

$$= \sum_{i=1}^{n} -y^{(i)} \frac{dv}{d\theta_j} + \frac{d \ln u}{du} \cdot \frac{du}{dv} \cdot \frac{dv}{d\theta_j}$$

Put Eq8, Eq9 and Eq10 into Eq11:

Eq12:

$$\frac{dL(\theta)}{d\theta_j} = \sum_{i=1}^{n} -y^{(i)}\ x_{(i)j} + \frac{1}{1+e^{\theta^T x^{(i)}}} \cdot e^{\theta^T x^{(i)}} \cdot x_{(i)j}$$
$$= \sum_{i=1}^{n} -y^{(i)}\ x_{(i)j} + \frac{1}{1+e^{-\theta^T x^{(i)}}} \cdot x_{(i)j}$$
$$= \sum_{i=1}^{n} -y^{(i)}\ x_{(i)j} + f\left(\theta^T x^{(i)}\right) \cdot x_{(i)j}$$

3. Based on this gradient, express the Stochastic Gradient Descent (SGD) update rule that uses a single sample <x(i),y(i)> at a time.

From Eq12, the gradient of a single sample <x(i),y(i)> should be:
Eq13:

$$\frac{dL\left(\theta, x^{(i)}, y^{(i)}\right)}{d\theta_j} = -y^{(i)}\ x_{(i)j} + f\left(\theta^T x^{(i)}\right) \cdot x_{(i)j}$$

And the SGD update rule that uses a single sample <x(i),y(i)> at a time is:
Eq14:

$$\theta_j := \theta_j - \alpha \cdot \frac{dL\left(\theta, x^{(i)}, y^{(i)}\right)}{d\theta_j}$$

Where α is a parameter called learning rate. Put Eq13 into Eq14 we finally have:
Eq15:

$$\theta_j := \theta_j + \alpha \cdot \left[y^{(i)} - f\left(\theta^T x^{(i)}\right)\right] \cdot x_{(i)j}$$

4. Write pseudocode for training a model using Logistic Regression and SGD.

procedure LogisticRegressionTraining(x, y, θ, α)
  Randomly initialize parameters θ and learning rate α
  while *Not Converged* **do**
    randomize order of examples in training set(x, y)
    for $i = 1, 2, \cdots, n$ **do**
      for $j = 1, 2, \cdots, d$ **do**
        $\theta_j := \theta_j + \alpha \cdot \left[y^{(i)} - f\left(\theta^T x^{(i)}\right)\right] \cdot x_{(i)j}$
      end for
    end for
  end while
end procedure

5. Estimate the number of operations per epoch of SGD, where an epoch is one complete iteration through all the training samples. Express this is Big-O notation, in terms of the number of samples (n) and the dimensionality of each sample (d).

The number of operations per epoch of SGD equals to the number of samples (n) times the dimensionality of each sample (d), or O(n * d) if we use Big-O notation.