Question 1

1(a)

Linear regression assumes uncertainty in measurement in only dependent variable (Y). Because independent variables (X) is treated as fixed values.

1(b)

A simple solution to tackle the sensitivity of Linear Regression model towards outliers is to use a Robust Estimation technique. One good example can be Theil-sen Estimator.

This estimator for a set of two-dimensional points (x_i, y_i) is the median 'm' of the slopes $(y_j - y_i)/(x_j - x_i)$ determined by all pairs of sample points. The median of slopes is taken only from pairs having distinct x - coordinates. Once the slope 'm' has been determined, a line can be determined form the sample points by setting the y-intercept b to be the median of the values $y_i - mx_i$. This method thus takes into account of the outliers and given a better decision boundary.

1(c)

Regression coefficients can have lesser absolute values. but still add very high weight-age to corresponding features.
Lets assume Age as a parameter. If age is mentioned in years instead of months, the absolute values might decrease. But the importance of Age still remains the same. It wouldnt mean that age has become more prominent.

1(d)

When some independent variables are perfect linear combination of other independent variables, we would not be able to obtain a unique estimate of coefficients as some columns in the matrix will reduce to all zeros. This in turn will bring down the efficiency of the Linear Regression model.

1(e)

Here, 1-to-K encoding uses K-bits, each to represent one of the K values. I would suggest we use **Dummy Coding**, which is a minor tweak to the 1-to-K encoding.

In Dummy Coding, we need just K-1 parameters to represent K values. We achieve this by starting to encode each value from $X \in [1 \text{ to } (K-1)^{th}]$ by turning ON just the X^{th} bit. But the change we do is, K^{th} value type is represented by using all 0's in all $(K-1)$ bits.

1(f)

Highly correlated independent variables can have undue effect on coefficients. Some of the effects can be :
i) Errors of effected coefficients tend to be very large. Also, Regression model can become very sensitive to even small changes in the input data.
ii) When new test data differs greatly from trained data, then there can be large errors in predictions of the coefficient values which in turn bring down the overall accuracy of the model.
iii) We can have imprecise estimate of the effect of change in variables on the coefficients as the variables are related to each other.

1(g)

Since Logistic regression is a special case of the more generalized linear regression method, in Linear regression if there are not too many errors in calculating the coefficients, then we would get the same accuracy in linear regression model as compared to the logistic regression (almost same). Since the results of logistic regression are limited between 0 and 1, the error terms in linear regression are not normally distributed in most cases. Nor is the error variance constant. Because of this, in some cases if error terms are huge, linear regression would not give same accuracy. Further calculations can be made to make the linear regression model mimic the logistic regression model performance wise. When we want to calculate a continuous measure, we use linear regression. But if we want to predict categories based on a value, we use logistic regression. For eg, if you want to see how body mass index predicts cholesterol values (a continuous value) , linear regression is better. But if you want to see how body mass index predicts the odds of being diabetic (a binary diagnosis), logistic regression is preferred.

1(h)

Suppose we have as few as 10 samples with over 100 features of each sample. In Linear Regression, we consider each sample as a linear equation with over 100 unknown coefficients. Such a situation leads to a under-determined system of equations. And such a system cannot be solved by simple mathematical techniques. But, some of them may be solved by using few estimation and elimination techniques.

Question 2

2(a)

We are give with samples from two classes.

$$P(x|y=c1) = \mathcal{N}(0,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} \tag{1}$$

$$P(x|y=c2) = \mathcal{N}(0,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\delta/2)^2}{2\sigma^2}} \tag{2}$$

We have the formula for μ and σ^2 as follows:

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

$$\bar{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mu})(x_i - \bar{\mu})^T \tag{4}$$

Since Variance of both the classes are same, we just have uni-variate data. Hence GDA is simplified to LDA for this problem.

Solutions for Class-1 and Class-2 Using formula (3) and (4)
Class-1
Here,

$$\mu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

$$\sigma_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_1)(x_i - \mu_1)^T = \sigma^2$$

$$\text{Where } X_i = (x_{i1}, x_{i2}, x_{i2} \ldots x_{i2D})$$

Class-2
Here,

$$\mu_2 = \frac{1}{n}\sum_{i=1}^{n} X_i = 0$$

$$\sigma_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_2)(x_i - \mu_2)^T = \sigma^2$$

$$\text{Where } X_i = (x_{i1}, x_{i2}, x_{i2} \ldots x_{i2D})$$

The μ changes if δ changes.

2(b)

Prove Gaussian is Logistic

Given,

$$P(x|y=c_1) = \mathcal{N}(\mu_1, \Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{\left(\frac{-1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)} \tag{5}$$

$$P(x|y=c_2) = \mathcal{N}(\mu_2, \Sigma) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{\left(\frac{-1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)} \tag{6}$$

For any value of y. By bayes theorem and eqns (5) and (6),

$$P(y=c_1|x) = \frac{P(x|y=c_1)P(y=c_1)}{P(x|y=c_1)P(y=c_1) + P(x|y=c_2)P(y=c_2)} = \frac{1}{1+\frac{P(x|y=c_2)P(y=c_2)}{P(x|y=c_1)P(y=c_1)}} \tag{7}$$

$$= \frac{1}{1+E} \tag{8}$$

From above derivation, we have,

$$E = \frac{P(x|y=c_2)P(y=c_2)}{P(x|y=c_1)P(y=c_1)}$$

$$= \frac{\frac{\pi_{c_2}}{\sqrt{(2\pi)^k|\Sigma|}} e^{\left(\frac{-1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right)}}{\frac{\pi_{c_1}}{\sqrt{(2\pi)^k|\Sigma|}} e^{\left(\frac{-1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)}}$$

$$E = \frac{\pi_{c_2}}{\pi_{c_1}} e^{\left[\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]}$$

Applying LOG on both sides,

$$ln(E) = ln(\frac{\pi_{c_2}}{\pi_{c_1}}) + \left[\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]$$

$$\Rightarrow E = e^{ln(\frac{\pi_{c_2}}{\pi_{c_1}}) + \left[\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]}$$

$$= e^{-\left[(\Sigma^{-1}(\mu_1-\mu_2))^T X + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\frac{\pi_{c_1}}{\pi_{c_2}})\right]}$$

Solving above value of E in eqn (4),

$$P(y|x) = \frac{1}{1 + e^{-\left[(\Sigma^{-1}(\mu_1-\mu_2))^T X + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\frac{\pi_{c_1}}{\pi_{c_2}})\right]}}$$

where,

$$\Theta = (\Sigma^{-1}(\mu_1 - \mu_2))$$
$$\beta = \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + ln(\frac{\pi_{c_1}}{\pi_{c_2}})$$

Prove Logistic need not always be Multivariate Gaussian

We need to show that there are other distributions apart from multivariate Gaussian that follows logistic function.
Lets assume Poisson Distribution for two classes c_1 an c_2

$$P(x, y = c_1) = \frac{\lambda_1^x}{x!}e^{-\lambda_1} \quad (9)$$
$$P(x, y = c_2) = \frac{\lambda_2^x}{x!}e^{-\lambda_2} \quad (10)$$

From Bayes theorem, we get,

$$P(y = c_1|x) = \frac{P(x|y = c_1)P(y = c_1)}{P(x|y = c_1)P(y = c_1) + P(x|y = c_2)P(y = c_2)} = \frac{1}{1+\frac{P(x|y=c_2)P(y=c_2)}{P(x|y=c_1)P(y=c_1)}}$$
$$= \frac{1}{1 + E}$$

From above derivation, we have,

$$E = \frac{P(x|y=c_2)P(y=c_2)}{P(x|y=c_1)P(y=c_1)}$$
$$= \frac{\frac{\lambda_2^x}{x!}e^{-\lambda_2}}{\frac{\lambda_1^x}{x!}e^{-\lambda_1}}$$
$$E = \frac{\lambda_2^x e^{-\lambda_2}}{\lambda_1^x e^{-\lambda_1}}$$

Applying LOG on both sides,

$$ln(E) = xln(\lambda_2) - xln(\lambda_1) + \lambda_1 - \lambda_2$$

$$\Rightarrow E = e^{-[(ln\lambda_1 - ln\lambda_2)X + \lambda_2 - \lambda_1]}$$

Substituting the value of E, we get,

$$p(y|x) = \frac{1}{1+e^{-[(ln\lambda_1 - ln\lambda_2)X + (\lambda_2 - \lambda_1)]}}$$

Hence, we can prove that, Logistic function doesnt always imply Gaussian Distribution as we just obtained Logistic function from a Poisson Distribution.

Question 3

We are given the following,

$$\text{Need to maximize: } ||w_{i+1} - w_i||_2$$
$$\text{Where given } (w_{i+1}^T x_i) - y_i = 0$$

We can use Lagrange multiplier to solve the problem.

$$\mathcal{L}(w_{i+1}, \lambda) = ||w_{i+1} - w_i||_2 + \lambda((w_{i+1}^T x_i) - y_i) = (w_{i+1} - w_i)^T (w_{i+1} - w_i) + \lambda((w_{i+1}^T x_i) - y_i) \quad (11)$$

Now we find partial derivatives of eqn (5)

$$\frac{\partial \mathcal{L}(w_{i+1}, \lambda)}{\partial w_{i+1}} = \frac{\partial}{\partial w_{i+1}} [w_{i+1}^T w_{i+1} - w_{i+1}^T w_i - w_i^T w_{i+1} + w_i^T w_i + \lambda w_{i+1}^T x_i - \lambda y_i]$$
$$= [2w_{i+1} - w_i - w_i + \lambda x_i]$$

$$\frac{\partial \mathcal{L}(w_{i+1}, \lambda)}{\partial w_{i+1}} = [2w_{i+1} - 2w_i + \lambda x_i] = 0 \quad (12)$$
$$\frac{\partial \mathcal{L}(w_{i+1}, \lambda)}{\partial \lambda} = w_{i+1}^T x_i - y_i = 0 \quad (13)$$

From eqns (13) and (14), we get,

$$w_{i+1} = w_i - \tfrac{1}{2} \lambda x_i \quad (14)$$
$$w_{i+1}^T x_i = y_i \quad (15)$$

Question 4

4(a)

Missing Values

ID	Independent Variable	Missing Values
1	pclass	0
2	name	0
3	sex	0
4	age	263
5	sibsp	0
6	parch	0
7	ticket	0
8	fare	1
9	cabin	1014
10	embarked	2
11	boat	823
12	body	1188
13	home.dest	564

4(b)

From the graphs, we deduce below relationship of Probability of survival with each independent variable. **Monotonous Relationship**
1. PClass
Non-Monotonous Relationship
1. Age 2. Sibsp 3. Parch 4. Fare

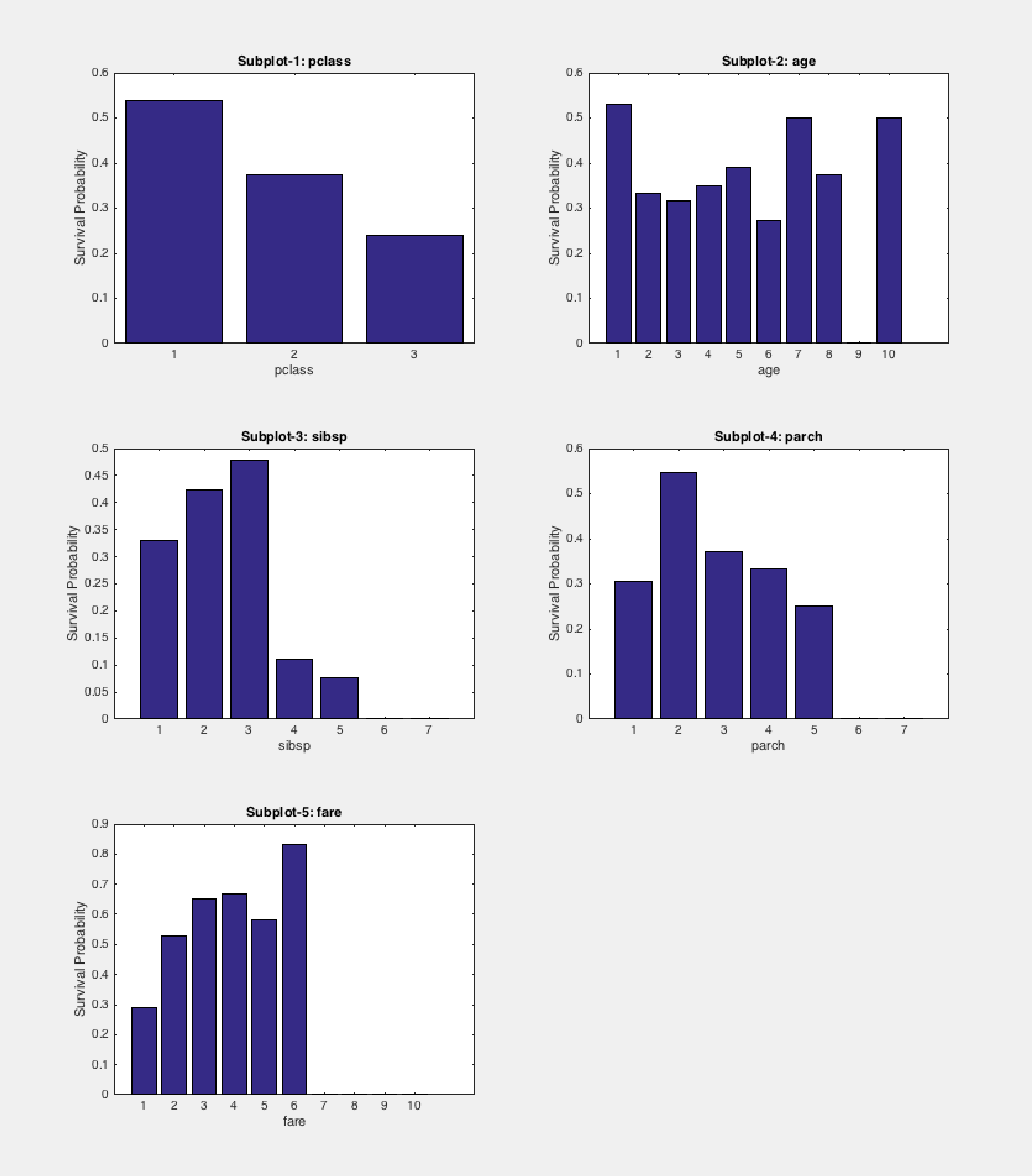

Figure 1: Monotonic Relationships

4(c)

ID	Independent Variable	Mutual Information
1	name	0.65785
2	ticket	0.57797
3	cabin	0.52282
4	home.dest	0.49965
5	sex	0.13964
6	fare	0.06747
7	pclass	0.061333
8	boat	0.04517
9	parch	0.020705
10	sibsp	0.017055
11	age	0.016515
12	embarked	0.0090031
13	body	0

4(d)

	Multiple Models	Substituting Values
Training Data:	81.68%	81.53%
Test Data:	77.68%	77.53%

Surprisingly, we see that both the models give almost same accuracy rates. Hence, we can say AGE variable doesnt have very high influence on the output Y.

4(e)

I used an encoding called **Dummy Coding**. In this method, we can encode K values using just K-1 binary bits. The Kth value is represented by all K-1 bits having 0's,

The number of rows were 903 when got reduced to **776** after deleting single value columns.

4(f)

ID	Training Accuracy	Test Accuracy
1	0.6565	0.5795
2	0.7557	0.6820
3	0.7557	0.6820
4	0.7756	0.6988
5	0.7618	0.6835
6	0.7664	0.7355
7	0.8260	0.7722
8	0.8244	0.7706
9	0.7588	0.6850
10	0.8275	0.7737

In the above graph and table, we see that, both Training and Test Accuracy are almost linearly increasing with the accuracy reaching almost 80% at 10 features. Hence this method of feature selection does seem to work to some extent.

Another pattern we see is that the accuracy seems to converge to some optimal point. So, we can assume that after reaching some particular feature count, the accuracy improvement might stop or become negligible. As far as the optimal number of features, the accuracy also depends on the combination of features and not just the number of features.

4(g)

As per the graph, we can say that the Convergence depends highly on **Stepsize** parameter. For **low Stepsize** values like 0.1, 0.2, 0.3, we see that the values converge very soon and there is a stable and steady convergence.

For **high Stepsize** values like, 0.5 and above, the convergence is mostly zig-zag and there is no guarantee that the the value would converge.

For Good stepsize (low) values, we see that we needed around 40 iterations to converge to optimum value.

4(h)

We see that, Newton Method converges faster than Gradient Descent. And the Accuracy is slightly greater than the glmfit accuracy we got. We see that it took only around 10 iterations to reach the optimum accuracy compared to 40 iterations in Gradient Descent.

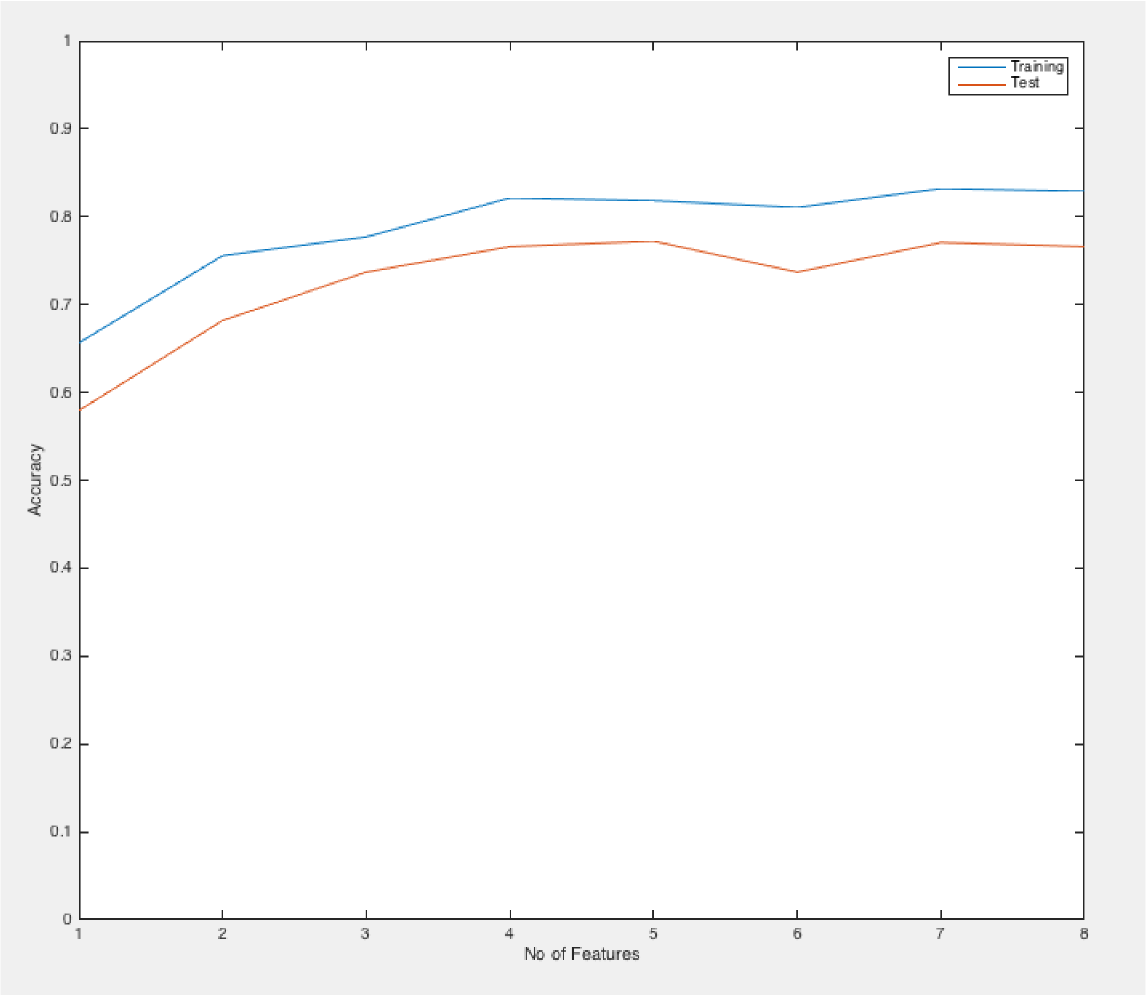

Figure 2: 4.(f) Sequential Feature Selection

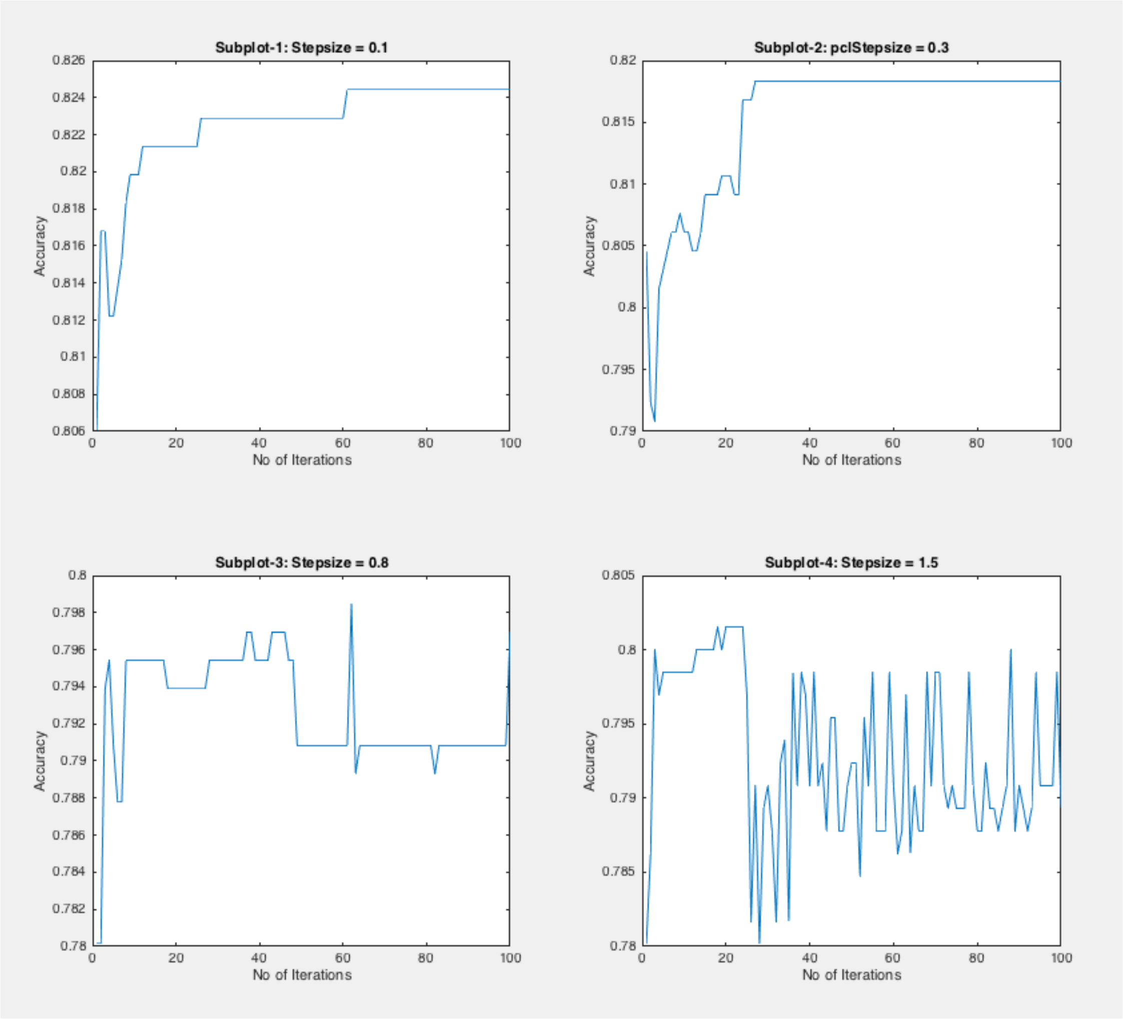

Figure 3: 4.(g) Gradient Descent Logistic Regression

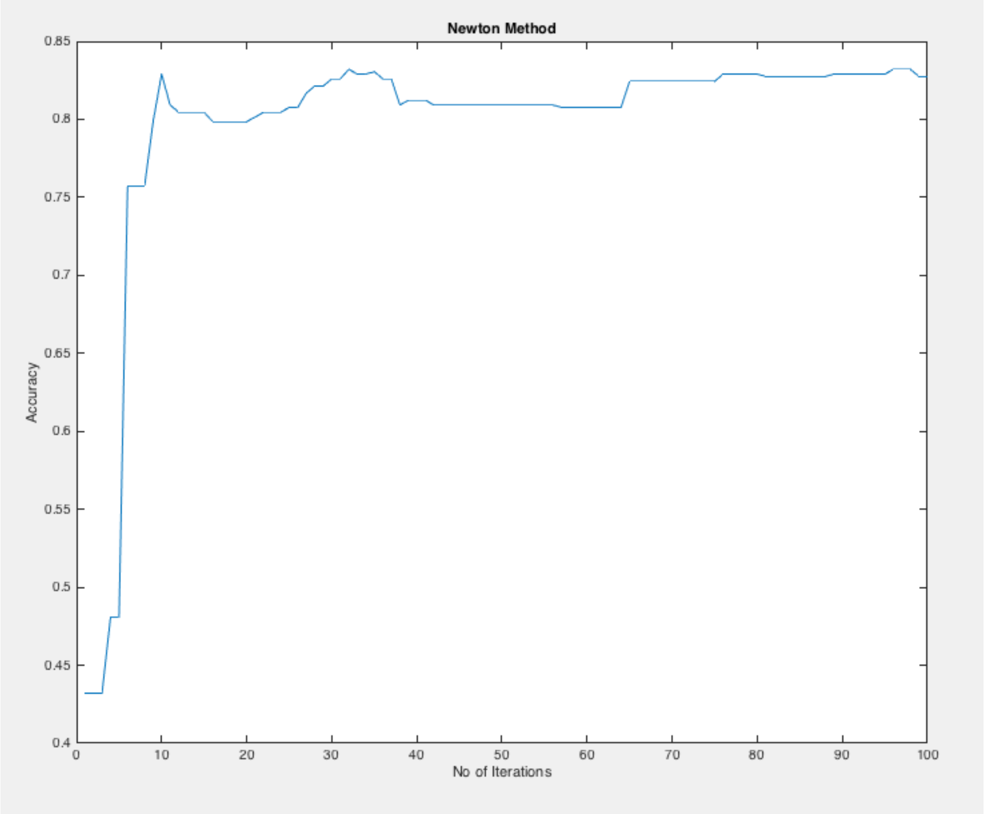

Figure 4: 4.(h) Newton Method Logistic Regression