1 (a)

$$\frac{\partial E(w)}{\partial w_j} = \sum_{i=1}^{N} (x_j)^{(i)} y^{(i)} (1 - \sigma(w^T x^{(i)})) - (x_j)^{(i)} \sigma(w^T x^{(i)}) + (x_j)^{(i)} y^{(i)} \sigma(w^T x^{(i)})$$

$$= \sum_{i=1}^{N} (x_j)^{(i)} y^{(i)} - (x_j)^{(i)} \sigma(w^T x^{(i)})$$

$$\frac{\partial^2 E(w)}{\partial (w_j)^2} = -\sum_{i=1}^{N} (x_j)^{(i)} \sigma(w^T x^{(i)})(1 - \sigma(w^T x^{(i)}))(x_j)^{(i)}$$

$$\frac{\partial^2 E(w)}{\partial w_j w_k} = -\sum_{i=1}^{N} (x_j)^{(i)} \sigma(w^T x^{(i)})(1 - \sigma(w^T x^{(i)}))(x_k)^{(i)}$$

Let $X \in \mathbb{R}^{n \times m}$ be the design matrix and $w \in \mathbb{R}^m$ be the weight vector, where n is the number of observations and m is the number of features. Let $\odot$ express the Hadamard product of its operands. We can express the second-order partial derivatives in matrix form as

$$X^T diag[-\sigma(Xw) \odot (1 - \sigma(Xw))]X$$

which gives us our Hessian matrix.

1 (b) Let $x, z \in \mathbb{R}^m$. Then

$$(x^T z)(x^T z) = \sum_{i=1}^{N} z_i x_i \sum_{j=1}^{N} x_j z_j = \sum_{i=1}^{N} \sum_{j=1}^{N} z_i x_i x_j z_j = (x^T z)^2$$

Consider

$$z^T X^T diag[\sigma(Xw) \odot (1 - \sigma(Xw))]Xz$$

Let D represent $X^T diag[\sigma(Xw)(1 - \sigma(Xw))]X$. $D \in \mathbb{R}^{m \times m}$. Let D_i be a column of D and $D^{(j)}$ be a row of D. we can now express the above as

$$-\sum_{i=1}^{m} [\sum_{j=1}^{m} z_j (D_i)^{(j)}] z_i$$

$$= -\sum_{i=1}^{m} [\sum_{j=1}^{m} z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}}] z_i = -\sum_{i=1}^{m} \sum_{j=1}^{m} z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}} z_i$$

Since in general $\sum_{i=1}^{N}\sum_{j=1}^{N} z_i x_i x_j z_j = (x^T z)^2 \geq 0$,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} z_j((D_i)^{(j)})^{\frac{1}{2}}((D_i)^{(j)})^{\frac{1}{2}} z_i \geq 0$$

therefore

$$-\sum_{i=1}^{m}[\sum_{j=1}^{m} z_j (D_i)^{(j)}] z_i \leq 0 \; \square$$

The Hessian matrix is negative semi-definite, and our original log-likelihood function is concave.

1 (c) We can invert the sign of the log-lihelihood function to get a loss function. This gives us the Hessian matrix $X^T diag[\sigma(Xw) \odot (1-\sigma(Xw))]X$. For the gradient of the loss function with respect to w, we can take the partial derivative expression from **1 (a)**

$$\sum_{i=1}^{N} (x_j)^{(i)} y^{(i)} - (x_j)^{(i)} \sigma(w^T x^{(i)})$$

invert the sign, and express it in matrix form:

$$\nabla_w E = X^T \sigma(Xw) - X^T y$$

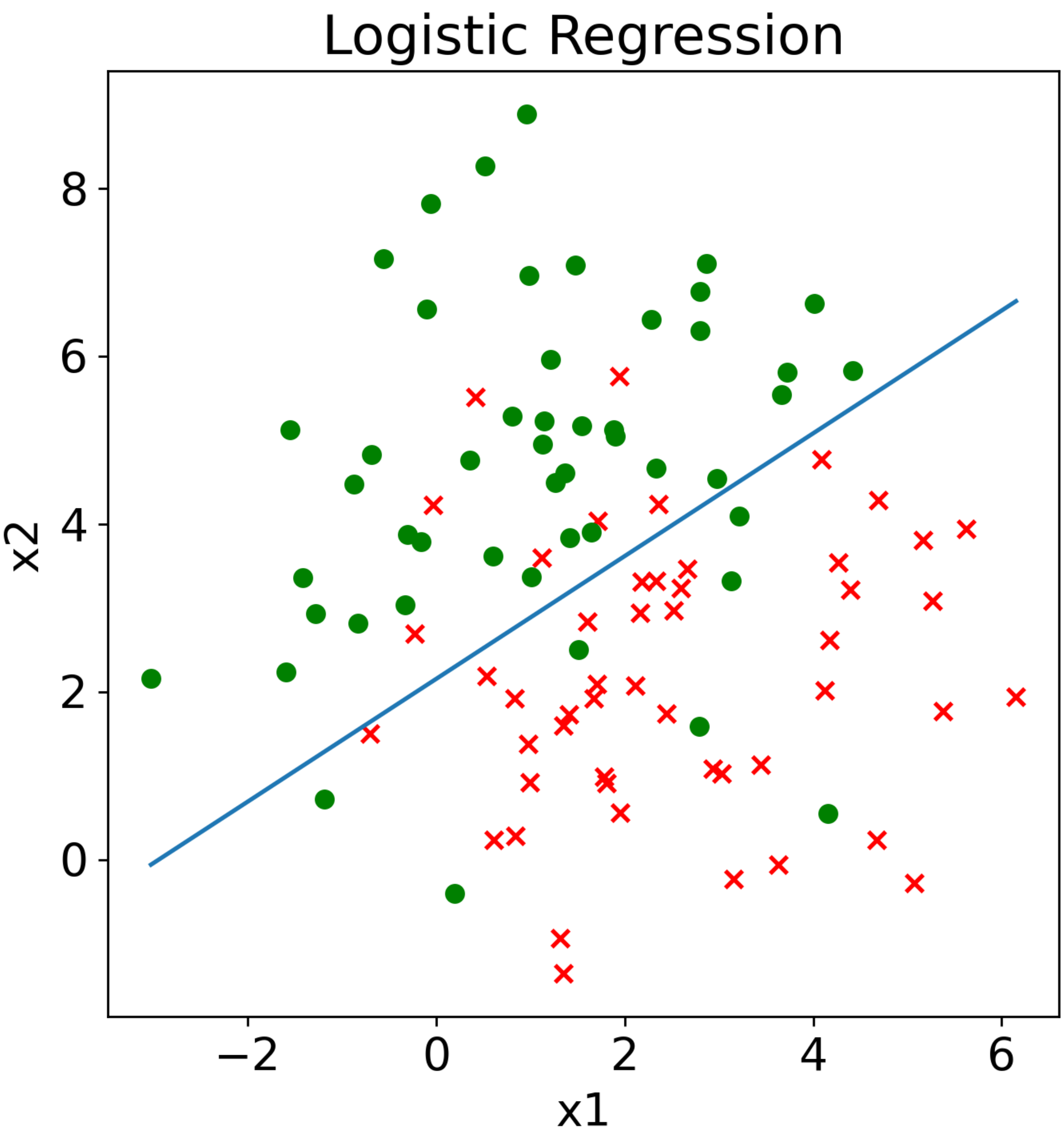
So we the update for Newton's method is

$$w_{new} = w_{old} - H^{-1} \nabla_w E$$

$$w_{new} = w_{old} - \{X^T diag[\sigma(Xw) \odot (1-\sigma(Xw))]X\}^{-1}[X^T \sigma(Xw) - X^T y]$$

1 (e) The weights from the logistic regression are
w_0: -1.84922892 w_1: -0.62814188 w_2: 0.85846843

1 (f)



2 (a) If we consider only w_m, we can express the log-likelihood as

$$\sum_{i=1}^{N} log([p(y^{(i)} = m|x^{(i)}, w)]^{\mathbb{I}\{y^{(i)}=m\}})$$

$$\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\} log([\frac{exp[w_m^T \phi(x^{(i)})]}{\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]}])$$

$\nabla w_m l(w) =$

$$\sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\} \frac{\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]}{exp[w_m^T \phi(x^{(i)})]} [\frac{exp[w_m^T \phi(x^{(i)})]\phi(x^{(i)})}{\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]} - \frac{(exp[w_m^T \phi(x^{(i)})])^2 \phi(x^{(i)})}{(\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})])^2}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\} [\frac{\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})] exp[w_m^T \phi(x^{(i)})]\phi(x^{(i)})}{exp[w_m^T \phi(x^{(i)})] \sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})]}$$

$$- \frac{\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})] (exp[w_m^T \phi(x^{(i)})])^2 \phi(x^{(i)})}{exp[w_m^T \phi(x^{(i)})] (\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})])^2}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\} [\phi(x^{(i)}) - \frac{(exp[w_m^T \phi(x^{(i)})])\phi(x^{(i)})}{(\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})])}]$$

$$= \sum_{i=1}^{N} \mathbb{I}\{y^{(i)} = m\} \phi(x^{(i)}) [1 - \frac{(exp[w_m^T \phi(x^{(i)})])}{(\sum_{j=1}^{K-1} exp[w_j^T \phi(x^{(i)})])}] \square$$