

1 (a)

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

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

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

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

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

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

$$\begin{aligned}
& - \sum_{i=1}^m \left[\sum_{j=1}^m z_j (D_i)^{(j)} \right] z_i \\
&= - \sum_{i=1}^m \left[\sum_{j=1}^m z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}} \right] z_i = - \sum_{i=1}^m \sum_{j=1}^m z_j ((D_i)^{(j)})^{\frac{1}{2}} ((D_i)^{(j)})^{\frac{1}{2}} z_i
\end{aligned}$$

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

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

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-likelihood function to get a loss function. This gives us the Hessian matrix $X^T \text{diag}[\sigma(Xw) \odot (1 - \sigma(Xw))]X$. 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 the update for Newton's method is

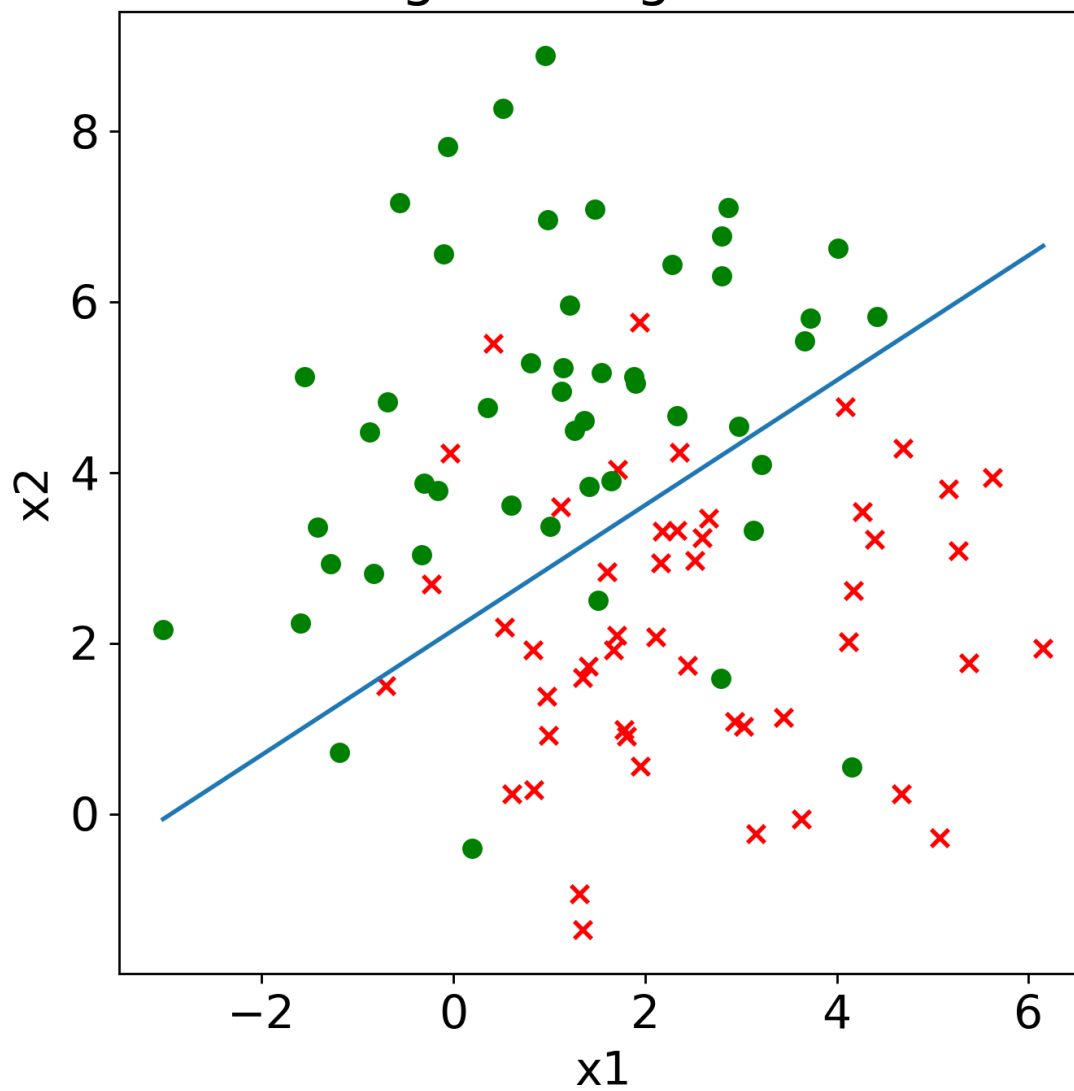
$$w_{new} = w_{old} - H^{-1} \nabla_w E$$

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

1 (e) The weights from the logistic regression are

w_0 : -1.84922892 w_1 : -0.62814188 w_2 : 0.85846843

Logistic Regression



2 (a) If we consider only the case where $k = w_m$, we can express the log-likelihood as

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

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

If we look only at this case, then the corresponding part of $\nabla w_m l(w)$ is

$$\begin{aligned} & \sum_{i=1}^N \mathbb{I}\{y^{(i)} = m\} \frac{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]}{\exp[w_m^T \phi(x^{(i)})]} \left[\frac{\exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} - \frac{(\exp[w_m^T \phi(x^{(i)})])^2 \phi(x^{(i)})}{(\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})])^2} \right] \\ &= \sum_{i=1}^N \mathbb{I}\{y^{(i)} = m\} \left[\frac{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})] \exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{\exp[w_m^T \phi(x^{(i)})] \sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \right. \\ & \quad \left. - \frac{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})] (\exp[w_m^T \phi(x^{(i)})])^2 \phi(x^{(i)})}{\exp[w_m^T \phi(x^{(i)})] (\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})])^2} \right] \\ &= \sum_{i=1}^N \mathbb{I}\{y^{(i)} = m\} \left[\phi(x^{(i)}) - \frac{(\exp[w_m^T \phi(x^{(i)})]) \phi(x^{(i)})}{(\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})])} \right] \\ &= \sum_{i=1}^N \phi(x^{(i)}) \mathbb{I}\{y^{(i)} = m\} \left[1 - \frac{(\exp[w_m^T \phi(x^{(i)})])}{(\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})])} \right] \end{aligned}$$

Now, let us look at the case where $k \neq m$. Here, the corresponding part of $\nabla w_m l(w)$ is

$$\begin{aligned} & \sum_{i=1}^{k \neq m} \mathbb{I}\{y^{(i)} = k\} \frac{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]}{\exp[w_k^T \phi(x^{(i)})]} \left[- \frac{\exp[w_k^T \phi(x^{(i)})] \exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{[\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]]^2} \right] \\ &= - \sum_{i=1}^{k \neq m} \mathbb{I}\{y^{(i)} = k\} \left[\frac{\exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{[\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]]} \right] \end{aligned}$$

If we incorporate both cases, we get

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

$$\begin{aligned}
&= \sum_{i=1}^N \phi(x^{(i)}) \mathbb{I}\{y^{(i)} = m\} \left[1 - \frac{\exp[w_m^T \phi(x^{(i)})]}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \right] - [1 - \mathbb{I}\{y^{(i)} = m\}] \left[\frac{\exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \right] \\
&= \sum_{i=1}^N \mathbb{I}\{y^{(i)} = m\} \phi(x^{(i)}) - \frac{\mathbb{I}\{y^{(i)} = m\} \exp[w_m^T \phi(x^{(i)})] (\phi(x^{(i)}))}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \\
&\quad - \left[\frac{\exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \right] + \frac{\mathbb{I}\{y^{(i)} = m\} \exp[w_m^T \phi(x^{(i)})] \phi(x^{(i)})}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \\
&= \sum_{i=1}^N \phi(x^{(i)}) [\mathbb{I}\{y^{(i)} = m\} - \left[\frac{\exp[w_m^T \phi(x^{(i)})]}{\sum_{j=1}^K \exp[w_j^T \phi(x^{(i)})]} \right]]
\end{aligned}$$

Because w_K is fixed at 0, we can write this as

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

So the update rule is

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

2 (c) Accuracy on the test set is 94%.

3 (a) With Bayes' theorem, we can express the posterior probability of the Gaussian discriminant analysis for $y = 1$ as

$$\begin{aligned}
 p(y = 1|x; \phi, \Sigma, m_0, m_1) &= \frac{p(x|y = 1)p(y = 1)}{p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)} \\
 &= \frac{\phi \frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]}{\phi \frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)] + (1 - \phi) \frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)]} \\
 &= \frac{\phi \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]}{\phi \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)] + (1 - \phi) \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)]} \\
 &= \frac{1}{1 + \frac{(1 - \phi) \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)]}{\phi \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]}}
 \end{aligned}$$

Consider

$$\begin{aligned}
 &\frac{(1 - \phi) \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)]}{\phi \exp[-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)]} \\
 &= \frac{1 - \phi}{\phi} \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)] \\
 &= \frac{1 - \phi}{\phi} \exp[-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)] \\
 &= \frac{1 - \phi}{\phi} \exp[x^T \Sigma^{-1} \mu_0 - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 - x^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1] \\
 &= \exp[\log(\frac{1 - \phi}{\phi})] \exp[x^T \Sigma^{-1} \mu_0 - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 - x^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1] \\
 &= \exp[\log(\frac{1 - \phi}{\phi}) + x^T \Sigma^{-1} \mu_0 - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 - x^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1] \\
 &= \exp[-(x^T (\Sigma^{-1} \mu_1 - \Sigma^{-1} \mu_0) + \frac{1}{2} (\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log(\frac{1 - \phi}{\phi}))]
 \end{aligned}$$

By incorporating this into our expression for $p(y = 1|x; \phi, \Sigma, m_0, m_1)$ above, we get

$$\frac{1}{1 + \exp[-(x^T (\Sigma^{-1} \mu_1 - \Sigma^{-1} \mu_0) + \frac{1}{2} (\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log(\frac{1 - \phi}{\phi}))]}$$

If we let w represent a function of ϕ, Σ, μ_0 , and μ_1 , then

$$p(y = 1|x; \phi, \Sigma, m_0, m_1) = \frac{1}{1 + \exp[-w^T x]} \quad \square$$

3 (b) $\ell(\phi, m_0, m_1, \Sigma) =$

$$\sum_{i=1}^N \log \left[\frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right] \phi^{y^{(i)}} (1 - \phi)^{1-y^{(i)}} \right]$$

$$\frac{\partial}{\partial \phi} \ell(\phi, \mu_0, \mu_1, \Sigma) =$$

$$\sum_{i=1}^N \frac{1}{\frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right] \phi^{y^{(i)}} (1 - \phi)^{1-y^{(i)}}}$$

$$\frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right]$$

$$[(1 - \phi)^{1-y^{(i)}} y^{(i)} \phi^{y^{(i)}-1} - \phi^{y^{(i)}} (1 - y^{(i)}) (1 - \phi)^{-y^{(i)}}]$$

$$= \sum_{i=1}^N y^{(i)} \phi^{-1} - \frac{(1 - y^{(i)}) (1 - \phi)^{-y^{(i)}}}{(1 - \phi)^{1-y^{(i)}}}$$

$$= \sum_{i=1}^N \frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi}$$

$$= \sum_{i=1}^N \frac{y^{(i)} (1 - \phi)}{\phi (1 - \phi)} - \frac{\phi - y^{(i)} \phi}{\phi (1 - \phi)}$$

$$0 = \sum_{i=1}^N \frac{y^{(i)}}{\phi (1 - \phi)} - \frac{\phi}{\phi (1 - \phi)}$$

$$N\phi = \sum_{i=1}^N y^{(i)}$$

$$\phi = \frac{1}{N} \sum_{i=1}^N \mathbb{I}\{y^{(i)} = 1\} \quad \square$$

$$\begin{aligned}
& \frac{\partial}{\partial \mu_0} \ell(\phi, \mu_0, \mu_1, \Sigma) = \\
& \sum_{i=1}^N \frac{1}{\frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x^{(i)} - \mu_{y(i)})^T \Sigma^{-1} (x^{(i)} - \mu_{y(i)})] \phi^{y(i)} (1 - \phi)^{1-y(i)}} \\
& \frac{1}{2\pi^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} \exp[-\frac{1}{2}(x^{(i)} - \mu_{y(i)})^T \Sigma^{-1} (x^{(i)} - \mu_{y(i)})] \phi^{y(i)} (1 - \phi)^{1-y(i)} \\
& \frac{\partial}{\partial \mu_{y(i)}} - \frac{1}{2} (x^{(i)} - \mu_{y(i)})^T \Sigma^{-1} (x^{(i)} - \mu_{y(i)}) \\
& = \frac{\partial}{\partial \mu_{y(i)}} \sum_{i=1}^N -\frac{1}{2} (x^{(i)} - \mu_{y(i)})^T \Sigma^{-1} (x^{(i)} - \mu_{y(i)}) \\
& = \frac{\partial}{\partial \mu_{y(i)}} \sum_{i=1}^N -\frac{1}{2} [x^{(i)T} \Sigma^{-1} x^{(i)} - x^{(i)T} \Sigma^{-1} \mu_{y(i)} - \mu_{y(i)}^T \Sigma^{-1} x^{(i)} + \mu_{y(i)}^T \Sigma^{-1} \mu_{y(i)}] \\
& \frac{\partial}{\partial \mu_0} = 0 = \sum_{i=1}^N \mathbb{I}\{y^{(i)} = 0\} x^{(i)T} \Sigma^{-1} - \mathbb{I}\{y^{(i)} = 0\} \mu_{y(i)}^T \Sigma^{-1} \\
& \sum_{i=1}^N \mathbb{I}\{y^{(i)} = 0\} \mu_0^T = \sum_{i=1}^N \mathbb{I}\{y^{(i)} = 0\} x^{(i)T} \\
& \mu_0 = \frac{\sum_{i=1}^N \mathbb{I}\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^N \mathbb{I}\{y^{(i)} = 0\}} \square
\end{aligned}$$

Similarly,

$$\mu_1 = \frac{\sum_{i=1}^N \mathbb{I}\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^N \mathbb{I}\{y^{(i)} = 1\}} \square$$