1 [21 points] Logistic regression

(a) **Answer[8 points]:** (Note we do things in a slightly shorter way here; this solution does not use the hint.) Recall that we have $g'(z) = g(z)(1 - g(z))$ where $g(z) = \sigma(z)$, and thus for $h(\mathbf{x}) = g(\mathbf{w}^T\mathbf{x})$, we have $\frac{\partial h(\mathbf{x})}{\partial \mathbf{w}_k} \partial h(\mathbf{x}) = h(\mathbf{x})(1 - h(\mathbf{x}))x_k$.

Remember we have shown in class:

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}_k} = \sum_{i=1}^{N} (y^{(i)} - h(\mathbf{x}^{(i)}))x_k^{(i)}$$

By taking second derivative, we get

$$\begin{aligned} H_{kl} &= \frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}_k \partial \mathbf{w}_l} \\ &= \sum_{i=1}^{N} -\frac{\partial h(\mathbf{x}^{(i)})}{\partial \mathbf{w}_l} x_k^{(i)} \\ &= \sum_{i=1}^{N} -h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)}))x_l^{(i)} x_k^{(i)} \end{aligned}$$

In a matrix form,

$$H = -\sum_{i=1}^{N} h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)}))\mathbf{x}^{(i)}\mathbf{x}^{(i)T}$$

To prove H is negative semidefinite, we show $\mathbf{z}^T H \mathbf{z} \leq 0$ for all $\mathbf{z}$.

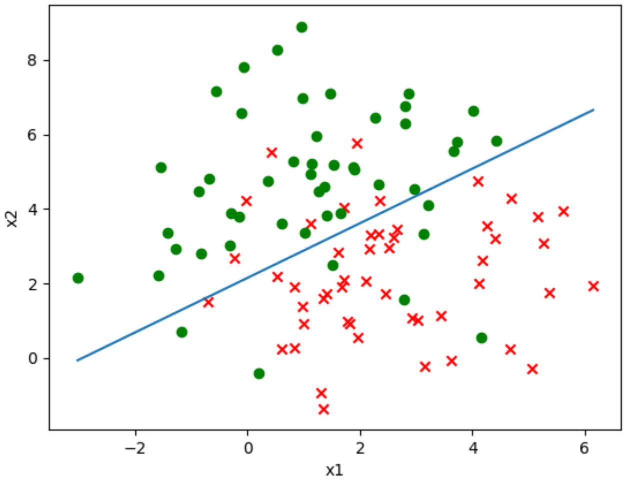
$$\begin{aligned} \mathbf{z}^T H \mathbf{z} &= \textbf{-}\mathbf{z}^T \left(\sum_{i=1}^{N} h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)}))\mathbf{x}^{(i)}\mathbf{x}^{(i)T} \right) \mathbf{z} \\ &= -\sum_{i=1}^{N} h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)}))\mathbf{z}^T\mathbf{x}^{(i)}\mathbf{x}^{(i)T}\mathbf{z} \\ &= -\sum_{i=1}^{N} h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)})) \left(\mathbf{z}^T\mathbf{x}^{(i)}\right)^2 \\ &\leq 0 \end{aligned}$$

with the last inequality holding, since $0 \leq h(\mathbf{x}^{(i)}) \leq 1$, which implies $h(\mathbf{x}^{(i)})(1 - h(\mathbf{x}^{(i)})) \geq 0$, and $(\mathbf{z}^T\mathbf{x}^{(i)})^2) \geq 0$.

(b) **Answer[8 points]:** $\mathbf{w} = (-1.8492, -0.6281, 0.8585)$ with the first entry corresponding to the intercept term.

See attached `q1_sol.py`.

(c) **Answer[5 points]:** As shown in the figure, the data sample $x^{(i)}$ with label $y^{(i)} = 0$ is plotted as red cross, and the data sample $x^{(i)}$ with label $y^{(i)} = 1$ is plotted as green dot.



8
6
4
2
0
x2
−2
0
2
4
6
x1

2 [27 points] Softmax Regression via Gradient Ascent

(a) **[13 points]** Derive the gradient ascent update rule for the log-likelihood of the training data.

We have:

$$l(\mathbf{w}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \log \left[p(y^{(i)} = k|\mathbf{x}^{(i)}, \mathbf{w})\right]^{\mathbf{I}(y^{(i)}=k)}$$

Taking gradient with respect to $\mathbf{w}_m$ ($p \leq m \leq K-1$):

$$\begin{aligned}
\nabla_{\mathbf{w}_m} l(\mathbf{w}) &= \nabla_{\mathbf{w}_m} \sum_{i=1}^{N}\sum_{k=1}^{K} \log \left[p(y^{(i)} = k|\mathbf{x}^{(i)}, \mathbf{w})\right]^{\mathbf{I}(y^{(i)}=k)} \\
&= \nabla_{\mathbf{w}_m} \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \log \left[p(y^{(i)} = k|\mathbf{x}^{(i)}, \mathbf{w})\right] \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \log \left[p(y^{(i)} = k|\mathbf{x}^{(i)}, \mathbf{w})\right] \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \Big[\log\Big(\frac{\exp(\mathbf{w}_k^T \phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))}\Big)\Big] \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \Big[\log\Big(\exp(\mathbf{w}_k^T \phi(\mathbf{x}^{(i)}))\Big) - \log\Big(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))\Big)\Big] \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \Big[\mathbf{w}_k^T \phi(\mathbf{x}^{(i)}) - \log\Big(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))\Big)\Big] \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \Big(\Big[\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)})\Big] - \Big[\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \log\big(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))\big)\Big]\Big)
\end{aligned}$$

As the log term on the right does not contain k, it can be taken out of the summation and since $\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) = 1$, we obtain the following:

$$\begin{aligned}
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \Big(\Big[\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)})\Big] - \log\Big(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))\Big)\Big) \\
&= \sum_{i=1}^{N} \nabla_{\mathbf{w}_m} \Big[\sum_{k=1}^{K} \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)})\Big] - \nabla_{\mathbf{w}_m} \log\Big(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))\Big)
\end{aligned}$$

The left term contains $\mathbf{w}_m$ iff $m = y^{(i)}$. Second term contains $\mathbf{w}_m$ (produced by summation)

$$\begin{aligned}
&= \sum_{i=1}^{N} \mathbf{I}(y^{(i)} = m)\phi(\mathbf{x}^{(i)}) - \frac{\exp(\mathbf{w}_m^T \phi(\mathbf{x}^{(i)}))\phi(\mathbf{x}^{(i)})}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))} \\
&= \sum_{i=1}^{N} \phi(\mathbf{x}^{(i)})[\mathbf{I}(y^{(i)} = m) - p(y^{(i)} = m|\mathbf{x}^{(i)})]
\end{aligned}$$

(b) See `hw2.py`. Instructor solution achieves 92.0% accuracy; students should be able to get an accuracy above 90%. SciKit-Learn gets an accuracy of 92-94%, depending on the version.

3 [22 points] Gaussian Discriminate Analysis

(a) **[8 points]**

Note, parameters can be omitted.

$p(y=1 \mid \mathbf{x}; \phi, \Sigma, \mu_0, \mu_1) = p(y=1 \mid \mathbf{x})$
$p(\mathbf{x} \mid y=1; \phi, \Sigma, \mu_0, \mu_1) = p(\mathbf{x} \mid y=1)$
$p(\mathbf{x}=1; \phi, \Sigma, \mu_0, \mu_1) = p(\mathbf{x}=1)$
$p(y=1; \phi, \Sigma, \mu_0, \mu_1) = p(y=1)$

Now,

$$
\begin{aligned}
p(y=1 \mid \mathbf{x}; \phi, \Sigma, \mu_0, \mu_1) &= p(y=1 \mid \mathbf{x}) \\
&= \frac{p(\mathbf{x} \mid y=1)p(y=1)}{p(\mathbf{x})} \\
&= \frac{p(\mathbf{x} \mid y=1)p(y=1)}{p(\mathbf{x} \mid y=1)p(y=1) + p(\mathbf{x} \mid y=0)p(y=0)} \\
&= \frac{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \Sigma^{-1}\left(\mathbf{x}-\mu_1\right)\right)\phi}{\exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \Sigma^{-1}\left(\mathbf{x}-\mu_1\right)\right)\phi + \exp\left(-\frac{1}{2}\left(\mathbf{x}-\mu_0\right)^T \Sigma^{-1}\left(\mathbf{x}-\mu_0\right)\right)(1-\phi)} \\
&= \frac{1}{1+\exp\left(\log\left(\frac{(1-\phi)}{\phi}\right) - \frac{1}{2}\left(\mathbf{x}-\mu_0\right)^T \Sigma^{-1}\left(\mathbf{x}-\mu_0\right) + \frac{1}{2}\left(\mathbf{x}-\mu_1\right)^T \Sigma^{-1}\left(\mathbf{x}-\mu_1\right)\right)} \\
&= \frac{1}{1+\exp\left(\log\left(\frac{1-\phi}{\phi}\right) + \mathbf{x}^T\Sigma^{-1}\mu_0 - \mathbf{x}^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1\right)} \\
&= \frac{1}{1+\exp\left(\log\left(\frac{1-\phi}{\phi}\right) + \mathbf{x}^T\Sigma^{-1}\left(\mu_0-\mu_1\right) - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \mu_1^T\Sigma^{-1}\mu_1\right)}
\end{aligned}
$$

By setting

$\mathbf{w}_0 = \frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right) - \log\frac{1-\phi}{\phi}$,
$\mathbf{w} = -\Sigma^{-1}\left(\mu_1-\mu_0\right)$,
A constant intercept term $\mathbf{x}_0 = 1$,

We get: $p(y=1 \mid \mathbf{x}) = \frac{1}{1+\exp(-(\mathbf{w}^T\mathbf{x}+\mathbf{w}_0\mathbf{x}_0))} = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$.

(b) **[8 points]** Question (b) is the special case of (c) with $M=1$. Let us derive the general case directly:

$$
\begin{aligned}
\ell\left(\phi, \mu_0, \mu_1, \Sigma\right) &= \log \prod_{i=1}^{N} p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{N} \log p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{N} \log p(y^{(i)}; \phi) \\
&= \sum_{i=1}^{N} \left[\log \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} - \frac{1}{2}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right) + \log \phi^{y^{(i)}} + \log\left(1-\phi\right)^{\left(1-y^{(i)}\right)}\right] \\
&\simeq \sum_{i=1}^{N} \left[-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right) + y^{(i)} \log \phi + \left(1 - y^{(i)}\right) \log\left(1-\phi\right)\right]
\end{aligned}
$$

(the constant term is independent of the parameters, thus removed.)

Then, the likelihood is maximized by setting the derivative with respect to each parameter to zero:

(1) with respect to ϕ:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \sum_{i=1}^{N} (\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1-\phi}) \\
&= \sum_{i=1}^{N} \frac{1(y^{(i)} = 1)}{\phi} + \frac{N - \sum_{i=1}^{N} 1(y^{(i)} = 1)}{1 - \phi}
\end{aligned}
$$

Therefore, $\phi = \frac{1}{N}\sum_{i=1}^{N} 1(y^{(i)} = 1)$, i.e. the percentage of the training examples such that $y^{(i)} = 1$.

(2) with respect to μ_0:

$$
\begin{aligned}
\nabla_{\mu_0} \ell &= -\frac{1}{2} \sum_{i:y^{(i)}=0} \nabla_{\mu_0} \left(\mathbf{x}^{(i)} - \mu_0\right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_0\right) \\
&= -\frac{1}{2} \sum_{i:y^{(i)}=0} \nabla_{\mu_0} \left[-2\mu_0^T \Sigma^{-1} \mathbf{x}^{(i)} + \mu_0^T \Sigma^{-1} \mu_0\right] \\
&= -\frac{1}{2} \sum_{i:y^{(i)}=0} \left[-2\Sigma^{-1}\mathbf{x}^{(i)} + 2\Sigma^{-1}\mu_0\right]
\end{aligned}
$$

By setting the gradient to zero,

$$
\begin{aligned}
\sum_{i:y^{(i)}=0} \left[\Sigma^{-1}\mathbf{x}^{(i)} - \Sigma^{-1}\mu_0\right] &= 0 \\
\sum_{i=1}^{N} 1\left\{y^{(i)} = 0\right\} \Sigma^{-1}\mathbf{x}^{(i)} - \sum_{i=1}^{N} 1\left\{y^{(i)} = 0\right\} \Sigma^{-1}\mu_0 &= 0
\end{aligned}
$$

Thus we obtain $\mu_0 = \frac{\sum_{i=1}^{N} 1\left\{y^{(i)}=0\right\}\mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\left\{y^{(i)}=0\right\}}$

(3) with respect to μ_1:

The calculations are similar for μ_1. The resulting maximum likelihood estimate is: $\mu_1 = \frac{\sum_{i=1}^{N} 1\left\{y^{(i)}=1\right\}\mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\left\{y^{(i)}=1\right\}}$

(4) with respect to Σ:
The last step is to calculate the gradient with respect to Σ. Here, we assume $M = 1$, i.e., $|\Sigma| = \sigma^2$.
The log-likelihood of the data then can be written:

$$\begin{aligned}
\ell\left(\phi, \mu_0, \mu_1, \Sigma\right) &\simeq \sum_{i=1}^{N}\left[-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T \Sigma^{-1}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right) + y^{(i)}\log\phi + \left(1 - y^{(i)}\right)\log\left(1-\phi\right)\right] \\
&= \sum_{i=1}^{N}\left[-\log\sigma - \frac{1}{2\sigma^2}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right) + y^{(i)}\log\phi + \left(1 - y^{(i)}\right)\log\left(1-\phi\right)\right]
\end{aligned}$$

By taking derivative with respect to σ and set it to zero:

$$\nabla_\sigma \ell = \sum_{i=1}^{N}\left[-\frac{1}{\sigma} + \frac{1}{\sigma^3}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)\right] = 0$$

You obtain: $\Sigma = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)^T\left(\mathbf{x}^{(i)} - \mu_{y^{(i)}}\right)$

(c) **[6 points]** Elaborated above.

4 [30 points] Logistic regression

See `q4.py` for instructor solution.

(a) A correct NB implementation achieves 1.625% accuracy exactly.

(b) Top 5 spam tokens are `['httpaddr' 'spam' 'unsubscrib' 'ebai' 'valet']`.

(c) These should be the accuracy breakdowns, with the training set of size 1400 giving the best accuracy.

- Training set size 50: Test set error = 3.875%
- Training set size 100: Test set error = 2.625%
- Training set size 200: Test set error = 2.625%
- Training set size 400: Test set error = 1.875%
- Training set size 800: Test set error = 1.75%
- Training set size 1400: Test set error = 1.625%