This handout includes space for every question that requires a written response. Please feel free to use it to handwrite your solutions (legibly, please). If you choose to typeset your solutions, the `README.md` for this assignment includes instructions to regenerate this handout with your typeset LaTeX solutions.

1.a

Since

$$g'(z) = g(z)(1 - g(z)) \text{ and}$$
$$h(x) = g(\theta^T x),$$

it follows that

$$\partial h(x)/\partial \theta_k = h(x)(1 - h(x))x_k.$$

Letting

$$h_\theta(x^{(i)}) = g(\theta^T x^{(i)}) = 1/(1 + \exp(-\theta^T x^{(i)})),$$

we have

$$\begin{aligned}
\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} &= \frac{1}{h(\theta^T x)} \times h(x)(1 - h(x))x_k \\
&= (1 - h(x))x^{(i)} \\
\frac{\partial \log(1 - h_\theta(x^{(i)}))}{\partial \theta_k} &= \frac{1}{1 - h(\theta^T x)} \times -1 \times h(x)(1 - h(x))x_k \\
&= -h_\theta(x)x^{(i)}
\end{aligned}$$

recalling that

$$J(\theta) = \sum_{i=1}^{n} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

Substituting into our equation for $J(\theta)$, we have

$$\begin{aligned}
\frac{\partial J(\theta)}{\partial \theta_k} &= \frac{1}{n} \frac{\partial \sum_{i=1}^{n} y^{(i)} log(h_\theta(x^{(i)})) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))}{\partial \theta} \\
&= \frac{1}{n} \sum_{i=1}^{n} y^{(i)}(1 - h(x))x^{(i)} + (1 - y^{(i)}) \times -h_\theta(x)x^{(i)} \\
&= \frac{1}{n} \sum_{i=1}^{n} y^{(i)}x^{(i)} - h(x)x^{(i)} \\
&= \frac{1}{n} \sum_{i=1}^{n} x^{(i)}(y^{(i)} - h(x))
\end{aligned}$$

Consequently, the (k, l) entry of the Hessian is given by

$$H_{kl} = \frac{\partial^2 J(\theta)}{\partial \theta_k \partial \theta_l} =$$

Using the fact that $X_{ij} = x_i x_j$ if and only if $X = xx^T$, we have

$$H =$$

To prove that H is positive semi-definite, show $z^T H z \geq 0$ for all $z \in \mathbb{R}^d$.

$$z^T H z =$$

1.c

For shorthand, we let $\mathcal{H} = \{\phi, \Sigma, \mu_0, \mu_1\}$ denote the parameters for the problem. Since the given formulae are conditioned on y, use Bayes rule to get:

$$
\begin{aligned}
p(y=1|x;\mathcal{H}) &= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x;\mathcal{H})} \\
&= \frac{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H}) + p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})}
\end{aligned}
$$

First note that

$$
\begin{aligned}
\frac{A}{A+B} &= \frac{1}{\frac{B+1}{A}} \\
&= \frac{1}{1+\frac{B}{A}}
\end{aligned}
$$

Now letting $A = p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})$
and $B = p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})$
We can continue as

$$
= \frac{1}{1+\frac{p(x|y=0;\mathcal{H})p(y=0;\mathcal{H})}{p(x|y=1;\mathcal{H})p(y=1;\mathcal{H})}}
$$

Noting that $p(y=1;\mathcal{H}) = \phi$ and $p(y=0;\mathcal{H}) = 1-\phi$

$$
= \frac{1}{1+\frac{p(x|y=0;\mathcal{H})(1-\phi)}{p(x|y=1;\mathcal{H})\phi}}
$$

and noting that

$$
p(x|y=i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}})} exp(-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i))
$$

the $\frac{1}{(2\pi)^{\frac{d}{2}}}$ terms will cancel leaving

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

1.d

First, derive the expression for the log-likelihood of the training data:

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma)p(y^{(i)}; \phi) \tag{1}$$

$$= \sum_{i=1}^{n} \log p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi) \tag{2}$$

$$= \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}})} exp(-\frac{1}{2}(x^{(i)} - \mu_i)^T\Sigma^{-1}(x^{(i)} - \mu_i)) + \sum_{i=1}^{n} \log p(y^{(i)}; \phi) \tag{3}$$

$$= \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}})} exp(-\frac{1}{2}(x^{(i)} - \mu_i)^T\Sigma^{-1}(x^{(i)} - \mu_i)) + \sum_{i=1}^{n} \log \phi^{y^{(i)}}(1-\phi)^{(1-y^{(i)})} \tag{4}$$

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

For ϕ:
Let n_i be the number of y values equal to i, for $i \in 0, 1$

$$\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \frac{n_1 \partial log(\phi)}{\partial \phi} + \frac{n_0 \partial log(1-\phi)}{\partial \phi} \\
&= \frac{n_1}{n\phi} - \frac{n_0}{n(1-\phi)} \\
&= \frac{n_1(1-\phi)}{n\phi(1-\phi)} - \frac{n_0\phi}{n\phi(1-\phi)} \\
&= \frac{n_1}{n\phi} - \frac{n_0}{n(1-\phi)} \\
&= \frac{n_1(1-\phi) - n_0\phi}{n\phi(1-\phi)}
\end{aligned}$$

Assuming ϕ is not 0, then this is zero when

$$\begin{aligned}
n_1(1-\phi) &= n_0\phi \\
n_1 - n_1\phi &= n_0\phi \\
n_1 &= \phi(n_1 + n_0) \\
\frac{n_1}{n_1 + n_0} &= \phi \\
n_1 &= \phi
\end{aligned}$$

Setting this equal to zero and solving for ϕ gives the maximum likelihood estimate.

For μ_0:
Hint: Remember that Σ (and thus Σ^{-1}) is symmetric.

$$\begin{aligned}
\nabla_{\mu_0}\ell &= \nabla_{\mu_0} \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}})} exp(-\frac{1}{2}(x^{(i)} - \mu_i)^T\Sigma^{-1}(x^{(i)} - \mu_i)) + \sum_{i=1}^{n} \log \phi^{y^{(i)}}(1-\phi)^{(1-y^{(i)})} \\
&= \nabla_{\mu_0} \sum_{i=1}^{n} \log \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}})} exp(-\frac{1}{2}(x^{(i)} - \mu_i)^T\Sigma^{-1}(x^{(i)} - \mu_i)) \\
&= K\nabla_{\mu_0} \sum_{i=1}^{n} -\frac{1}{2}(x^{(i)} - \mu_0)^T\Sigma^{-1}(x^{(i)} - \mu_0)) \text{ for the } y^{(i)} = 0 \text{ and some constant } K
\end{aligned}$$

Now my matrix calculus is rusty but I know we will get terms that look like

$$= \nabla_{\mu_0} \sum_{i=1}^{n} \frac{1}{2} x^2 - 2x^{(i)}\mu_0 + \mu_0^2$$

and taking the gradient wrt μ_0 we will get

$$= \sum_{i=1}^{n} -x^{(i)} + \mu_0 \text{ for the } y(i) = 0$$

and setting this to 0 will yield something like

$$\mu_0 = \frac{\sum x^{(i)} \text{ where } y^{(i)} = 0}{n_0}$$

where n_0 is the number of $y^{(i)} = 0$

Setting this gradient to zero gives the maximum likelihood estimate for μ_0.

For μ_1:

Hint: Remember that Σ (and thus Σ^{-1}) is symmetric. Similar to above

$$\mu_1 = \frac{\sum x^{(i)} \text{ where } y^{(i)} = 1}{n_1}$$

where n_1 is the number of $y^{(i)} = 1$

Setting this gradient to zero gives the maximum likelihood estimate for μ_1.

For Σ, we find the gradient with respect to $S = \Sigma^{-1}$ rather than Σ just to simplify the derivation (note that $|S| = \frac{1}{|\Sigma|}$). You should convince yourself that the maximum likelihood estimate S_n found in this way would correspond to the actual maximum likelihood estimate Σ_n as $S_n^{-1} = \Sigma_n$.

Hint: You may need the following identities:

$$\nabla_S |S| = |S|(S^{-1})^T$$

$$\nabla_S b_i^T S b_i = \nabla_S tr\left(b_i^T S b_i\right) = \nabla_S tr\left(S b_i b_i^T\right) = b_i b_i^T$$

$$\nabla_S \ell =$$

Next, substitute $\Sigma = S^{-1}$. Setting this gradient to zero gives the required maximum likelihood estimate for Σ.

1.f

1.g

1.h

2.c

2.d

2.e