1 (a) Prove that $w_0 = \bar{Y} - w_1\bar{X}$:

Let

$$f'(g(w_0)) = \frac{\partial L}{\partial \sum_{i=1}^{N}[y^{(i)} - (w_0 + w_1x^{(i)})]}$$

$$g'(w_0) = \frac{\partial \sum_{i=1}^{N}[y^{(i)} - (w_0 + w_1x^{(i)})]}{\partial w_0}$$

then

$$\frac{\partial L}{\partial w_0} = f'(g(w_0)) \cdot g'(w_0)$$

$$f'(g(w_0)) = \sum_{i=1}^{N}[y^{(i)} - w_0 - w_1x^{(i)}] = N\bar{Y} - Nw_0 - Nw_1\bar{X}$$

$$g'(w_0) = \sum_{i=1}^{N} -1 = -N$$

$$\frac{\partial L}{\partial w_0} = -N(N\bar{Y} - Nw_0 - Nw_1\bar{X}) = -N^2\bar{Y} + N^2w_0 + N^2w_1\bar{X}$$

$$0 = -N^2\bar{Y} + N^2w_0 + N^2w_1\bar{X}$$

$$w_0 = \bar{Y} - w_1\bar{X} \ \square$$

Prove that

$$w_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}y^{(i)} - \bar{Y}\bar{X}}{\frac{1}{N}\sum_{i=1}^{N}(x^{(i)})^2 - \bar{X}^2}$$

By the chain rule,

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{N}[-x^{(i)}(y^{(i)} - (w_0 + w_1x^{(i)}))]$$

Setting the partial derivative to 0, we get

$$0 = -\sum_{i=1}^{N} x^{(i)}y^{(i)} + \sum_{i=1}^{N} w_0x^{(i)} + \sum_{i=1}^{N}[w_1(x^{(i)})^2]$$

Since $w_0 = \bar{Y} - w_1\bar{X}$

$$0 = -\sum_{i=1}^{N} x^{(i)}y^{(i)} + \sum_{i=1}^{N}[(\bar{Y} - w_1\bar{X})x^{(i)}] + \sum_{i=1}^{N}[w_1(x^{(i)})^2]$$

$$= -\sum_{i=1}^{N} x^{(i)}y^{(i)} + \bar{Y}\sum_{i=1}^{N} x^{(i)} - \bar{X}w_1\sum_{i=1}^{N} x^{(i)} + \sum_{i=1}^{N}[w_1(x^{(i)})^2]$$

$$w_1[\bar{X}\sum_{i=1}^{N} x^{(i)} - \sum_{i=1}^{N}(x^{(i)})^2] = -\sum_{i=1}^{N} x^{(i)}y^{(i)} + \bar{Y}\sum_{i=1}^{N} x^{(i)}$$

$$w_1 = \frac{-\sum_{i=1}^{N} x^{(i)}y^{(i)} + \bar{Y}\sum_{i=1}^{N} x^{(i)}}{\bar{X}\sum_{i=1}^{N} x^{(i)} - \sum_{i=1}^{N}(x^{(i)})^2}$$

$$w_1 = \frac{\sum_{i=1}^{N} x^{(i)}y^{(i)} - N\bar{Y}\bar{X}}{\sum_{i=1}^{N}(x^{(i)})^2 - N\bar{X}^2}$$

$$w_1 = \frac{\frac{1}{N}\sum_{i=1}^{N} x^{(i)}y^{(i)} - \bar{Y}\bar{X}}{\frac{1}{N}\sum_{i=1}^{N}(x^{(i)})^2 - \bar{X}^2} \square$$

1 (b) i. Let us first show that if $\lambda_i > 0$ for all i, then A must be PD.

For any $z \neq 0 \in \mathbb{R}^d, z^T Az = z^T(U\Lambda U^T)z$. Let $y = U^T z$. Then

$$z^T Az = y^T \Lambda y = y_1^2\lambda_1 + y_2^2\lambda_2 + ... + y_d^2\lambda_d$$

Since U is an orthogonal matrix, no row or column of U can consist entirely of zeros, since each row and column must have a norm of 1. The entries of y can be written as

$$y_1 = u_1^T z, y_2 = u_2^T z, ..., y_d = u_d^T z$$

Since $z \neq 0$, then for at least one $i = \{1, 2, ..., d\}$, $y_i \neq 0$. We have assumed that for all i $\lambda_i > 0$. In the expression

$$y_1^2\lambda_1 + y_2^2\lambda_2 + ... + y_d^2\lambda_d$$

each term will be 0 if $y_i = 0$ and greater than 0 if $y_i \neq 0$. So

$$z^T Az > 0 \square$$

Now let us show that if A is PD then for all i $\lambda_i > 0$.
We have it that, for all values of i, $Au_i = \lambda_i u_i$, where u_i is a column of U. By multiplying both sides of the equation by u_i^T on the left, we get

$$u_i^T Au_i = u_i^T \lambda_i u_i$$

Because A is PD and $u_i \neq 0 \in \mathbb{R}^d$, $u_i^T A u_i > 0$. We can write $u_i^T \lambda_i u_i$ as

$$\lambda_i \sum_{j=1}^{d} u_{ji}^2$$

where u_{ji} is entry of U at the jth row and ith column. If, for any value of i, $\lambda_i = 0$, then

$$\lambda_i \sum_{j=1}^{d} u_{ji}^2 = 0 = u_i^T A u_i$$

which would contradict A being PD. Suppose instead that $\lambda_i < 0$. For all values of i and j, if $u_{ji} = 0$ then $u_{ji}^2 = 0$, and if $u_{ji} \neq 0$ then $u_{ji}^2 > 0$. Since the column vector u_i is orthonormal, for some value of j, $u_{ji}^2 > 0$. In this case,

$$\lambda_i \sum_{j=1}^{d} u_{ji}^2 = u_i^T A u_i < 0$$

which would also contradict A being PD. So it must be the case that if A is PD then, for all values of i,

$$\lambda_i > 0 \, \square$$

1 (b) ii. Let us start with the eigenvalues of $\Phi^T\Phi + \beta I$.
In effect, $\Phi^T\Phi + \beta I$ differs from $\Phi^T\Phi$ by having diagonal values shifted by β. So if, for $i = \{1, 2, ..., d\}$, the eigenvalues of $\Phi^T\Phi$ are λ_i, then the eigenvalues of $\Phi^T\Phi + \beta I$ are $\lambda_i + \beta$. We can see this in the following way. Let $A = \Phi^T\Phi$ and $B = \Phi^T\Phi + \beta I$, and μ_i expresses the eigenvalues of B. Let M be the diagonal matrix $diag(\mu_i)$. Analogously to $AU = U\Lambda$,

$$BU = UM$$

Since $B = A + \beta I$,

$$(A + \beta I)U = UM$$

$$AU + \beta U = UM$$

Since $AU = U\Lambda$,

$$U\Lambda + \beta U = UM$$

$$U\beta = UM - U\Lambda = U(M - \Lambda)$$

$$\beta = M - \Lambda \square$$

Therefore, the difference between the diagonal values of M and Λ is given by β, and the eigenvalues of B (that is, $\Phi^T\Phi + \beta I$) are given by $\lambda_i + \beta$.

Now let us show that A and B have the same eigenvectors. Let z be an eigenvector of B. Since the eigenvalues of B are given by $\lambda_i + \beta$, we can write

$$Bz = (\lambda_i + \beta)z = \lambda_i z + \beta z$$

$$Az + \beta I z - \beta I z = \lambda_i z$$

$$Az = \lambda_i z \square$$

By the definition of eigenvectors and eigenvalues, this means that z is also an eigenvector of A. If u_i is an eigenvector of $\Phi^T\Phi$, it is also an eigenvector of $\Phi^T\Phi + \beta I$.

To see that $\Phi^T\Phi + \beta I$ is PD if $\beta > 0$, we can first show that $\Phi^T\Phi$ is PSD. In general, for any matrix X, X^TX is PSD. For any vector $z \neq 0 \in \mathbb{R}^d$,

$$z^T(X^TX)z = (Xz)^T Xz = ||Xz||_2^2 \geq 0$$

From the proof in **1 (b) i.** we can see that for a PSD matrix, for all values of i, $\lambda_i \geq 0$. In this case, if $\beta > 0$,

$$\lambda_i + \beta > 0 \square$$

As we have seen, if all the eigenvalues of $\Phi^T\Phi + \beta I$ are positive, then $\Phi^T\Phi + \beta I$ is PD.

1 (c) We can write $\sum_{n=1}^{N} logP(y^{(n)}|x^{(n)})$ as

$$\sum_{n=1}^{N}\{\mathbb{I}(y^{(n)} = 1)logP(y^n = 1|x^{(n)}) + \mathbb{I}(y^{(n)} = -1)logP(y^{(n)} = -1|x^{(n)})\}$$

For the probabilities of the class labels {-1, 1}, we can use the standard sigmoid function for logistic regression $\sigma(w^Tx)$ and treat y as $y \in \{0, 1\}$:

$$\sum_{n=1}^{N}\{y^{(n)}log[\sigma(w^T\phi(x^{(n)}))] + (1 - y^{(n)})log[1 - \sigma(w^T\phi(x^{(n)}))]^{1-y^{(n)}}\}$$

and then set the partial derivative $f'(w)$ to 0, switching the expression's sign to make it a convex optimization problem:

$$0 = -\sum_{n=1}^{N}\{y^{(n)}\frac{1}{\sigma[w^T\phi(x^{(n)})]}\sigma[w^T\phi(x^{(n)})][1 - \sigma(w^T\phi(x^{(n)}))]\phi(x^{(n)})$$

$$+ (1 - y^{(n)}) \frac{1}{1 - \sigma[w^T \phi(x^{(n)})]} [-\sigma(w^T \phi(x^{(n)}))][1 - \sigma(w^T \phi(x^{(n)}))] \phi(x^{(n)})\}$$

$$0 = -\sum_{n=1}^{N} \{y^{(n)}[1 - \sigma(w^T \phi(x^{(n)}))]\phi(x^{(n)})$$

$$- (1 - y^{(n)})[\sigma(w^T \phi(x^{(n)}))]\phi(x^{(n)})\}$$

$$0 = -\sum_{n=1}^{N} \{[y^{(n)} - \sigma(w^T \phi(x^{(n)}))]\phi(x^{(n)})\}$$

By setting the partial derivative of expression (4) with respect to w to 0 we get

$$0 = -\sum_{n=1}^{N} \{ \frac{exp[-yw^T \phi(x^{(n)})]}{1 + exp[-yw^T \phi(x^{(n)})]} y\phi(x^{(n)})\}$$

$$0 = -\sum_{n=1}^{N} \{ \frac{1}{exp[yw^T \phi(x^{(n)})] + 1} y\phi(x^{(n)})\}$$

If the derivative equation for the negative log likelihood above has $y = 0$, then the equation becomes

$$0 = -\sum_{n=1}^{N} \{[-\sigma(w^T \phi(x^{(n)}))]\phi(x^{(n)})\}$$

which is also what we get from the partial derivative equation for the loss function we derived from (4) if $y = -1$. So maximizing the partial derivative of the log-likelihood function with respect to w is equivalent to minimizing the partial derivative of loss function (4) with respect to w when $y = -1$. If $y = 1$, then the derivative equation for the negative log likelihood becomes

$$0 = -\sum_{n=1}^{N} \{[1 - \sigma(w^T \phi(x^{(n)}))]\phi(x^{(n)})\}$$

Since $1 - \sigma(w^T \phi(x^{(n)})) = \frac{exp[-w^T \phi(x^{(n)}))]}{1+exp[-w^T \phi(x^{(n)}))]}$

$$0 = -\sum_{n=1}^{N} \{ \frac{exp[-w^T \phi(x^{(n)}))]}{1 + exp[-w^T \phi(x^{(n)}))]} \phi(x^{(n)})\}$$

We can use the same expression the partial derivative equation for the loss function in the same way if $y = 1$:

$$0 = -\sum_{n=1}^{N} \{ \frac{1}{exp[w^T \phi(x^{(n)})] + 1} \phi(x^{(n)})\} = -\sum_{n=1}^{N} \{ \frac{exp[-w^T \phi(x^{(n)}))]}{1 + exp[-w^T \phi(x^{(n)}))]} \phi(x^{(n)})\}$$

Thefore, since w has the same value at the maximum of the log-likelihood function and the minimum of loss function (4) for all values of y, maximizing the log-likelihood is equivalent to minimizing this loss function.