Question 1-1

If the number of training samples (n) is less than the number of input dimensions plus one ($D+1$) (because of the bias term), $\mathbf{X}\mathbf{X}^\intercal$ is non-invertible: $Rank(\mathbf{X}) = Rank(\mathbf{X}\mathbf{X}^\intercal) \leq n < D+1$

Question 1-2

$RSS = \sum_n (y_n - (b + \sum_d w_d x_{nd}))^2$
$\frac{\partial RSS}{\partial b} = 0 \Rightarrow \sum_n y_n = N * b + \sum_d w_d x_{nd}$
Because of the preprocessing step, which centers the data along each dimension, we can assume that $\sum_d w_d x_{nd} = 0$, because for all ds we have $\sum_n x_{nd} = 0$, therefore: $b = 1/N \sum_n y_n$

Question 2-1

The optimal classifier will be a constant value: $p(y = 1) = \sigma(b)$, because we do not have the x features. The objective function will be:
$\varepsilon(b) = \min_b - \sum_n (y_n log(\sigma(b)) + (1 - y_n) log(1 - \sigma(b)))$
$\frac{\partial \varepsilon(b)}{\partial b} = 0 \Rightarrow - \sum_n y_n (1 - \sigma(b)) - (1 - y_n)\sigma(b) = 0 \Rightarrow \sum_n y_n - \sigma(b) = 0 \Rightarrow \sigma(b) = 1/N \sum_n y_n \Rightarrow$
$b = \sigma^{-1}(1/N \sum_n y_n) = \frac{log(1/N \sum_n y_n)}{1 - log(1/N \sum_n y_n)}$