1(a).

1(a).

$$l(w) = \sum_{i=1}^{N} y^{(i)} \log h(x^{(i)}) + (1-y^{(i)}) \log(1-h(x^{(i)}))$$

$$\nabla_w l = \sum_{i=1}^{N} (y^{(i)} - \sigma(w^T x^{(i)}))\, x^{(i)}$$

$$\frac{\partial l^2}{\partial^2 w} = \sum_{i=1}^{N} -\sigma(w^T x^{(i)})(1-\sigma(w^T x^{(i)}))\, x^{(i)} \cdot x^{(i)}$$, since $\sigma' = \sigma(1-\sigma)$

$$= -\sum_{i=1}^{N} x^{(i)} \sigma(w^T x^{(i)})(1-\sigma(w^T x^{(i)}))\, x^{(i)}$$

So the Hession matrix can be represented as:

$$H = -X^T \Sigma X, \quad \Sigma = \mathrm{diag}(\sigma^{(n)}(1-\sigma^{(n)})) = \begin{pmatrix} \sigma(w^T x^{(1)})(1-\sigma(w^T x^{(1)})) & & 0 \\ & \ddots & \\ 0 & & \sigma(w^T x^{(m)})(1-\sigma(w^T x^{(m)})) \end{pmatrix}$$

$$\forall z, \quad z^T H z = -z^T X^T \Sigma X z$$

$$= -(Xz)^T \Sigma (Xz) = -\sum_{i=1}^{N} \sigma^{(i)}(1-\sigma^{(i)})(X^{(i)} z)^2$$

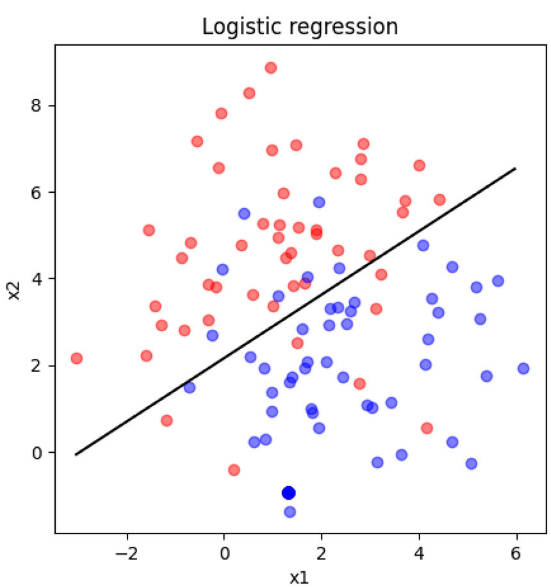
Since $\sigma(x) = \frac{1}{1+\exp(-x)} \in (0,1)$, $z^T H z \le 0$, negative semi-definite.

1(b).

The coefficients w resulting from my fit is [-1.84922892, -0.62814188, 0.85846843]

1(c).

The figure below shows the training data and decision boundary. The red dots represent the true samples, and the blue dots represent the negative samples.



2(a).

2/(a)
$$\nabla_{W_m} l(w) = \nabla_{W_m} \sum_{i=1}^{N} \sum_{k=1}^{K} \log [p(y^{(i)}=k \mid x^{(i)}, w)]^{I(y^{(i)}=k)}$$

$$= \nabla_{W_m} \sum_{i=1}^{N} I(y^{(i)}=m) \log (p(y^{(i)}=m \mid x^{(i)}, w)) + \nabla_{W_m} \sum_{i=1}^{N} \sum_{\substack{k \neq m \\ k \neq K}}^{K} I(y^{(i)}=k) \log (p^{(i)}=k \mid x^{(i)}, w)$$

$$+ \nabla_{W_m} \sum_{i=1}^{N} I(y^{(i)}=K) \log (p(y^{(i)}=K \mid x^{(i)}, w)$$

Only $I(y^{(i)}=m)$ can be non negative.

$$\nabla_{W_m} \exp(w_m^T \phi(x^{(i)})) = \phi(x^{(i)}) \exp(w_m^T \phi(x^{(i)}))$$

So
$$\nabla_{W_m} l(w) = \sum_{i=1}^{N} I(y^{(i)}=m) \nabla_{W_m} \log \frac{\exp(w_m^T \phi(x))}{1+\sum_{j=1}^{K-1} \exp(w_j^T \phi(x))}$$

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

$$= \sum_{i=1}^{N} \phi(x^{(i)}) [I(y^{(i)}=m) - p(y^{(i)}=m \mid x^{(i)}, w)]$$

2(b).

The SoftMax regression model implemented using the results in part (a) has an accuracy of 94.0 %

The logistic regression function in sklearn got 92% accuracy.

3(a).

3/(a)
$$p(y^{(i)}=0 \mid x^{(i)}) = p(x^{(i)} \mid y^{(i)}=0)\, p(y^{(i)}=0) = \frac{1-\phi}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\tfrac{1}{2}(x^{(i)}-\mu_0)^T \Sigma^{-1} (x^{(i)}-\mu_0))$$

$$p(y^{(i)}=1 \mid x^{(i)}) = p(x^{(i)} \mid y^{(i)}=1)\, p(y^{(i)}=1) = \frac{\phi}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp(-\tfrac{1}{2}(x^{(i)}-\mu_1)^T \Sigma^{-1} (x^{(i)}-\mu_1))$$

Since $\sigma(a) = \frac{1}{1+\exp(-a)}$, $a = \ln(\frac{\sigma}{1-\sigma})$

$$\Rightarrow a = \ln \frac{p(y^{(i)}=1 \mid x^{(i)})}{p(y^{(i)}=0 \mid x^{(i)})} = \ln \frac{\exp(-\frac{1}{2}(x^{(i)}-\mu_0)^T \Sigma^{-1} (x^{(i)}-\mu_0))}{\exp(-\frac{1}{2}(x^{(i)}-\mu_1)^T \Sigma^{-1} (x^{(i)}-\mu_1))} + \ln \frac{\phi}{1-\phi}$$

$$= (-\tfrac{1}{2}(x^{(i)}-\mu_1)^T \Sigma^{-1} (x^{(i)}-\mu_1)) - (-\tfrac{1}{2}(x^{(i)}-\mu_0)^T \Sigma^{-1} (x^{(i)}-\mu_0)) + \ln \frac{\phi}{1-\phi}$$

$$= (\mu_1-\mu_0)^T \Sigma^{-1} x^{(i)} - \tfrac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \tfrac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \ln \frac{\phi}{1-\phi}$$

$$= W^T x + w_0,$$

$w^T = (\mu_1-\mu_0)^T \Sigma^{-1}$, $\quad w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \ln \frac{\phi}{1-\phi}$

refine x to M+1 dim by adding $x_0=1$, then the above can be written as

$$w^T x + w_0 x_0 = w^T x_{new}, \quad x_{new} \in R^{M+1 \times m}$$

$$\Rightarrow p(y=1 \mid x, \phi, \Sigma, \mu_0, \mu_1) = \frac{1}{1+\exp(-a)} = \frac{1}{1+\exp(-w^T x_{new})}$$

3(b).

3/(b) $\ell(\quad, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{N} p(x^{(i)}|y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)\, p(y^{(i)}; \phi)$

$$= \sum_{i=1}^{N} -\log(\sigma(2\pi)^{\frac{1}{2}}) - \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)$$

$$\frac{\partial \ell}{\partial \phi} = 0 \Rightarrow \sum_{i=1}^{N}\left(\frac{y^{(i)}}{\phi} + \frac{1-y^{(i)}}{\phi-1}\right) = \sum_{i=1}^{N} \frac{\phi - y^{(i)}}{\phi(\phi-1)} = 0 \Rightarrow \phi = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} = \frac{1}{N}\sum_{i=1}^{N} 1(y^{(i)}=1)$$

$$\frac{\partial \ell}{\partial \mu_0} = 0 \Rightarrow \sum_{i=1}^{N} \Sigma^{-1}(x^{(i)} - \mu_0)\, 1\{y^{(i)}=0\} = 0 \Rightarrow \mu_0 = \frac{\sum_{i=1}^{N} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)}=0\}}$$

$$\frac{\partial \ell}{\partial \mu_1} = 0 \Rightarrow \sum_{i=1}^{N} \Sigma^{-1}(x^{(i)} - \mu_1)\, 1\{y^{(i)}=1\} = 0 \Rightarrow \mu_1 = \frac{\sum_{i=1}^{N} 1\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)}=1\}}$$

$$\frac{\partial \ell}{\partial \Sigma} = 0 \Rightarrow \text{Let } S = \Sigma^{-1},\ -\frac{1}{2}\sum_{i=1}^{N} \nabla_{\Sigma}\left[-\log|S| + (x^{(i)} - \mu_{y^{(i)}})^T S (x^{(i)} - \mu_{y^{(i)}})\right] = 0$$

$$\Rightarrow -\frac{1}{2}\sum_{i=1}^{N}\left(-S^{-1} + (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\right)$$

$$= \sum_{i=1}^{N} \frac{1}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T = 0$$

$$\Rightarrow \frac{1}{2} N \Sigma = \frac{1}{2}\sum_{i=1}^{N} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

$$\Rightarrow \Sigma = \frac{1}{N}\sum_{i=1}^{N} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

3(c).

3/(c). ϕ remains the same as (b): $\phi = \frac{1}{N}\sum_{i=1}^{N} 1\{y^{(i)}=1\}$.

$$\frac{d\ell}{d\mu_0} = \sum_{i=1}^{N} -\frac{1}{2}\frac{d}{d\mu_0}\left[(x^{(i)} - \mu_0)^T \Sigma^{-1}(x^{(i)} - \mu_0)\right] 1(y^{(i)}=0)$$

$$= \sum_{i=1}^{N} -\Sigma^{-1}((x^{(i)} - \mu_0)\, 1\{y^{(i)}=0\}$$

$$\frac{d\ell}{d\mu_1} = \sum_{i=1}^{N} -\frac{1}{2}\frac{d}{d\mu_1}\left[(x^{(i)} - \mu_1)^T \Sigma^{-1}(x^{(i)} - \mu_1)\right] 1(y^{(i)}=1)$$

$$= \sum_{i=1}^{N} -\Sigma^{-1}((x^{(i)} - \mu_1)\, 1\{y^{(i)}=1\}$$

$$\Rightarrow \mu_0 = \frac{\sum_{i=1}^{N} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)}=0\}}, \quad \mu_1 = \frac{\sum_{i=1}^{N} 1\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)}=1\}}$$

4(a).

The Naive Bayes model has a test error rate of 1.625 %

4(b).

Using the model trained in (a), the top 5 tags that are most indicative of spam categories are "httpaddr", "spam", "unsubscrib", "ebai", "valet"

4(c).

Below is a plot of the test error with respect to the training set size. It can be seen that the training set containing 1400 texts gives the smallest test error of 1.6250%