Problem 1

by the properties of metrices, it's clear that

$$\begin{aligned}&\sum_{i=1}^{n}(f_{\boldsymbol{w},b}(\boldsymbol{x}_i)-y_i)^2\\=&|(\boldsymbol{w}^T\boldsymbol{X})^T-\boldsymbol{y}|^2\\=&\left(\boldsymbol{w}^T\boldsymbol{X}+b\cdot\mathbf{1}^T-\boldsymbol{y}^T\right)\left(\boldsymbol{X}^T\boldsymbol{w}+b\cdot\mathbf{1}-\boldsymbol{y}\right)\end{aligned}\tag{1}$$

let

$$\begin{aligned}\tilde{\boldsymbol{w}}&=\begin{pmatrix}b\\\boldsymbol{w}\end{pmatrix}\\\tilde{\boldsymbol{x}}&=\begin{pmatrix}1\\\boldsymbol{x}\end{pmatrix}\\\tilde{\boldsymbol{X}}&=(\tilde{\boldsymbol{x}}_1,\dots,\tilde{\boldsymbol{x}_n})\end{aligned}\tag{2}$$

with the notations above, we can rewrite the given conditions

$$\begin{aligned}f_{\boldsymbol{w},b}(\boldsymbol{x}_i)&=\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{x}}\\\sum_{i=1}^{n}(f_{\boldsymbol{w},b}(\boldsymbol{x}_i)-y_i)^2=&|(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}})^T-\boldsymbol{y}|^2\\=&\left(\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}}-\boldsymbol{y}^T\right)\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{w}}-\boldsymbol{y}\right)\\=&\boldsymbol{y}^T\boldsymbol{y}-\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{y}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{w}}+\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{w}}\end{aligned}\tag{3}$$

in order to minimize the expression above, there should be

$$\begin{aligned}&\frac{\partial}{\partial\tilde{\boldsymbol{w}}}\left(\sum_{i=1}^{n}(f_{\boldsymbol{w},b}(\boldsymbol{x}_i)-y_i)^2\right)=0\\\Longrightarrow\ &2\tilde{\boldsymbol{w}^{*}}^T\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T-2\boldsymbol{y}^T\tilde{\boldsymbol{X}}^T=0\\\Longrightarrow\ &\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{w}^*}-\boldsymbol{y}\right)^T\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{w}^*}\right)=0\\\Longrightarrow\ &y-\hat{y}\perp\hat{y}\end{aligned}\tag{4}$$

where $\tilde{\boldsymbol{w}}^{\boldsymbol{*}}$ contains the best solution $\boldsymbol{w}^*$ and b^*. Q.E.D.

Problem 2

Noting that there's only two classes of the data points, the objective function can be written as

$$\begin{aligned}J(\boldsymbol{w},b)=&\frac{1}{N}\sum_{i=1}^{N}L\big(f(\boldsymbol{x}_i;\boldsymbol{w},b)\big)\\=&\frac{1}{N}\left[\sum_{i_1=1}^{N_1}L\big(f(\boldsymbol{x}_{i_1};\boldsymbol{w},b)\big)+\sum_{i_2=1}^{N_2}L\big(f(\boldsymbol{x}_{i_2};\boldsymbol{w},b)\big)\right]\\=&\frac{1}{N}\left[\sum_{i_1=1}^{N_1}\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{x}_{i_1}+b-\frac{N}{N_1})^2+\sum_{i_2=1}^{N_2}\frac{1}{2}(\boldsymbol{w}^T\boldsymbol{x}_{i_2}+b+\frac{N}{N_2})^2\right]\end{aligned}\tag{5}$$

to get the desired solution

$$\boldsymbol{w}^*,b^*=\mathrm{argmin}_{\boldsymbol{w},b}J(\boldsymbol{w})\tag{6}$$

we firstly take the derivative w.r.t. b:

$$\begin{aligned} &\frac{\partial J(\boldsymbol{w}, b)}{\partial b} = 0 \\ \Longrightarrow\ &\frac{1}{N}\left[\sum_{i_1=1}^{N_1}(\boldsymbol{w}^T\boldsymbol{x}_{i_1} + b - \frac{N}{N_1}) + \sum_{i_2=1}^{N_2}(\boldsymbol{w}^T\boldsymbol{x}_{i_2} + b + \frac{N}{N_2}) = 0\right] \\ \Longrightarrow\ &b = -\boldsymbol{w}^T\boldsymbol{m} \end{aligned} \tag{7}$$

where

$$\boldsymbol{m} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{8}$$

is the overall mean value of $\boldsymbol{x}$.

take the derivative w.r.t. $\boldsymbol{w}$ of $J(\boldsymbol{w})$, substitute $b = -\boldsymbol{w}^T\boldsymbol{m}$:

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{w}}J(\boldsymbol{w}) = 0 \\ \Longrightarrow\ &\frac{1}{N}\left[\sum_{i_1=1}^{N_1}(\boldsymbol{w}^{*T}\boldsymbol{x}_{i_1} - \boldsymbol{w}^{*T}\boldsymbol{m} - \frac{N}{N_1})\boldsymbol{x}_{i_1}^T + \sum_{i_2=1}^{N_2}(\boldsymbol{w}^{*T}\boldsymbol{x}_{i_2} - \boldsymbol{w}^{*T}\boldsymbol{m} + \frac{N}{N_2})\boldsymbol{x}_{i_2}^T\right] = 0 \\ \Longrightarrow\ &(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T + \frac{1}{N}\boldsymbol{w}^{*T}\left[\sum_{i_1=1}^{N_1}\boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}^T + \sum_{i_2=1}^{N_2}\boldsymbol{x}_{i_2}\boldsymbol{x}_{i_2}^T - \boldsymbol{m}\sum_{i=1}^{N}\boldsymbol{x}_i^T\right] = 0 \\ \Longrightarrow\ &\left[\sum_{i_1=1}^{N_1}\boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}^T + \sum_{i_2=1}^{N_2}\boldsymbol{x}_{i_2}\boldsymbol{x}_{i_2}^T - N\boldsymbol{m}\boldsymbol{m}^T\right]\boldsymbol{w}^* = N(\boldsymbol{m}_1 - \boldsymbol{m}_2) \end{aligned} \tag{9}$$

while

$$\begin{aligned} S_W &= \sum_{i_1=1}^{N_1}(\boldsymbol{x}_{i_1} - \boldsymbol{m}_1)(\boldsymbol{x}_{i_1} - \boldsymbol{m}_1)^T + \sum_{i_2=1}^{N_2}(\boldsymbol{x}_{i_2} - \boldsymbol{m}_2)(\boldsymbol{x}_{i_2} - \boldsymbol{m}_2)^T \\ &= \sum_{i_1=1}^{N_1}\boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}^T - N_1\boldsymbol{m}_1\boldsymbol{m}_1^T + \sum_{i_2=1}^{N_2}\boldsymbol{x}_{i_2}\boldsymbol{x}_{i_2}^T - N_2\boldsymbol{m}_2\boldsymbol{m}_2^T \\ &= \sum_{i_1=1}^{N_1}\boldsymbol{x}_{i_1}\boldsymbol{x}_{i_1}^T + \sum_{i_2=1}^{N_2}\boldsymbol{x}_{i_2}\boldsymbol{x}_{i_2}^T - N\boldsymbol{m}\boldsymbol{m}^T \end{aligned} \tag{10}$$

so if S_W is not singular, we can obtain that

$$\boldsymbol{w}^* = N \cdot S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)\ //\ S_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{11}$$

Q.E.D.

Problem 3

Part 1

Two class softmax classification

$$P = P(y = 1|\boldsymbol{x}) = \frac{\exp\left(f_1(\boldsymbol{x})\right)}{\exp\left(f_1(\boldsymbol{x})\right) + 1} \tag{12}$$

Using the maximum likelihood estimation, the two class softmax classification

$$\begin{aligned} &\max_{\boldsymbol{w}}\prod_{i=1}^{N}\left[P^{\mathbf{1}(y_i=1)}(1 - P)^{\mathbf{1}(y_i=0)}\right] \\ \Longrightarrow\ &\max_{\boldsymbol{w}}\prod_{i=1}^{N}\left[\left(\frac{\exp\left(\boldsymbol{w}_1^T\boldsymbol{x}_i + b_1\right)}{\exp\left(\boldsymbol{w}_1^T\boldsymbol{x}_i + b_1\right) + 1}\right)^{\mathbf{1}(y_i=1)}\left(\frac{1}{\exp\left(\boldsymbol{w}_1^T\boldsymbol{x}_i + b_1\right) + 1}\right)^{\mathbf{1}(y_i=0)}\right. \end{aligned} \tag{13}$$

applying negative log to the likelihood function, we get the objective function

$$\begin{aligned}\min_{\boldsymbol{w}} J(\boldsymbol{w}) &= \min_{\boldsymbol{w}} -\sum_{i=1}^{N}\left\{y_i\log\left(\frac{e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}}{e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}+1}\right)+(1-y_i)\log\left(\frac{1}{e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}+1}\right)\right\}\\ &= \min_{\boldsymbol{w}} -\sum_{i=1}^{N}\left\{y_i(\boldsymbol{w}_1^T\boldsymbol{x_i}+b_1)-\log\left(e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}+1\right)\right\}\end{aligned} \tag{14}$$

the logistic regression with target value 1 and -1 can be

$$\min_{\boldsymbol{w}} J(\boldsymbol{w}) = \min_{\boldsymbol{w}}\sum_{i=1}^{N}\log\left(1+e^{-\tilde{y}_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)}\right) \tag{15}$$

where $\tilde{y}_i \in \{-1,1\}$. Corresponding that with $y_i \in \{0,1\}$, we can simply compare the related terms. The terms with $y_i = 1$ should be

$$\begin{aligned}&-\left(\boldsymbol{w}_1^T\boldsymbol{x}+b_1-\log\left(e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}+1\right)\right)\\ =&\log\left(1+e^{-\boldsymbol{w}_1^T\boldsymbol{x}_i-b_1}\right)\\ =&\log\left(1+e^{-1\cdot(\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1)}\right)\end{aligned} \tag{16}$$

and similarly we can do this for the terms with $y_i = 0$

$$\begin{aligned}&-\left(0-\log\left(e^{\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1}+1\right)\right)\\ =&\log\left(1+e^{1\cdot(\boldsymbol{w}_1^T\boldsymbol{x}_i+b_1)}\right)\end{aligned} \tag{17}$$

Q.E.D.

Part 2

the gradient of loss function w.r.t. $\boldsymbol{w}_i$

$$\begin{aligned}&\frac{\partial}{\partial\boldsymbol{w}_i}\left(-\log P(y=k|\boldsymbol{x})\right)\\ =&-\frac{\sum_{j=1}^{K}\exp\left(f_j(x)\right)}{\exp\left(f_k(\boldsymbol{x})\right)}\frac{\partial}{\partial\boldsymbol{w}_i}\left[\frac{\exp\left(f_k(\boldsymbol{x})\right)}{\sum_{j=1}^{K}\exp\left(f_j(x)\right)}\right]\\ =&-\frac{\sum_{j=1}^{K}\exp\left(f_j(x)\right)}{\exp\left(f_k(\boldsymbol{x})\right)}\left[\frac{\exp\left(f_k(\boldsymbol{x})\right)}{\sum_{j=1}^{K}\exp\left(f_j(x)\right)}\boldsymbol{x}^T-\frac{\left(\exp\left(f_k(x)\right)\right)^2}{\left(\sum_{j=1}^{K}\exp\left(f_j(x)\right)\right)^2}\boldsymbol{x}^T\right]\\ =&(P-1)\boldsymbol{x}^T\end{aligned} \tag{18}$$

Part 3

Noting that only the direction of $\boldsymbol{w}$ matters, and the magnitude of b decreases when the magnitude of $\boldsymbol{w}$ decreases (maintaining the direction of $\boldsymbol{w}$), so we can simply scale the logits by a small number e.g. 0.01, so that therefore overflow would not occur, and a moderate threshold should be set to adapt to the change.

Problem 4

Part 1

Adopting Fisher's criterion to the data `breast-cancer-wisconsin.txt` , and taking the first 70 data points as training set, the last 600 data points as test set, we finally get the result as below:

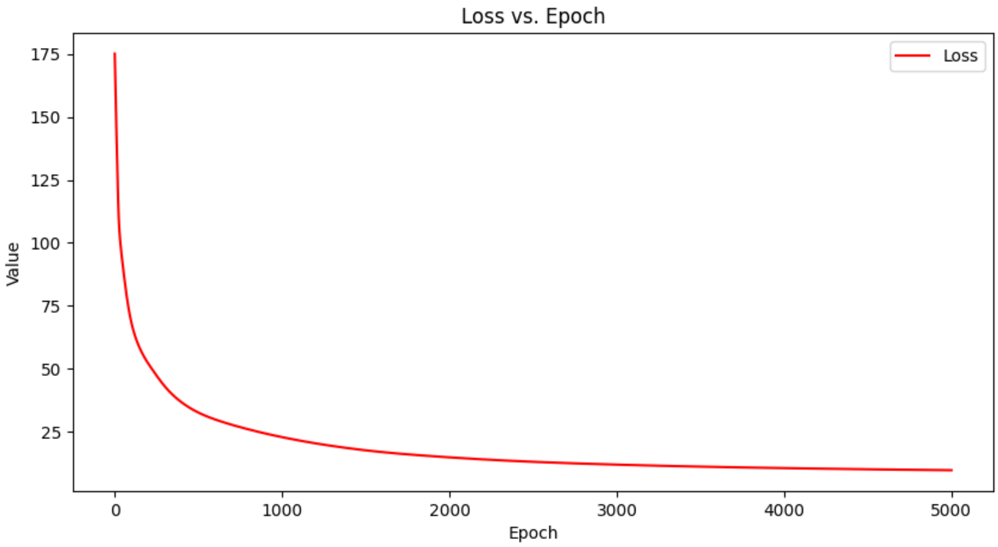
normalized 9-dimensional vector $\boldsymbol{w}^*$

$$\boldsymbol{w}^* = \begin{pmatrix} 0.62795825 \\ -0.17865309 \\ 0.27499601 \\ 0.23786778 \\ 0.35489863 \\ 0.25557686 \\ -0.00713858 \\ 0.44309115 \\ 0.23241913 \end{pmatrix} \tag{19}$$

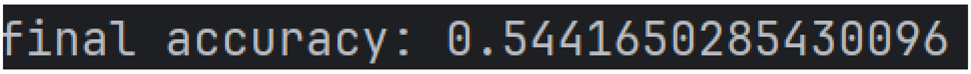
and the accuracy is `0.55445` .

Part 2

in my program, the ρ used is $\rho = 0.1$, and the loss value against iterations are shown below:



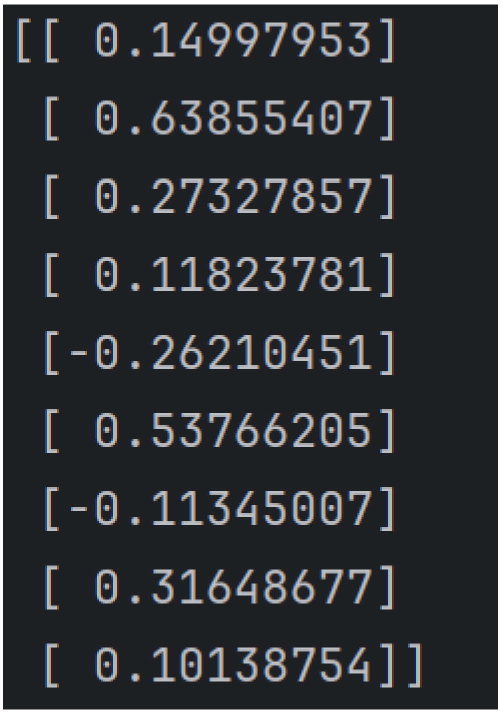
and the final classification accuracy is



```
final accuracy: 0.5441650285430096
```

Part 3

the w for logistic LDA



```
[[ 0.14997953]
 [ 0.63855407]
 [ 0.27327857]
 [ 0.11823781]
 [-0.26210451]
 [ 0.53766205]
 [-0.11345007]
 [ 0.31648677]
 [ 0.10138754]]
```

then we can compute the cosine between the two w^* is `0.2924` , indicating they are quite not similar to each other. That's because they used different objective functions, and the determining directions are therefore different.

The most indicative feature should have the greatest absolute value of the corresponding components of w^* , from the two w^*, we can easily find the most indicative feature is feature 2.