Question 1

(a)

From the lecture, we know in classification problem, the weight update formula is

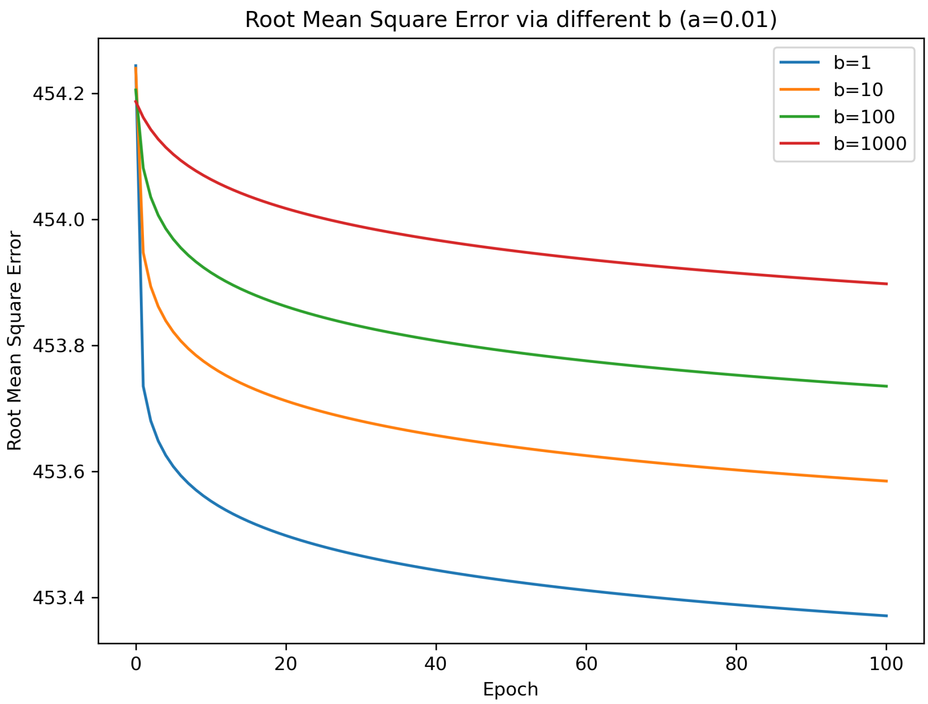
$$w(i+1) \leftarrow w(i) - \eta(i)(w(i)^T X - b)X \tag{1}$$

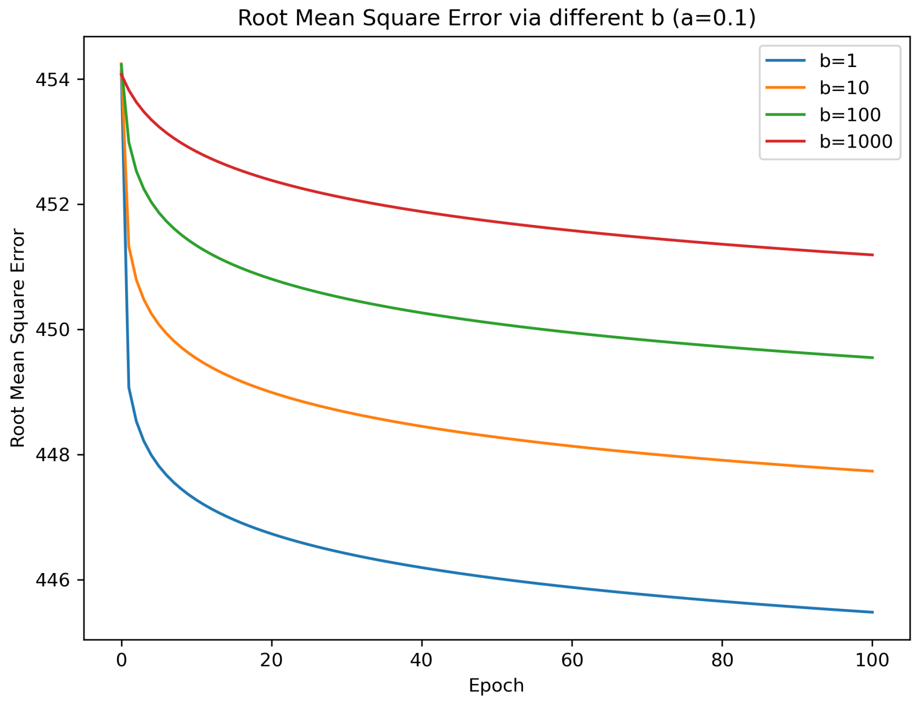
In the regression problem, the weight update formula is

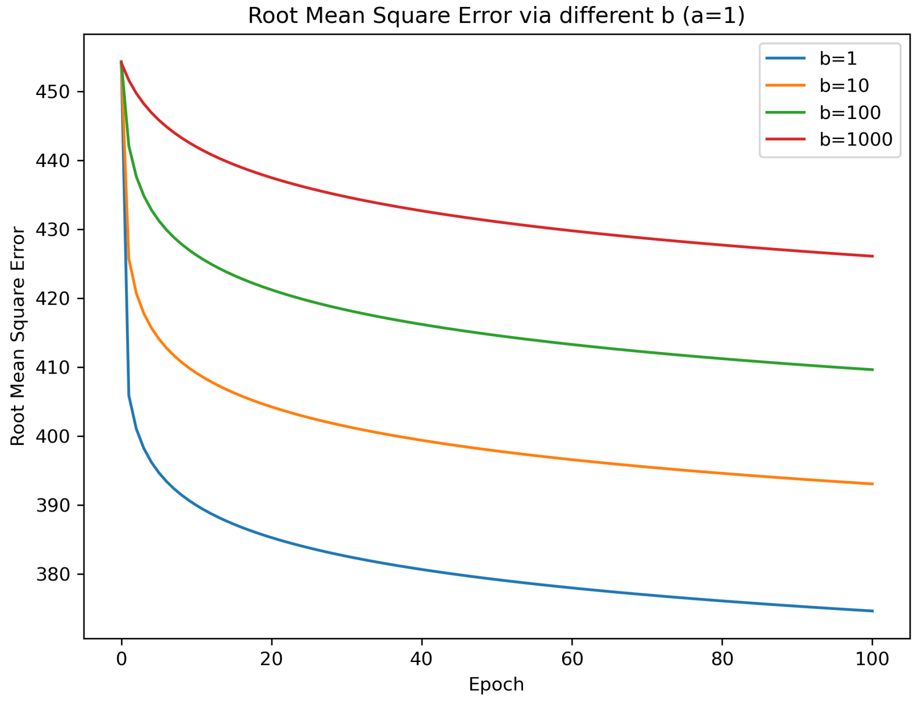
$$w(i+1) \leftarrow w(i) - \eta(i)(w(i)^T X - y)X \tag{2}$$

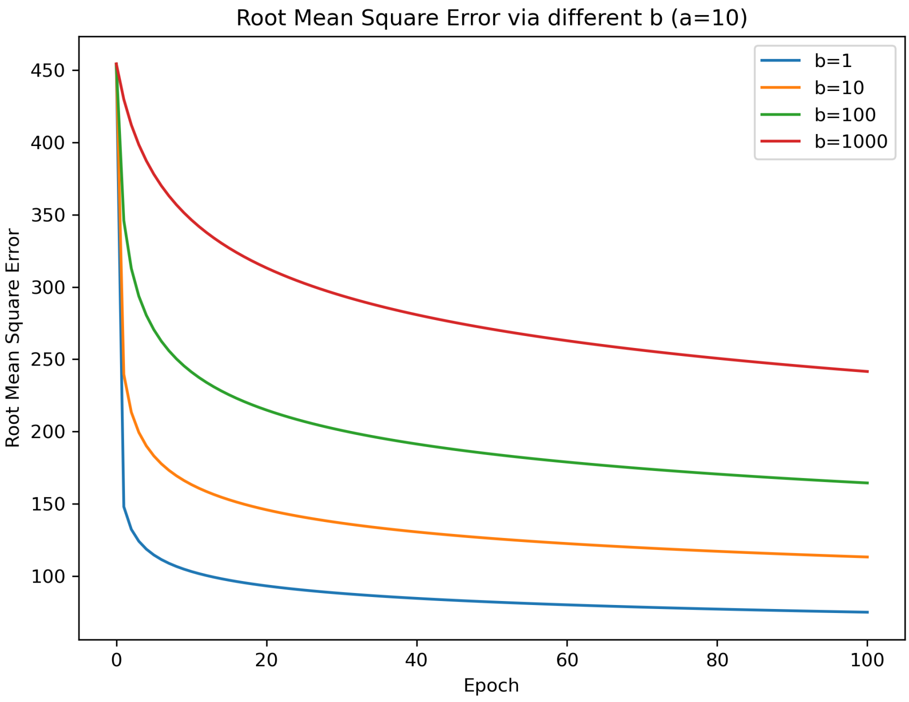
Because the weight in the classification problem will be convergence, which means $\eta(i)(w(i)^T X - b)X < \epsilon, i \to \infty$. Because X, y, b is constant, then we can obtain that $\eta(i)(w(i)^T X - y)X < \epsilon, i \to \infty$, it means that the weight in the regression problem is also convergence.

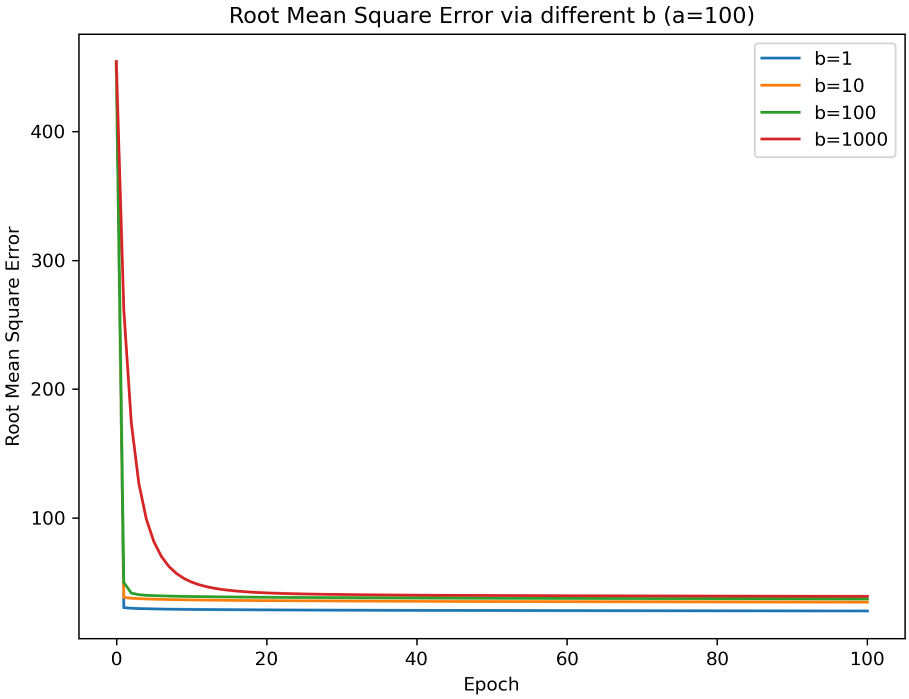
(b)




Root Mean Square Error via different b (a=0.1)
b=1
b=10
b=100
b=1000
Root Mean Square Error
Epoch


Root Mean Square Error via different b (a=1)
b=1
b=10
b=100
b=1000
Root Mean Square Error
Epoch





(c)

From the figures above, we can find that the learning curve is dependent on A and B. While A is large, the rse loss is also very large. And the rse loss is low if b is small.

(d)

The best pair is (A=100, B=1)

The best rse loss is 30.667

(e)

The regressor's error is not substantially lower than the error of this trivial regressor.

Question 2

If we apply the regularization on the non-augmented weight vector, and the bias term is only used to minimize the MSE, then we can obtain the objective function below:

$$J(w) = ||\underline{X}^{+}w - y||^2 + \lambda||I^{'}w||^2 \tag{3}$$

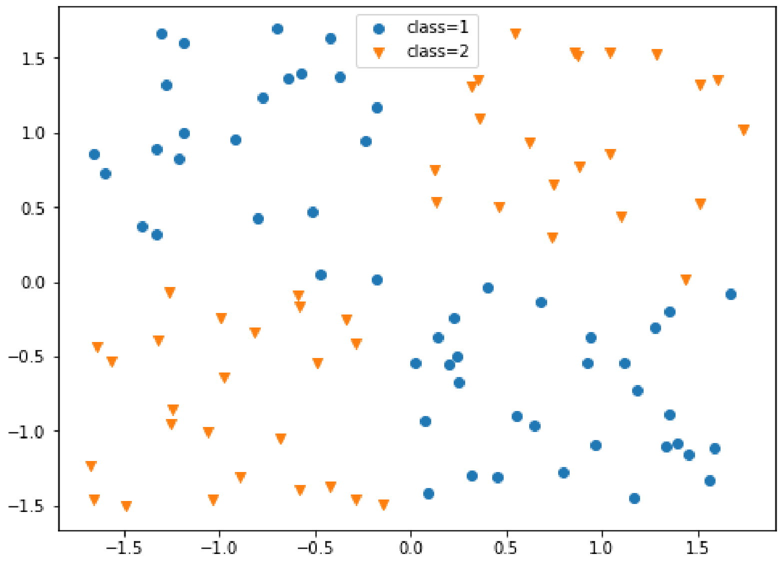
Compute the gradient of it, we can get that:

$$\begin{aligned}
\frac{\partial J(\underline{w})}{\partial \underline{w}} &= \frac{\partial(||\underline{X}^{(+)}\underline{w} - \underline{y}||^2 + \lambda||\underline{I}^{'}\underline{w}||^2)}{\partial \underline{w}} \\
&= \frac{\partial(\underline{w}^{(+)T}\underline{X}^{(+)^T}\underline{X}^{(+)}\underline{w} - \underline{y}^T\underline{X}^{(+)}\underline{w} - \underline{w}^T\underline{X}^{(+)^T}y + y^Ty + \lambda\underline{w}^T\underline{I}^{'T}\underline{I}^{'}\underline{w})}{\partial \underline{w}} \\
&= 2\underline{X}^{(+)^T}\underline{X}^{(+)}\underline{w} - 2\underline{X}^{(+)^T}y + 2\lambda\underline{I}^{'T}\underline{I}^{'}\underline{w} = 0
\end{aligned} \quad (\cdot$$

Thus, we can get that the optimal $\hat{\underline{w}} = (\underline{X}^{(+)^T}\underline{X}^{(+)} + \lambda\underline{I}^{'T}\underline{I}^{'})^{-1}\underline{X}^{(+)^T}\underline{y} = (\underline{X}^{(+)^T}\underline{X}^{(+)} + \lambda\underline{I}^{'})^{-1}\underline{X}^{(+)^T}\underline{y},$ where $\underline{I}^{'} = \underline{X}^{(+)}\underline{X}^{(+)^T}\text{diag}(0, 1, 1, \cdots, 1)$

Question 3

(a)

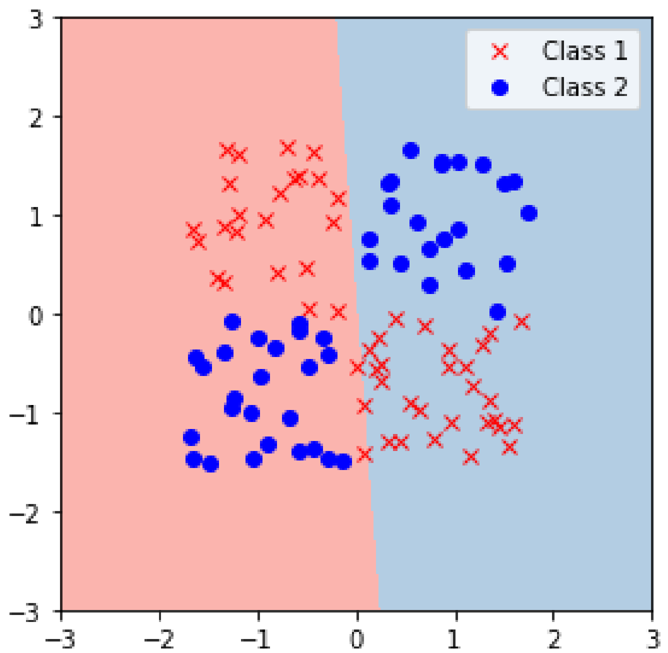


From the figure above, the data is not linearly separable in this feature space.

(b)

The original accuracy score is 0.530

(c)



(d)

The accuracy of the new data is 1.000

The new dataset is linearly separable in this feature space.