Part1

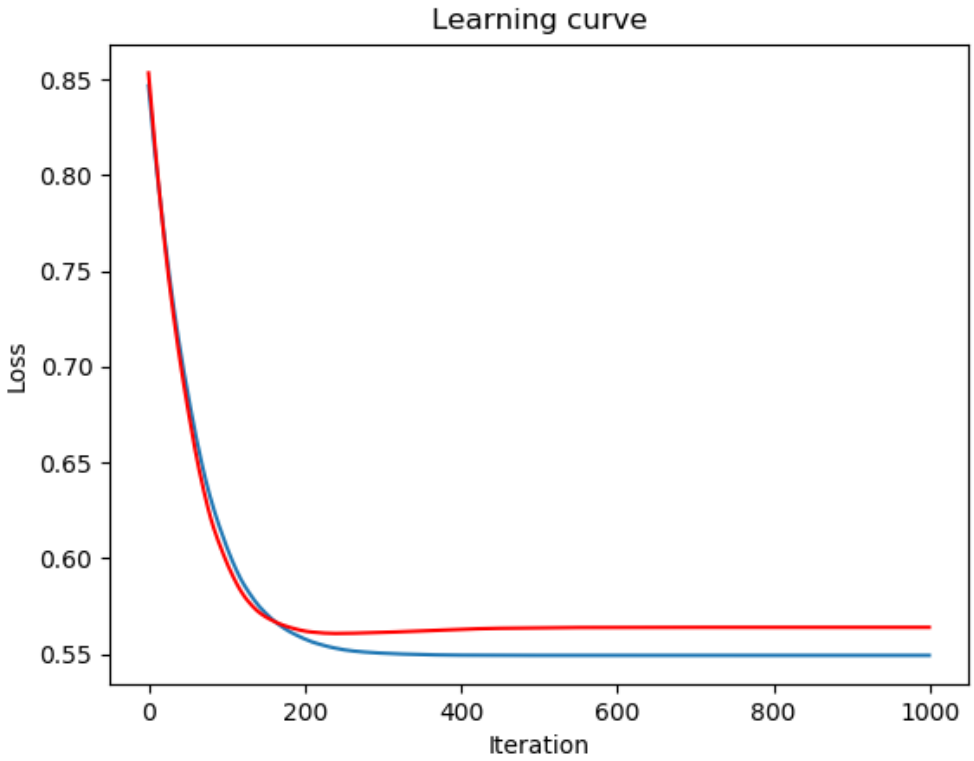
1.

Main 裡的 error_type 可以設”MAE” or “MSE”

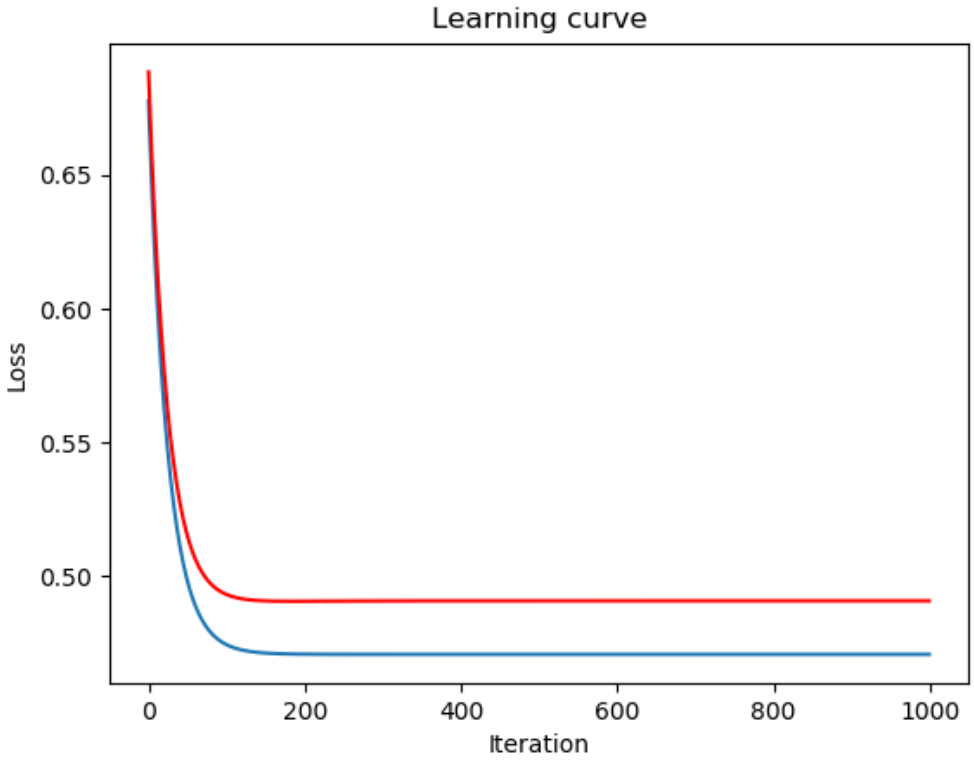
2.

紅線是 test_data

MAE



MSE



3.

MAE:0.56391

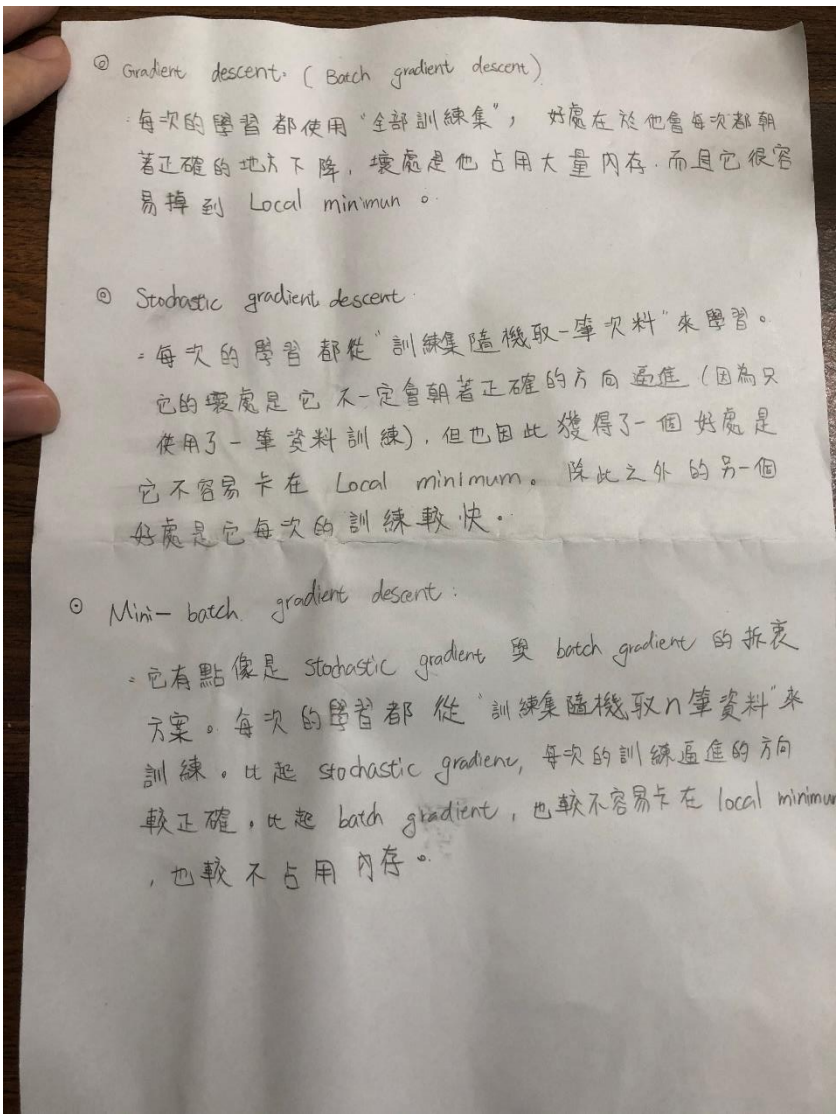
MSE:0.4909

4.

MSE: (β1,β0) = (0.45273457 , -0.0012672)

MAE: (β1,β0) = (0.43505747 , -0.03806984)

5.


Gradient descent. (Batch gradient descent)
每次的學習都使用"全部訓練集"，好處在於他會每次都朝著正確的地方下降，壞處是他占用大量內存，而且它很容易掉到 Local minimun。
Stochastic gradient descent
每次的學習都從"訓練集隨機取一筆次料"來學習。它的壞處是它不一定會朝著正確的方向逼近（因為只使用了一筆資料訓練），但也因此獲得了一個好處是它不容易卡在 Local minimum。除此之外的另一個好處是它每次的訓練較快。
Mini- batch gradient descent
它有點像是 Stochastic gradient 與 batch gradient 的折衷方案。每次的學習都從"訓練集隨機取n筆資料"來訓練。比起 stochastic gradient，每次的訓練逼近的方向較正確，比起 batch gradient，也較不容易卡在 local minimum，也較不占用內存。

Part2

1

① $P(R)\cdot\frac{3}{10} + P(B)\cdot\frac{2}{4} + P(G)\cdot\frac{4}{20}$

$= \frac{2}{10}\cdot\frac{3}{10} + \frac{4}{10}\cdot\frac{2}{4} + \frac{4}{10}\cdot\frac{4}{20}$

$= \frac{6}{100} + \frac{20}{100} + \frac{8}{100}$

$= \frac{34}{100} = 0.34$ #

② $P(R)\cdot\frac{3}{10} + P(B)\cdot\frac{2}{4} + P(G)\cdot\frac{12}{20}$

$= \frac{2}{10}\cdot\frac{3}{10} + \frac{4}{10}\cdot\frac{2}{4} + \frac{4}{10}\cdot\frac{12}{20}$

$= \frac{50}{100}$

$\frac{P(B)\cdot\frac{2}{4}}{\frac{50}{100}} = \frac{4}{10} = 0.4$ #

2. $\mathrm{Var}(f) = E(f(x) - E(f(x))^2)$

$= E(f(x)^2 - 2f(x)\cdot E(f(x)) + E(f(x))^2)$

$= E(f(x)^2) - 2E(f(x))\cdot E(f(x)) + E(f(x))^2$

$= E(f(x)^2) - 2E(f(x))^2 + E(f(x))^2$

$= E(f(x)^2) - E(f(x))^2$ #

3. $E_y(E_x(X|Y)) = \sum_y E(X|Y=y)\cdot P(Y=y)$

$= \sum_y \left(\sum_x x\cdot P(X=x|Y=y)\right)\cdot P(Y=y)$

$= \sum_y \sum_x x\cdot P(X=x|Y=y)\cdot P(Y=y)$

$= \sum_y \sum_x \left(x\cdot \frac{P(X=x, Y=y)}{P(Y=y)}\right)\cdot P(Y=y)$

$= \sum_x x \sum_y P(X=x, Y=y)$

$= \sum_x x\cdot P(X=x)$

$= E(X)$