Pattern Recognition HW1

M093781 趙宇涵

Part 1. Coding (60%)

Please find the following link to check my coding:
https://github.com/honey0703/CS_AT0828/blob/main/HW1/HW1.ipynb
Here I edited the hyper parameters to leaning rate = 1e-3, iterations = 10000 to get a better training result.

1. (15%) Plot the learning curve of the training, you should find that loss decreases after a few iterations (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)

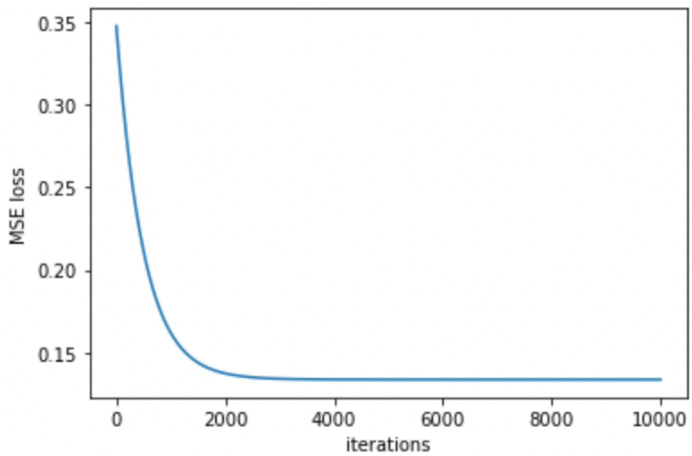


Fig.1 Learning curve of training data

2. (15%) What's the Mean Square Error of your prediction and ground truth (prediction=model(x_test), ground truth=y_test)
   After I repeat training for more than 5 times, my MSE losses are about **0.03435**

3. (15%) What're the weights and intercepts of your linear model?
   Weight: 0.81795508
   Intercepts: 0.7845605

```
In [13]: # Problem 3. Weight and Intercepts.
         print ('Weight: ', theta, '\nIntercepts: ', bias)

         Weight:  [0.81795508]
         Intercepts:  [0.7845605]
```

Fig.2 weight and intercepts

4. (10%) What's the difference between Gradient Descent, Mini-Batch Gradient Descent, and Stochastic Gradient Descent?

The main difference between these three methods is the different size of training dataset. Here I introduce their training data and their pros and cons individually.

- ♦ Gradient Descent:

  The hole training dataset is used during training process. It updates the model after calculating all loss of entire training dataset.

  - Pros: It can get a smooth curve, such as Fig.1. It can continually reach the lowest loss score.
  - Cons: The progress will take long time because all training data need to be considered during 1 update.

- ♦ Mini-Batch Gradient Descent:

  It samples few examples from training dataset during 1 update. These sample data are called a mini batch.

  - Pros: It combines the pros of GD and SGD. Training process is faster than GD, and few fluctuating than SGD.
  - Cons: It combines the cons of GD and SGD. Training process is slower than SGD, and more fluctuating than GD.

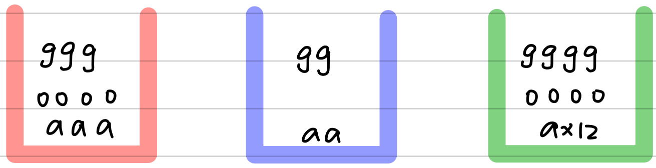
- ♦ Stochastic Gradient Descent

  It only uses 1 example in 1 update.

  - Pros: Training process is much faster than GD and mini-batch GD.
  - Cons: Due to updating with one example, the loss won't always decrease. Thus, the learning curve will fluctuate seriously.

Part 2. Questions (40%) M093781 趙宇涵

1.



$P(R)=0.2$ $P(B)=0.4$ $P(G)=0.4$

(1) Probability of selecting guava?

$P(g|R) + P(g|B) + P(g|G)$

$= (0.2\times 0.3) + (0.4\times 0.5) + (0.4\times 0.2)$

$= 0.34$ #

(2) Select apple, the probability that it came from blue box?

From Baye's Theorem:

$$P(B|a) = \frac{P(a|b)\,P(B)}{P(a)}$$

$$= \frac{0.5 \times 0.4}{0.2\times 0.3 + 0.4\times 0.5 + 0.4\times 0.6}$$

$$= 0.4 \ \#$$

2. <known> $a \le b$, $a \le (ab)^{1/2}$

From textbook P40. (1.78)

$P(\text{mistake}) = \int_{R_1} P(x, C_2)\,dx + \int_{R_2} P(x, C_1)\,dx$

<prof> $P(\text{mistake}) \le \int \{P(x,C_1)\,P(x,C_2)\}^{1/2}\,dx$

<sol>

Seperate P(mistake) to 2 part: $\int_{R_1} P(x,C_2)\,dx$, $\int_{R_2} P(x,C_1)\,dx$

From textbook Fig 1.24

in range R_1: $\because p(x,C_2) \le P(x,C_1)$

$\therefore p(x,C_2) \le \{p(x,C_2)\,p(x,C_1)\}^{1/2}$

$\int_{R_1} P(x,C_2)\,dx \le \int_{R_1} \{P(x,C_2)\,p(x,C_1)\}^{1/2}\,dx$ .... ①

in range R_2 $\because p(x,C_1) \le p(x,C_2)$

$\therefore p(x,C_1) \le \{p(x,C_1)\,p(x,C_2)\}^{1/2}$

$\int_{R_2} P(x,C_1)\,dx \le \int_{R_2} \{P(x,C_1)\,P(x,C_2)\}^{1/2}\,dx$ ...②

sum ① ②

$$\int_{R_1} p(x,C_2)\,dx + \int_{R_2} p(x,C_1)\,dx \le \int \{P(x,C_1)\,P(x,C_2)\}^{1/2}\,dx$$

$$P(\text{mistake}) \le \int \{P(x,C_1)\,P(x,C_2)\}^{1/2}\,dx \ \#$$

3. <prof> 1. $E[x] = E_y[E_x[x|y]]$

2. $var[x] = E_y[var_x[x|y]] + var_y[E_x[x|y]]$

<sol> 1. $E_y[E_x[x|y]] = \int_{-\infty}^{\infty} E(X|Y=y)\, f_Y(y)\, dy$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, dx \;\; f_Y(y)\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x|y)\, f_Y(y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dx\, dy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy\, dx$$

$$= \int_{-\infty}^{\infty} x f_X(x)\, dx$$

$$= E(x) \quad \#$$

2. $\because E(x) = \int E(X|Y=y) \times p(y)\, dy$

$= E[E(X|Y)]$

$\therefore E[E(X|Y)] = E(X)$

$E[E(X^2|Y)] = E(X^2)$

$Var[E(X|Y)] + E[Var(X|Y)]$

$= E\{E(X|Y) - E[E(X|Y)]\}^2 + E\{E(X^2|Y) - [E(X|Y)]^2\}$

$= E[E(X|Y)]^2 - \{E[E(X|Y)]\}^2 + E\{E(X^2|Y) - [E(X|Y)]^2\}$

$= E[\cancel{E(X|Y)}]^2 - \{E[E(X|Y)]\}^2 + E[E(X^2|Y)] - E[\cancel{E(X|Y)}]^2$

$= E[E(X^2|Y)] - \{E[E(X|Y)]\}^2$

$= E(X^2) - E(X)^2$

$= Var[X] \quad \#$