Pattern Recognition, Homework 1

周芝妤 **0886004**

Part. 1, Coding (70%):

In this coding assignment, you are required to implement linear regression by using only NumPy, then train your model using **gradient descent** on the provided dataset, and evaluate the performance on testing data. Find the sample code and data on the GitHub page https://github.com/NCTU-VRDL/CS_ILE5065/tree/main/HW1

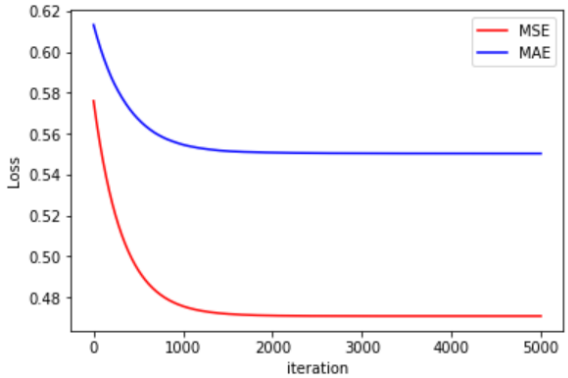
Please note that only NumPy can be used to implement your model, you will get 0 points by calling sklearn.linear_model.LinearRegression. Moreover, please train your regression model using gradient descent, not the closed-form solution.

1. (15%) Implement the linear regression model and train it by using **gradient descent** with **mean absolute error** and **mean square error** as the objective function, respectively

   Ans: As shown in my .ipynb file.

2. (15%) Plot the learning curve of the training with both losses in the same figure, you should find that loss decreases and converges after a few iterations (x-axis=iteration, y-axis=loss, Matplotlib or other plot tools is available to use)

   Ans: As the following picture which copied from my jupytor notebook. Red line represents the mean square error (MSE) and blue line represents the mean absolute error (MAE).



3. (15%) What're the mean square error and mean absolute error between your predictions and the ground truths on the testing data (prediction=model(x_test), ground truth=y_test)

   Ans:
   MSE = 0.49089334870602513, MAE = 0.5613094758418951

4. (10%) What're the weights (β_1) and intercepts (β_0) of your linear model trained from both losses?

   Ans:
   My weights (β_1) is [0.45289541] and my Bias (aka intercepts, $\beta 0$) is [-0.00123045]

5. (10%) Please explain the difference between gradient descent, mini-batch gradient descent, and stochastic gradient descent?
   Ans:

	The biggest difference
gradient descent	Need to go through all training data for deciding every step (gradient).
mini-batch	Cutting the training data as "mini-batches". For example, we cut 5,000,000 data into 1000 mini-batches (5000 data in one batch), so one iteration (epoch) can finish 1000 gradient update. We can use this method when we have a large training data, and we need to check the one batch can fit in the CPU/GPU.
stochastic	This method is just like mini-batch gradient decent, but only randomly run one "mini-batch" to update the gradient. The advantage of this method is the less computation and running, but sometimes it is less efficient.

6. (5%) All your codes should follow the PEP8 coding style and with clear comments

Part. 2, Questions (30%):

The answer of Part2 is handed written and showed in following pictures.

1. (10%) Suppose that we have three colored boxes R (red), B (blue), and G (green). Box R contains 3 apples, 4 oranges, and 3 guavas, box B contains 2 apples, 0 orange, and 2 guavas, and box G contains 12 apples, 4 oranges, and 4 guavas. If a box is chosen at random with probabilities p(R)=0.2, p(B)=0.4, p(G)=0.4, and a piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting guava? If we observe that the selected fruit is in fact an apple, what is the probability that it came from the blue box?

2. (10%) Using the definition $\mathrm{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$ show that $\mathrm{var}[X]$ satisfies $\mathrm{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

3. (10%) Consider two variables X and Y with joint distribution $p(X, Y)$. Prove the following result

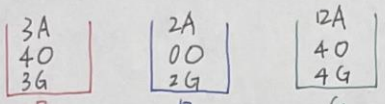
$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$$

   Here $\mathbb{E}_X[X|Y]$ denotes the expectation of X under the conditional $p(X|Y)$, with a similar notation for the conditional variance.

   Hint: Please check the definitions of the expectation operator, the sum rule, and the product rule.

Part 2

1.



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

(a) A piece of fruit is removed, the probability of selecting guava:

$0.2 \times \frac{3}{10} + 0.4 \times \frac{2}{4} + 0.4 \times \frac{4}{20}$ (marginal probability I think).

$= 0.2 \times 0.3 + 0.4 \times 0.5 + 0.4 \times 0.2 = 0.34$ ✻

(b) Apple is selected, probability of coming from blue box:

I think this case is conditional probability. $P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$

$$P(\text{Blue box} \mid \text{Apple}) = \frac{P(\text{Apple} \mid \text{Blue box})\, P(\text{Blue box})}{P(\text{Apple})} = \frac{2/4 \times 0.4}{17/34}$$

$$= 0.4$$ ✻

2、 $Var[f] = E[(f(x) - E[f(x)])^2] \Rightarrow Var[f] = E[f(x)^2] - E[f(x)]^2$

According to page 57 in the slides.

We know the average value of a random variable x is:

$$E[x] = \int N(x \mid \mu, \sigma^2)\, x\, dx = \mu \quad \text{(mean)}$$

And the second order moment of variable x is:

$$E[x^2] = \int N(x \mid \mu, \sigma^2)\, x^2\, dx = \mu^2 + \sigma^2 \quad \text{(mean plus variance)}$$

So the variance of x ($f(x) = x$ in this case)

$Var[f(x)] = Var[x] = E[(f(x) - E[f(x)])^2]$ and it also equal to σ^2 (variance)

$\Rightarrow \sigma^2 = (\mu^2 + \sigma^2) - \mu^2 = E[x^2] - (E[x])^2$

So $Var[f] = E[f(x)^2] - (E[f(x)])^2$ make sense ✻

3. $P(x|y)$ is the joint probability of variable x, y, prove $E[x] = E_y[E_x[x|y]]$

Let $f(x) = x$, $f(y) = E_x[x|y]$

$$E_y[E_x[x|y]] = E_y[f(y)] = \sum_y p(y) f(y) = \sum_y p(y) E_x[x|y]$$

$$= \sum_y p(y) \sum_x p(x|y) x = \sum_y \sum_x p(y) p(x|y) x = \sum_y \sum_x p(y, x) x$$

$$= \sum_x p(x) x$$

Sum rule:
$P(x) = \sum_y p(x, y)$

Cause $E[x] = \sum_x p(x) x$, So $E[x] = E_y[E_x[x|y]]$ ✗

↳ $= \sum_x p(x) x$.